



# Proceedings of ICIQ 2012

The 17th International Conference on Information Quality

Paris, November 16-17, 2012

Laure Berti-Équille, Isabelle Comyn-Wattiau, Monica Scannapieco, Editors



DATAFLUX  
data management



ataccama







**The 17th International Conference on Information Quality**  
Cnam, Paris, November 16-17, 2012

<http://iciq2012.cnam.fr/>

**TABLE OF CONTENTS**

WELCOME MESSAGE FROM GENERAL CHAIRS .....	ii
WELCOME MESSAGE FROM PROGRAMME COMMITTEE CHAIRS .....	iii
2012 BALLOU-PAZER IQ DISSERTATION AWARD COMPETITION .....	iv
2012 STUART ELLIOT MADNICK BEST PAPER AWARD.....	v
CONFERENCE ORGANIZATION.....	vi
PROGRAMME COMMITTEE .....	vii
CONFERENCE VENUE.....	ix
INVITED TALKS AND KEYNOTES.....	x
ICIQ 2012 CONFERENCE PROGRAM .....	xiii
MAIN TRACK RESEARCH PAPERS .....	xvi
Session 1 – IQ and Organizations .....	1
Session 2 – IQ and Knowledge.....	46
Session 3 – Information Accuracy .....	85
Parallel Sessions: Session 4 – IQ Improvement.....	126
Parallel Sessions: Session 5 – IQ Dimensions .....	178
Parallel Sessions: Session 6 – Measurement of IQ.....	229
Parallel Sessions: Session 7 – IQ and Social Media .....	297
AUTHOR INDEX.....	341
ICIQ 2012 SPONSORS.....	342

## WELCOME MESSAGE FROM GENERAL CHAIRS

On behalf of the conference committee for ICIQ 2012, it is our pleasure to welcome you to Paris for the 17th International Conference on Information Quality. This is the second time that this conference is being held in Europe; the first time was 3 years ago in 2009 in Germany. The conference is jointly organized by the CNAM, *Conservatoire National des Arts et Métiers*, Paris and EXQI, the French association for Data Quality and Governance.

ICIQ is a premier annual international forum for data and information quality management researchers, practitioners, vendors, and application developers. The conference will feature research talks and industry presentations. It will cover current issues in information quality management in database and information systems research and development.

The conference would not have been possible if not for the efforts of many people. Thanks are due to the Organization committee and Program chairs – Dr. Laure Berti-Équille, Dr. Isabelle Comyn-Wattiau and Dr. Monica Scannapieco and their PC members for producing an exciting programme. Thanks also to the efforts of the industrial Track chair – Sylvaine Nugier, the Publicity chair – Delphine Clément, and the Local Organization Committee, in particular: Alexandre, Anne, Natacha, Yura, Frédéric, Sarah, Amina, Samira, Nadine, Joël and Henri.

We would like also to warmly thank Andy Koronios and Jing Gao from University of South Australia (ICIQ-2011 Co-Chairs) and John Talburt and Liz Pierce from University of Arkansas at Little Rock, USA (ICIQ-2010 Co-Chairs) for organizing the previous two ICIQ conferences and the MIT IQ Programme liaison – Richard Wang. We wish the 17<sup>th</sup> conference (ICIQ-2012) to be as successful as the previous ones, and continue the tradition for ICIQ-2013 in Little Rock, USA again, and ICIQ-2014 in Xi’An, China to be hosted by the School of Management, Xi’An Jiaotong University.

We are grateful for the generous support of our Platinum Sponsors – IBM, SAS Dataflux, Gold Sponsors – Ataccama, EDF, GDE, ScoringData and Silver Sponsors – REVER and Steria.

We wish to thank our academic distinguished speakers: Stuard Madnick from MIT and Felix Naumann from Hasso Plattner Institut der Universität Potsdam in Germany and our industry keynote speaker: Marielle Vo-Van from Bouygues Telecom.

We also wish to thank the supporting organization – EXQI.

Last but not least, our sincere thanks go to the authors of the papers, the speakers, and all the participants of ICIQ 2012 who have made this conference a resounding success.

Welcome and enjoy the conference and have a good time in Paris!

Jacky Akoka, CNAM, France

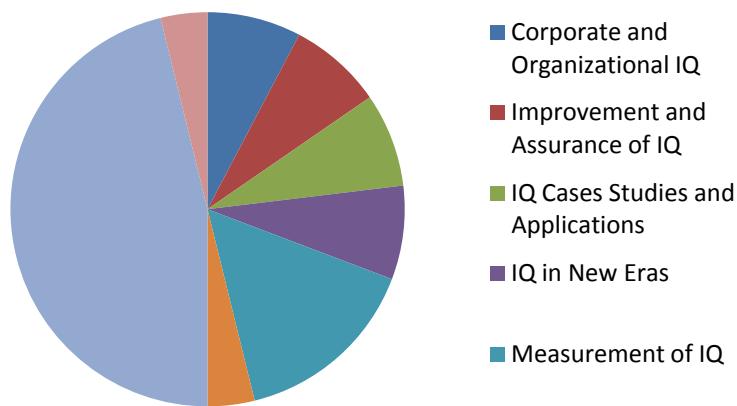
Brigitte Laboisse, ExQI, France

**ICIQ 2012 General Chairs**

## WELCOME MESSAGE FROM PROGRAMME COMMITTEE CHAIRS

The program and the organization of the conference are the result of a huge effort by many people who contributed to the success of ICIQ 2012 and we want to warmly thank them all. First, we would like to thank the authors of all submitted papers, both accepted and rejected ones.

The 24 papers collected in this volume, out of 48 papers that were submitted to the Main Track of ICIQ 2012 Conference, are a significant sample of recent achievements in the various areas of information and data quality, ranging from information quality models and evaluation frameworks to data cleaning and quality of social media data. In the following we report a pie chart representation of the distribution of the number of accepted papers by primary subject area.



Seven sessions in the Main Track along with three sessions for the Industrial Track have been proposed for the conference program. The acceptance rate of the Main Track of ICIQ 2012 (50%) is slightly higher than ICIQ 2011 (46%). The 61 members of the Program Committee were very thorough and zealous. Between three and four reviews were requested for every paper, and the selection process was regulated only by technical factors. We really wish to thank the program committee members for the reviewing work they did to ensure high-quality papers.

We wish to thank our distinguished keynote speakers: Professor Stuart Madnick from Massachusetts Institute of Technology, USA, and Professor Felix Naumann from Hasso Plattner Institut der Universität Potsdam in Germany, currently at Qatar Computing Research Institute, Qatar for their enlightening talks.

In addition, we would like to thank all the people who volunteered their time to help us organize the conference.

Finally, we thank you for attending the ICIQ 2012 conference. We sincerely hope that you find the program very exciting and enjoy the conference environment.

Laure Berti-Équille

Isabelle Comyn-Wattiau

Monica Scannapieco

**ICIQ 2012 PC Chairs**

## **2012 BALLOU-PAZER IQ DISSERTATION AWARD COMPETITION**

Following are the results for this year's best dissertations for the 2012 Ballou-Pazer IQ Dissertation Award Competition:

1. **Dr. Mohamed Yakout**, Purdue University (Advisor: Prof. Ahmed K. Elmagarmid)

Dissertation Title: *Guided Data Cleaning*

2. **Dr. Ahmed Abu Halimeh\***, University of Arkansas at Little Rock (Advisor: Dr. Mihail Tudoreanu)

Dissertation Title: *Integrating Information Quality in Visual Analytics*

and

- Dr. **Yinle Zhou\***, University of Arkansas at Little Rock (Advisor: Dr. John Talburt)

Dissertation Title: *Modeling and Design of Entity Identity Information in Entity Resolution Systems*

\*alphabetical order only

Dr. Yakout clearly had the highest ranking (.2 mean rankings points ahead) based on the vote by the members of the 2012 Ballou-Pazer IQ Dissertation Award Competition Committee and thus won the competition receiving the first prize. Dr. Halimeh and Dr. Zhu had exactly the same rankings and they thus receive both the second prize. Therefore the Committee decided this year to give a first prize followed by two second prizes for the second place. Each contestant will receive a certificate at this year's ICIQ Conference in Paris. In addition, Dr. Yakout will receive a check for US\$1,000 and Dr. Halimeh and Dr. Zhu each will receive a check for US\$250.

The Committee is expressing its congratulations to all three finalists!

Respectfully submitted:

Rolf Wigand,

**Chair,**

**2012 Ballou-Pazer IQ Dissertation Award Competition Committee**

**2012 STUART ELLIOT MADNICK BEST PAPER AWARD**

The committee of the Stuart Elliot Madnick Best Paper Award competition congratulates the recipients of ICIQ 2012 Stuart Elliot Madnick Best Paper Award (\$1000) chosen from the top-ranked accepted articles of the 17<sup>th</sup> International Conference on Information Quality 2012.

The ICIQ 2012 Stuart Elliot Madnick Best Paper Award for 2012 goes to:

*Key-based Blocking of Duplicates in Entity-Independent Probabilistic Data*

**Fabian Panse, Wolfram Wingerath, Steffen Friedrich, Norbert Ritter**

**University of Hamburg, Germany**

The choice for the award was made by Prof. Jacky Akoka, Brigitte Laboisson, Prof. Isabelle Comyn-Wattiau, Dr Laure Berti-Équille, Dr Monica Scannapieco (ICIQ-2012 Chairs), Prof. John Talburt (2013 ICIQ Chair, USA), Prof. Yang W. Lee (ICIQ Committee Chair), Prof. Wayne Huang (2014 ICIQ Chair, China), Dr Richard Wang (ICIQ General Chair).

Warmest congratulations to the winners!

**ICIQ 2012 Stuart Elliot Madnick Best Paper Award Competition Committee,**

## CONFERENCE ORGANIZATION

### CONFERENCE CHAIRS

Jacky Akoka, CNAM, France

Andy Koronios, University of South Australia, Australia

Brigitte Laboisse, BDQS, France

John Talburt , University of Arkansas at Little Rock, USA

### PROGRAM COMMITTEE CHAIRS

Laure Berti-Équille, IRD, France

Isabelle Comyn-Wattiau, CNAM, France

Monica Scannapieco, Istat, Italy

### DOCTORAL CONSORTIUM CHAIRS

Samira Si-said Cherfi, CNAM, France

Raul Ruggia, University of the Republic of Uruguay, Uruguay

### INDUSTRIAL CHAIR

Sylvaine Nugier, ExQi, France

### PUBLICITY CHAIR

Delphine Clément, Microsoft, France

**PROGRAMME COMMITTEE**

Stuart Ainsworth, UniSA, Australia  
Carlo Batini, University of Milan, Italy  
Laure Berti-Équille, IRD Institute of Research for Development, France  
Mikhaila Burgess, University of Glamorgan  
Ismael Caballero, UCLM, Spain  
Cinzia Cappiello, Politecnico di Milano, Italy  
InduShobha Chengalur-Smith, SUNY at Albany, USA  
Chia-Chu Chiang, University of Arkansas at Little Rock, USA  
Isabelle Comyn-Wattiau, CNAM, France  
Olivier Coppet, SNCD, France  
Tamraparni Dasu, AT&T Labs Research, USA  
Bruce Davidson, Cedars-Sinai Health System, USA  
Claudio di Cicco, Universita di Roma La Sapienza, Italy  
Adir Even, Ben-Gurion University of the Negev, Israel  
Craig Fisher, Marist College, USA  
Zbigniew Gackowski, California State University Stanislaus, USA  
Jing Gao, University of South Australia, Australia  
Marcus Gebauer, Arcor, Germany  
Michael Gertz, Ruprecht-Karls-University Heidelberg, Germany  
Markus Helfert, Dublin City University, Ireland  
Beverly Kahn, Suffolk University, USA  
Barbara Klein, University of Michigan at Dearborn, USA  
Akihisa Kodate, Tsuda College, Japan  
Jochen Kokemüller, Fraunhofer IAO, Germany  
Eitel Lauria, Marist College, USA  
Peggy Leonowich-Graham, USMA, USA  
Jiuyong Li, University of South Australia, Australia  
Helina Melkas, Lappeenranta University of Technology, Finland  
Mariofanna Milanova, University of Arkansas at Little Rock, USA  
Paolo Missier, University of Manchester, United Kingdom  
Mukesh Mohania, IBM, India

Agarwal Nitil, University of Arkansas at Little Rock, USA  
Paulo Jorge Oliveira, Politecnico do Porto, Portugal  
Ajith Parlikad, University of Cambridge, United Kingdom  
Oscar Pastor, Valencia University of Technology, Spain  
Barbara Pernici, Politecnico di Milano, Italy  
Leo Pipino, University of Massachusetts Lowell, USA  
Geert Poels, University of Ghent, Belgium  
Robert Pokorny, XSB Inc., USA  
Srinivasa Ramaswamy, University of Arkansas at Little Rock, USA  
Tom Redman, dataqualitysolutions, USA  
Grant Robinson, New South Wales Office of Water, Australia  
David Rowlands, Direkt Consulting Pty Ltd, Australia  
Laura Rusu, IBM Research Australia, Australia  
Shazia Sadiq, University of Queensland, Australia  
Kai-Uwe Sattler, Ilmenau University of Technology, Germany  
Monica Scannapieco, Italian National Institute of Statistics (Istat), Italy  
Scott Schumacher, IBM, USA  
Valerie Sessions, Charleston Southern University, USA  
Ganesan Shankaranarayanan, Babson College, USA  
John (Skip) Slone, Lockheed Martin Corp., USA  
Besiki Stvilia, Florida State University, USA  
Giri Kumar Tayi, SUNY at Albany, USA  
Mihail E. Tudoreanu, University of Arkansas at Little Rock, USA  
Rolf Wigand, University of Arkansas at Little Rock, USA  
Philip Woodall, University of Cambridge, United Kingdom  
Ningning Wu, University of Arkansas at Little Rock, USA  
C. Lwanga Yonke, Aera Energy LLC, Australia  
Diego Zardetto, Italian National Institute of Statistics (Istat), Italy  
Harry Zhu, Old Dominion University, USA

## CONFERENCE VENUE

ICIQ 2012 will take place in Paris, France. Paris is the capital and the largest city in France. It is situated on the river Seine, in northern France, at the heart of the Ile-de-France region. The city of Paris, within its administrative limits (the 20 arrondissements) largely unchanged since 1860, and is one of the most populated metropolitan areas in Europe. Paris is today one of the world's leading business and cultural centres, and its influences in politics, education, entertainment, media, fashion, science, and the arts all contribute to its status as one of the world's major global cities.

ICIQ 2012 will be held at CNAM Paris.

The CNAM is an institution dedicated to life-long higher education. It is a Public Scientific, Cultural and Professional Institution, classed as a *grand établissement*, among France's top higher education establishments. It is supervised by the French Minister for Higher Education. The Cnam was created in 1794, during the French Revolution, on the location of a medieval monastery, the royal abbey of Saint-Martin des Champs. Nowadays, thanks to its integrated network, the Cnam spreads higher adult education and life-long training, in France and abroad.

How to get there?

The **Cnam** is located at the heart of Paris, close to the Louvre, Notre Dame and the Pompidou Centre.  
Le **Cnam**, 292 rue Saint-Martin - 75003 PARIS, France.

<http://the.cnam.eu/>

Metro stations: Arts-et-Métiers or Réaumur-Sébastopol.



## INVITED TALKS AND KEYNOTES

### Keynote Talk 1: BIG Data Must Overcome BIG Data Quality Challenges

**Professor Stuart MADNICK, John Norris Maguire Professor of Information Technology,  
Sloan School of Management & Professor of Engineering Systems, School of  
Engineering, Massachusetts Institute of Technology**

**16<sup>th</sup> November (Friday), 9:30 AM – 10:30 AM**

**Location:** *Amphi C (Abbé Grégoire)*

In this talk I will describe: (1) the recent excitement and new opportunities about data under the "Big Data" theme, (2) the history of Data Quality research at MIT and elsewhere, and (3) how those two topics intersect. Big Data has rapidly become an extremely important topic in both academia and industry. Some recent examples to be presented include how the granularity and combinations of data now available make new kinds of analysis possible, such as the ability to anticipate (a) what you will buy next or (b) where you will go next. Examples such as these have led to concerns about the usage and privacy of social media and other personal data. Some data quality research issues to be discussed include: (a) the multiple dimensions of data quality, (b) the need for organizational data quality assessment, and (c) the interplay of data quality and data semantics, including data provenance.

As the title of this talk states: ""Big Data Must Overcome Big Data Quality Challenges." This is illustrated by a remark already heard from many Executives: "I now have more and more information, that I know less and less about ..." Since Big Data provides even more data, including personal data, from even more diverse sources, to get true and effective value from Big Data, it must be high quality Big Data. In order to do that, you need to know the quality of the data and the origin (provenance) of the data.

*Professor Stuart Madnick has been on the faculty at MIT since 1972 and served as the head of MIT's Information Technologies Group for more than twenty years. He is the co-author of over 380 books, articles, or technical reports. He co-heads the MIT Total Data Quality Management (TDQM) research program. He has been active in industry as a developer and consultant. He has also been the co-founder of several high-tech firms. Dr. Madnick has degrees in Electrical Engineering (BS and MS), Management (MS), and Computer Science (PhD) from MIT. He has been a Visiting Professor/Scholar at 8 institutions, including Conservatoire National des Arts et Métiers (Paris) and the European Research Consortium for Informatics and Mathematics (Nice.)*



Keynote Talk 2: The five legged sheep: Bouygues Telecom, data quality and governance case study

**Marielle Vo-Van, Customer Insight and Campaign management Director, Bouygues Telecom, France**

**16<sup>th</sup> November (Friday), 3:00 PM – 3:45 PM**

**Location: Amphi C (Abbé Grégoire)**

*The experience feedback of an operator of telecommunications in the implementation of data governance and administration: how Bouygues Telecom had the opportunity to make become aware of the importance of the data during its project of revision of the DWH. Marielle VO-VAN will paint the portrait of data manager and will announce us her best practice.*

*Marielle VO-VAN LIGER, 48, is Bouygues Telecom's Director of Customer Insight and CRM. With more than 20 years of professional experience in Direct Marketing, Marielle has led the design and implementation of the first statistical analysis tools at Bouygues Telecom, in order to better understand the customer's behaviour and needs. A definitive CRM-addict, Marielle leads the design and development of decision-making tools for the operational departments (Marketing, Sales, Customer Service,...), in order to help them increase margins and revenue, and prioritise customer interactions. These tools range from very simple to utterly sophisticated (segmentation, scoring, Customer Life Time Value,...). Through her rich experience, Marielle has acquired many proofs of the high business value of information derived from detailed data, (and most notably customer data), and has been a key advocate for the implementation of an Enterprise Datawarehouse within Bouygues Telecom.*



### Keynote Talk 3: The Quality of Web Data

**Professor Felix NAUMANN, Hasso-Plattner-Institut für Softwaresystemtechnik,  
Germany**

**17<sup>th</sup> November (Saturday), 9:15 AM – 10:15 AM**

**Location: Amphi C (Abbé Grégoire)**

The wealth of freely available, structured information on the Web is constantly growing. Driving domains are public data from and about governments and administrations, scientific data, and data about media, such as articles, books and albums. In addition, general-purpose datasets, such as DBpedia and Freebase from the linked open data community, serve as a focal point for many more data sets. Thus, it is possible to query or integrate data from multiple sources and create new, integrated data sets with added value.

Yet integration is far from simple: It happens at technical level by ingesting data in various formats, at structural level by providing a common ontology and mapping the data source structures to it, and at semantic level by linking multiple records about same real world entities and fusing these representations into a clean and consistent record. The talk highlights the extreme heterogeneity and poor quality of web data and points to methods to overcome them including a multitude of tasks that must be completed: source selection to identify appropriate and high quality sources, data extraction to create structured data, scrubbing to standardize and clean data, entity matching to associate different occurrences of the same entity, and finally data transformation and data fusion to combine all data about an entity in a single, consistent representation.

*Felix Naumann studied mathematics, economy, and computer sciences at the University of Technology in Berlin. After receiving his diploma in joined the graduate school at Humboldt University of Berlin. He completed his PhD. thesis on data quality in 2000. Before moving to the University of Potsdam, he worked at the IBM Almaden Research Center and served as an assistant professor for information integration at the Humboldt-University of Berlin. Since 2006 he holds the chair of Information Systems at the Hasso Plattner Institute (HPI) and is currently on leave at the Qatar Computing Research Institute (QCRI).*



**ICIQ 2012 CONFERENCE PROGRAM**

<b>Thursday November 15, 2012</b>		
	Workshop Data Excellence Paris 2012	
18:30	ICIQ 2012 Welcome Reception	
<b>Friday November 16, 2012 Morning</b>		
8:00-9:00	Registration Morning Refreshments	<i>Location: Salle des Textiles</i>
9:00-9:15	Conference Welcome and Recognitions <b>Jacky AKOKA, CNAM, France</b> <b>Brigitte LABOISSE, EXQI, France</b>	<i>Location: Amphi C (Abbé Grégoire)</i>
9:15-9:30	Conference Program Presentation <b>Laure BERTI-EQUILLE, IRD, France</b> <b>Isabelle COMYN-WATTIAU, CNAM, France</b> <b>Monica SCANNAPIECO, ISTAT, Italy</b>	<i>Location: Amphi C (Abbé Grégoire)</i>
9:30-10:30	Keynote 1 – BIG Data Must Overcome BIG Data Quality Challenges <b>Stuart MADNICK, Professor, MIT, USA</b> <i>Location: Amphi C (Abbé Grégoire)</i>	
10:30-10:45	Break	<i>Location: Salle des Textiles</i>
10:45-12:15	Parallel Sessions	
Room 21.2.31	<b>Session 1 – IQ and Organizations</b> <ul style="list-style-type: none"> <li>• Organizational Issues in Establishing Master Data Management Function, <b>Riikka Vilminko-Heikkilä, Samuli Pekkola</b></li> <li>• The State of Information and Data Quality Efforts in Today's Organizations, <b>Elizabeth Pierce, C. Lwanga Yonke, Piyush Malik, Chitra Kagathur Nargaraj</b></li> <li>• Designing Business Processes Able to Satisfy Data Quality Requirements, <b>Angélica Caro, Alfonso Rodriguez, Cinzia Cappiello, Ismael Caballero</b></li> </ul>	<i>Session Chair: Carlo Batini</i>
Room 21.2.37	<b>Session 2 – IQ and Knowledge</b> <ul style="list-style-type: none"> <li>• Knowledge Acquisition from and Semantic Variability in Schizophrenia Clinical Trial Data, <b>Meredith Nahm</b></li> <li>• Towards Expertise Modelling for Routing Data Cleaning Tasks within a Community of Knowledge Workers, <b>Umair ul Hassan, Sean O'Riain, Edward Curry</b></li> <li>• Domain Knowledge Based Quality for Business Process Models, <b>Sarah Ayad, Samira Si-said Cherfi</b></li> </ul>	<i>Session Chair: Philip Woodall</i>
Room 21.2.44	<b>Session 3 – Information Accuracy</b> <ul style="list-style-type: none"> <li>• APC-SIMULATOR: Demonstrating the Effects of Technical and Semantic Errors in the Accuracy of Hospital Reporting, <b>Sami Laine</b></li> <li>• Assessing Accuracy Degradation over Time with a Markov-Chain Model, <b>Alisa Wechsler, Adir Even</b></li> <li>• Determinants of Accuracy in the Context of Clinical Study Data, <b>Meredith Nahm, Joseph Bonner, Philip L. Reed, Kit Howard</b></li> </ul>	<i>Session Chair: Bruce Davidson</i>
12:15-13:45	Lunch	<i>Location: Salle des Textiles</i>

Friday November 16, 2012 Afternoon	
13:45-14:45	Parallel Sessions
Room 21.2.31	<p><b><i>Industrial Track - Session I</i></b>  <i>Session Chair: Olivier Coppet (GDE France)</i></p> <ul style="list-style-type: none"> <li>• Data Provenance and Financial Systemic Risk, <b>Len Seligman, Shaun Brady, MITRE</b></li> <li>• An Industry Study Case of Data Governance Program in Health Information: the Medtronic MCRI Initiative in Data Management, <b>Marie-Astrid Cartron-Mizeracki, UALR/MEDTRONIC</b></li> </ul>
Room 21.2.37	<p><b><i>Industrial Track - Session II</i></b>  <i>Session Chair: Jean-Michel Derelle (LAFARGE)</i></p> <ul style="list-style-type: none"> <li>• How the Emergence of Open Data Impacts the Data Quality Routines of a Data Service Provider? <b>Soumaya Ben Hassine, AID, Andrea Micheaux, University of Lille 1, Eric Sommervogel, AID</b></li> <li>• Looking back 10 years – Evolution of the Data Management Organization at Microsoft <b>Delphine Clément, Ronan Corre, MICROSOFT</b></li> <li>• The Role of Information Quality Management in Achieving Organizational Performance Excellence: An IQ-Focused Examination of the Baldrige Framework with Examples from the Health Care Industry <b>Bruce Davidson, CEDARS-SINAI Health System</b></li> </ul>
14:45-15:00	Break <i>Location: Salle des Textiles</i>
15:00-15:45	Keynote 2 – The five legged sheep: Bouygues Telecom, data quality and governance case study, <b>Marielle Vo-Van, Customer Insight and Campaign Management Director, Bouygues Telecom, France</b> <i>Location: Amphi C (Abbé Grégoire)</i>
15:45-15:50	IQ Associations' Presentation <i>Location: Amphi C (Abbé Grégoire)</i>
16:00-18:00	Parallel Sessions
Room 21.2.31	<p><b><i>Session 4 – IQ Improvement</i></b>  <i>Session Chair: Ismael Cabellero</i></p> <ul style="list-style-type: none"> <li>• Customized Data Quality Improvement, <b>Philip Woodall, Alexander Borek, Ajith Kumar Parlikad</b></li> <li>• Checking and Repairing the Quality of Information in Databases by Inconsistency Metrics, <b>Hendrik Decker</b></li> <li>• Introducing Data and Information Quality Principles in Today's College Curriculum via an Introductory Probability and Statistics Course, <b>William Rybolt, Leo Pipino</b></li> <li>• Towards the Use of Model Checking for Performing Data Consistency Evaluation and Cleansing, <b>Mario Mezzanzanica, Mirko Cesarini, Fabio Mercurio, Roberto Boselli</b></li> </ul>
Room 21.2.37	<p><b><i>Session 5 – IQ Dimensions</i></b>  <i>Session Chair: Samira Si-said Cherfi</i></p> <ul style="list-style-type: none"> <li>• IQ : Purpose and Dimensions, <b>Phyllis Illari, Luciano Floridi</b></li> <li>• An Investigation into Data Quality Root Cause Analysis, <b>Philip Woodall, Andy Koronios, Jing Gao, Ajith Kumar Parlikad, Elaine George</b></li> <li>• Impact of Conceptual Modeling Approaches on Information Quality: Theory and Empirical Evidence, <b>Roman Lukyanenko, Jeffrey Parsons</b></li> <li>• The Many Faces of Information and their Impact on Information Quality, <b>Carlo Batini, Matteo Palmonari, Giuanluigi Viscusi</b></li> </ul>

## Friday November 16, 2012 Afternoon

16:00-18:00	Parallel Sessions
Room 21.2.44	<p><b>Session 6 – Measurement of IQ</b>  <i>Session Chair: Cinzia Cappiello</i></p> <ul style="list-style-type: none"> <li>The Effect of Missing Data on Classification Quality, <b>Michael Feldman, Adir Even, Yisrael Parmet</b></li> <li><del>Information Quality Assessment in Korean Asset Managing Organization – Using a Product Perspective, Abrar Haider, Snag Hyun Lee</del></li> <li>CALYDAT : A Methodology for Evaluating Data Quality Dimensions based on Data Profiling Techniques, <b>Yonelbys Iznaga, César Guerra, Ismael Caballero</b></li> <li>Key-based Blocking of Duplicates in Entity-Independent Probabilistic Data, <b>Fabian Panse, Wolfram Wingerath, Steffen Friedrich, Norbert Ritter</b></li> </ul>
20:00	Conference Banquet and Awards <i>Location: Restaurant Chez Georges aux Halles</i>

## Saturday November 17, 2012 Morning

9:00-9:15	Morning Refreshments	<i>Location: Salle des Textiles</i>
9:15-10:15	Keynote 3 – The Quality of Web Data, <b>Felix NAUMANN, Professor, Hasso-Plattner-Institut für Softwaresystemtechnik, Germany</b> <i>Location: Amphi C (Abbé Grégoire)</i>	
10:20-10:30	ACM JDIQ Journal Presentation (by L. Raschid)	<i>Location: Amphi C (Abbé Grégoire)</i>
10:30-10:45	Break	<i>Location: Salle des Textiles</i>
10:45-12:15	Parallel Sessions	
Room 21.2.28	<p><b>Session 7 – IQ and Social Media</b>  <i>Session Chair: Andrea Maurino</i></p> <ul style="list-style-type: none"> <li>Research on the Role of Social Media and Motivation to Use in the Local Community – Index of Information Quality and Private Space Function, <b>Yasuhiro Tanaka, Akihisa Kodate</b></li> <li>Quality of Social Media Data and Implications of Social Media for Data Quality, <b>G. Shankaranarayanan, Bala Iyer, Donna Stoddard</b></li> <li>Measuring Information Quality on the Internet – A User Perspective, <b>Olivier Blattmann, Patrick Kaltenrieder, Patrizia Haupt, Thomas Myrach</b></li> </ul>	
Room 21.2.40	<p><b>Industrial Track - Session III</b>  <i>Session Chair: Sylvaine Nugier, Groupe EDF, France</i></p> <ul style="list-style-type: none"> <li>Using Lean to Improve Information Quality, <b>C. Lwanga Yonke, AERA ENERGY LLC</b></li> <li>Towards High-Quality Automative Product Configuration Data Using Meta-Rules, <b>Dirk Zitterell, Ruediger Berndt, AUDI</b></li> <li>Master Data Cleansing for SAP Implementation Project Large Power Generation Company, <b>Reinhard Schiel, PILOG INTERNATIONAL</b></li> <li>RDFREDUCE : Supporting Diverse Knowledge Perspectives on Heterogeneous Data Sources in the Industrial Systems Domain, <b>Mario Pichler, SCCH</b></li> </ul>	
12:15	Official End of ICIQ 2012	<i>Location: Amphi C (Abbé Grégoire)</i>

## MAIN TRACK RESEARCH PAPERS

Session 1 – IQ and Organizations	<i>page</i>
— <b>Riikka Vilminko-Heikkinen, Samuli Pekkola</b> , Organizational Issues in Establishing Master Data Management Function.....	1
— <b>Elizabeth Pierce, C. Lwanga Yonke, Piyush Malik, Chitra Kagathur Nargaraj</b> , The State of Information and Data Quality Efforts in Today's Organizations.....	14
— <b>Angélica Caro, Alfonso Rodriguez, Cinzia Cappiello, Ismael Caballero</b> , Designing Business Processes Able to Satisfy Data Quality Requirements.....	31
Session 2 – IQ and Knowledge	
— <b>Meredith Nahm</b> , Knowledge Acquisition from and Semantic Variability in Schizophrenia Clinical Trial Data.....	46
— <b>Umair ul Hassan, Sean O'Riain, Edward Curry</b> , Towards Expertise Modelling for Routing Data Cleaning Tasks within a Community of Knowledge Workers.....	58
— <b>Sarah Ayad, Samira Si-said Cherfi</b> , Domain Knowledge Based Quality for Business Process Models.....	70
Session 3 – Information Accuracy	
— <b>Sami Laine</b> , APC-SIMULATOR: Demonstrating the Effects of Technical and Semantic Errors in the Accuracy of Hospital Reporting.....	85
— <b>Alisa Wechsler, Adir Even</b> , Assessing Accuracy Degradation over Time with a Markov-Chain Model.....	99
— <b>Meredith Nahm, Joseph Bonner, Philip L. Reed, Kit Howard</b> , Determinants of Accuracy in the Context of Clinical Study Data.....	111
Parallel Sessions: Session 4 – IQ Improvement	
— <b>Philip Woodall, Alexander Borek, Ajith Kumar Parlikad</b> , Customized Data Quality Improvement.....	126
— <b>Hendrik Decker</b> , Checking and Repairing the Quality of Information in Databases by Inconsistency Metrics.....	139
— <b>William Rybolt, Leo Pipino</b> , Introducing Data and Information Quality Principles in Today's College Curriculum via an Introductory Probability and Statistics Course.....	151
— <b>Mario Mezzanzanica, Mirko Cesarini, Fabio Mercurio, Roberto Boselli</b> , Towards the Use of Model Checking for Performing Data Consistency Evaluation and Cleansing.....	163

Parallel Sessions: Session 5 – IQ Dimensions

— <i>Phyllis Illari, Luciano Floridi</i> , IQ : Purpose and Dimensions .....	178
— <i>Philip Woodall, Andy Koronios, Jing Gao, Ajith Kumar Parlikad, Elaine George</i> , An Investigation into Data Quality Root Cause Analysis.....	193
— <i>Roman Lukyanenko, Jeffrey Parsons</i> , Impact of Conceptual Modeling Approaches on Information Quality: Theory and Empirical Evidence.....	206
— <i>Carlo Batini, Matteo Palmonari, Giuanluigi Viscusi</i> , The Many Faces of Information and their Impact on Information Quality.....	212

Parallel Sessions: Session 6 – Measurement of IQ

— <i>Michael Feldman, Adir Even, Yisrael Parmet</i> , The Effect of Missing Data on Classification Quality.....	229
— <i>Abrar Haider, Snag Hyun Lee</i> , <del>Information Quality Assessment in Korean Asset Managing Organization – Using a Product Perspective.</del> .....	243
— <i>Yonelby Iznaga, César Guerra, Ismael Caballero</i> , CALYDAT: A Methodology for Evaluating Data Quality Dimensions based on Data Profiling Techniques.....	260
— <i>Fabian Panse, Wolfram Wingerath, Steffen Friedrich, Norbert Ritter</i> , Key-based Blocking of Duplicates in Entity-Independent Probabilistic Data.....	278

Parallel Sessions: Session 7 – IQ and Social Media

— <i>Yasuhiro Tanaka, Akihisa Kodate</i> , Research on the Role of Social Media and Motivation to Use in the Local Community – Index of Information Quality and Private Space Function.....	297
— <i>Ganesan Shankaranarayanan, Bala Iyer, Donna Stoddard</i> , Quality of Social Media Data and Implications of Social Media for Data Quality.....	311
— <i>Olivier Blattmann, Patrick Kaltenrieder, Patrizia Haupt, Thomas Myrach</i> , Measuring Information Quality on the Internet – A User Perspective.....	326

# **ORGANIZATIONAL ISSUES IN ESTABLISHING MASTER DATA MANAGEMENT FUNCTION**

(Research Paper)

**Riikka Vilminko-Heikkinen**

Department of Information Management and Logistics  
Tampere University of Technology, Finland  
[riikka.vilminko-heikkinen@tut.fi](mailto:riikka.vilminko-heikkinen@tut.fi)

**Samuli Pekkola**

Department of Information Management and Logistics  
Tampere University of Technology, Finland  
[samuli.pekkola@tut.fi](mailto:samuli.pekkola@tut.fi)

**Abstract:** Master data management (MDM) provides an access to the consistent views of the organization's most important data, also referred to as master data. In addition to technical issues, there are many organizational items related to MDM and its organizational implementation. However, current academic literature lacks empirical studies on organizational challenges influencing the MDM initiatives. Consequently organizational issues in establishing master data management function in an organization are studied in this paper. Data collection is conducted by participatory observations of a year-long MDM project. Reflecting the findings to the literature shows that several new issues have emerged. These indicate that the implementation of MDM is also affected by the organization's ability to identify data owners and associate them with appropriate roles and responsibilities, and to create a unified understanding of the key terms and concepts regarding MDM. Also the importance of communication is emphasized.

**Key Words:** Master data management, MDM, organizational issues, organizational implementation, data quality, qualitative research

## **INTRODUCTION**

Data has been developed in silos over the years. This and the fact that the amount of data has increased rapidly, have caused the data to be stored in numerous information systems (IS) and databases. It is also common that multiple information systems hold the same or nearly the same data [16]. Disparate systems and applications create segregated information. This results in duplicate, incomplete and inaccurate data that leads to inappropriate analytics and, at the end, inaccurate business decisions [25]. Problems with data quality and reliability have thus emerged. These problems create additional costs for organizations and make it problematic for them to use the data [20]. The quality of transactional and inventory data depends directly on the quality of master data [15]. Another angle on the subject is that still 40 % of organizations are unaware of the problems with their data [29].

In order to cope with several data siloes and vast amounts of data quality problems, data is often organized according to its business criticality. To manage business critical data, a new concept, master data as the organization's core data that forms the basis for business processes [19] has been introduced. Its typical characteristics are stability [26], reuse [5] and high value for the organization [17]. Common examples of master data are customers, products, and vendors.

Loshin [17] describes master data management (MDM) as a collection of data management practices that are orchestrated by key stakeholders, participants, and business clients. They utilize business applications, information management methods, and data management tools to implement policies, services, and infrastructures to support the capturing, integrating, and sharing accurate, timely, consistent, and complete master data. MDM aims at supporting the organization's functions by providing an access to consis-

tent views of uniquely identifiable master data entities across the operational application infrastructures [17]. MDM is consequently a method, or an ensemble of methods to, target fragmented data that is stored in various data databases and siloes in the organization [27]. Therefore, MDM contributes to maintaining information quality [18].

MDM is often conceived as a technical term, even though the literature states its challenges are mostly concerned with people in the organizations [1]. For example cultural impedance is creating difficulties [1]. Yet in general the literature on non-technical issues is scarce [36]. This study thus opportunistically focuses on organizational issues that a MDM initiative may face.

In this study, we aim at identifying organizational obstacles and issues that an organization may encounter when establishing its MDM function. Consequently we supplement current literature. The data for the case study is collected through an ethnographical study within a year-long case project.

Before going to the case and its description, the challenges from the prior research are identified. Then the case study settings and our findings from the case are presented. The paper ends with discussion and concluding chapters.

## **RELATED RESEARCH AND THEORETICAL BACKGROUND**

Introducing and further establishing MDM into an organization is a complex process with numerous steps and viewpoints [17]. With this initiative, many issues, that may even conflict, emerge along the way. Earlier technical issues have been identified and studied (c.f. [36]). Those include choosing and creating MDM solutions that would take into account the organizational demands [2], and challenges that appear in the context of complex enterprise resource planning landscapes [21, 30]. Also different MDM architectural design challenges have been identified (e.g. [21, 8]). Although the studies have, by large extent, emphasized technical aspects, they have also touched some decisive organizational issues in introducing MDM.

Generally speaking, the literature on MDM is scarce. MDM has mainly been seen as a technical concept [31]. Although apparently there is a lack of academic research, there are many industry experts that have contemplated the subject from many angles. Both academic and practitioner-oriented literature imply that simple treatments of MDM just as a technical concept is one of the reasons why the projects fail, and why MDM has not delivered expected results (e.g. [28, 23]). From the technical perspective, a successful MDM project can be well implemented but still not being able to fulfil the business objectives. Under the circumstances Andriole [3] describes MDM as being partly technology, partly governance, and partly philosophy, not just as technology.

Identifying a primary business owner for data item has been identified as one of the key issues when implementing MDM [33]. This also means that stakeholders must be involved in the MDM initiatives [33]. However, often the definitions for data ownership are inadequate or completely missing. The challenge emerges when the data ownership is not emphasized in the organization's culture [32]. Data ownership can easily be regarded to as IT unit's task as the data is associated with certain information systems and its databases. Yet the owner has to be found from the business processes. He/she has to understand the responsibilities the role brings. Unclear data ownership can cause, for example, inadequate process definitions, making data maintenance very difficult or even impossible [32].

Fung-A-Fat [9] argues that when identifying data owners, some surprises can emerge in the organization. Those may quickly reveal some confusing and contradictory processes and interactions. Moss [23] has listed several decisions that the data owners should make for their data. Those are related to the domains

and valid values, data availability and accessibility and of their timescales and actors, security policies, and the frequency of updates.

MDM has been identified as an initiative that involves different processes and functions of the organization. Radcliffe [28] underlines strong alignment with the organization's business vision and MDM initiatives. It is thus essential to have multidisciplinary teams, i.e. participants from all business lines, when implementing MDM [12]. This emphasizes a need for high level coordination to control the involved parties [19]. Yet it might be difficult to find a coordinator as he/she needs to be neural pacesetter that steers the initiative and considers the organizations' different viewpoints as well as ensures that MDM supports the business. Because MDM crosses over organizational boundaries, it might become very difficult to collaborate between different business operations, functions and departments.

Because of the novelty of MDM concept, MDM terminology is not shared [27, 23, 9]. The absence of commonly agreed terms becomes an issue when, e.g. mutual understanding of the terms of customer or product, are missing [27, 4]. This may lead to situations where data sets with ambiguous definitions cannot be comprehended from the MDM perspective.

The role of management and their commitment has been recognized as a key issue when establishing MDM. The challenge is to convince the management about MDM and to get their support [17]. This emphasizes the importance of executive sponsorship [7]. Executive sponsorship is also needed for ensuring the resources for the initiative [31]. Yet executive or general management support alone is not sufficient. Taking the initiative forward requires commitment from the whole organization. For example collaboration with the broad spectrum of business and IT people across the organization is important. This includes, e.g., CIO and IT staff, business owners, data integrators, application developers, as well as executive sponsorship [8].

The management's lack of commitment is a result of the limited understanding of the data quality problems [34, 33]. As MDM is a very challenging concept, it is hard to detach it from general data management practices [32]. Yet caring the data and its quality should be considered as important business activities [14]. This necessitates a shared understanding of master data as a common asset [8]. The management should thus ensure that the importance of the relationship between business processes and data is evident to each and every party [14].

Business needs set requirements for governing the master data and its availability, usability, integrity, and security [31]. Yet those responsibilities are rarely defined when starting the MDM project [9]. This again underlines the importance of mutually shared understanding and responsibility of both MDM and master data within the whole organization [14].

Altogether, marketing MDM initiatives inside the organization is seen difficult. Almost all activities involve the use of data [11]. Yet MDM is not just data. It also involves the management, process owners, and those who enter the data into information systems. Recurrent communication is consequently important. Many MDM initiatives fail because the expectations are not communicated nor understood. This decreases motivation and results the lack of interest and commitment towards the initiative [16].

Problems with responsibilities are barriers for MDM [11, 28]. MDM often requires changes, such as new practices, disciplines, methods, roles, responsibilities, policies, and procedures [23, 34], in the organization and its operations. Finding appropriate data governance roles is essential [35]. This becomes particularly problematic if explicit data governance roles have not been set. Organizations need to define data governance policies and procedures to oversee MDM processes [17, 5]. Yet it is hard to evaluate the

organization's preparedness for MDM [17]. MDM specific maturity models to assess this do not exist.

Table 1 summarizes the organizational issues in establishing MDM in organizations.

<b>ISSUE</b>	<b>REFERENCE</b>
Communicating the idea of MDM	Lee et al. [16]
Data owners	Fung-A-Fat [9], Smith and McKeen [33], Silvola et al. [32]
Engaging people in the project	Shankar [31], Dreibelbis et al.[8]
Lack of high-level coordination	Loser, Legner and Gizanis [19]
Management support	Loshin [17], Snow [34]
Organizational changes	Berson and Dubov [5], Loshin [17], McKnight [22]
Organizational responsibilities	Fung-A-Fat [9], Silvola et al. [32], Radcliffe [28], Haug and Arlbjørn [11]
Unified Terms and Concepts	Moss [23], Fung-A-Fat [9], Poolet [27]

**Table 1. Organizational issues identified in the literature**

## **RESEARCH METHOD**

The subject for the study is a public sector organization with approximately 16 000 employees. Municipality's services are produced using a multiple provider model. This means that external companies and communities provide services alongside the city's own service provision. The operational model separates the service purchaser from the provider. The organization consists of central administration, purchasing unit, welfare services, municipal corporations and several subsidiaries. The MDM project was mainly conducted in the central administration and its IT unit.

Motivation for starting the MDM project was seen already in 2008 when problems with data that was considered of being organization's important core data, were dispersed. Clear data quality problems were indicated. At first, the most obvious problems concerned data duplicates and issues in maintaining the data access. Master data management was considered a solution that would solve the problems comprehensively as they were perceived to origin from the maintaining processes and several applications.

Already in 2008, the business objectives for MDM were identified for the first time. These included enabling more effective work by streamlining work processes and the organization, improving reporting and achieving better interoperability with service-oriented architecture (SOA). Also some MDM objectives were identified. These were to provide processes for data collection, integration, consolidation, quality assurance, and distribution to ensure data integrity, maintenance, and application of information usage control. This set the original goal for the MDM project: to discover what was the organization's master data and how it should be considered in their MDM development, and to plan how the development should proceed. The project excluded technical solutions and the implementation of MDM.

Both the study and MDM project started in November 2010 and ended in October 2011. Overall duration was thus 12 months. The first four months were devoted to the procurement phase, followed by the actual MDM phase. The project organization included three different groups: project group, steering group and expert group. Altogether 33 persons were involved. They represented organizations' different functions, e.g., IT, human resources, business, and procurement and all the core processes. Few experts were from municipal corporations and two vendors acted as a consulting party. IT unit, where the first author was employed as project coordinator, was responsible for the implementation of the project.

The study is based on ethnographic research, which aims at understanding human activities in a particular environment and context. Data collection was done by participating in project group meetings, steer-

ing group meetings, kick-off and closing seminars, and other project-related meetings and informal discussions. The first author was actively involved in the project as member of the steering group and as a member of the expert group. The situation offered a unique opportunity to observe and understand the project while also participating in it. Ethnographical observations were recorded to personal diaries and notes. The first author made entries to her diary at least weekly, usually daily, whenever she encountered issues that were related to MDM or its implementation. In addition to ethnographical data, also project documentations such as procurement documentation, project plan, monthly status reports, different memos (working group, steering group, project portfolio group, stakeholder groups, kick-off and closing seminars) were used.

The data analysis was conducted by adopting the principles of grounded theory as an analysis tool. First the researcher familiarized herself with the data. The goal was to gain an impression of the material. After a time being, the focus on organizational issues of MDM emerged. After this individual themes were identified and gathered from the data. This allowed classifications of similar issues being expressed in various ways.

Ethnography is never neutral. The role of the researcher thus affects the final results [6]. Excessive subjectivity is avoided by giving detailed descriptions of the subject. The researcher is responsible for analysing and interpreting the results [6]. Even though the first author made systematic entries to the diaries and annotations to the documents throughout the project, all materials were analysed "at once", at the end of the project. This means that the first 11 months can be referred to as a data collection period where entries related to MDM were made. They were not limited or affected by the analysis of earlier entries. This was done to minimize the unintended manipulation of the entries as one may easily make subconscious decisions what to record. The analysis of the data can thus be regarded to as content analysis, where an external researcher makes his or her own interpretations of the phenomena. However, as the researcher had also collected the data and "lived with the tribe", she was able to complement and interpret it in the organizational context. This made it easier to understand the organization culture and social structures and their impacts, and to theorize the subject more richly and in more complex ways [13].

## FINDINGS

The analysis of empirical data revealed 12 factors as organizational issues influencing the implementation of the MDM.

### ***MDM and related concepts***

The business people linked MDM to the organization's attempt to refine knowledge management. In the last few years, the importance of knowledge management had been brought up by the business. Still the vision was not clear: "*Managing knowledge is a concept that has not been defined*" (Closing seminar 1.11.2011). This had an implication in identifying the vision for MDM and separating it from knowledge management. The issue of related concepts is thus evident. The first step is thus to clarify what MDM is, and, as entered in the researcher's diary (9.5.2011): "*[It is] important also to discuss what MDM is not*".

### ***Consensus about the objectives***

Confusion about the term MDM and how it relates to similar concepts also resulted as difficulties in unified understanding of the objectives. There were many different kinds of expectations towards the MDM project. The purchasing unit saw the initiative as an enabler for a larger process development work while business people perceived it as a solution for reporting, being more related to data warehousing. For example "*they (business unit) see the project as an enabler for data warehouse more than anything*" (Diary 14.4.2011). Many parties were also expecting quick technical solutions being implemented during the project. Contrary to these expectations, the project focused on establishing MDM and its practices, and

included only a brief preliminary study on technical solutions. Generally speaking MDM was very strongly perceived merely as a technical solution.

### ***Identifying the needed parties***

At the beginning of the project, identifying relevant business functions and processes, and naming participants to expert groups was considered as a challenge. Especially the level of expertise was difficult to distinguish and articulate. The problem can be conceptualized in a fact that participants needed to be positions where they knew enough about their business processes and functions, and about those information needs and usages.

### ***Engaging organization to the project***

MDM implementation project was generally seen as an IT-project. This made it difficult to encourage and engage participants across the whole organization at the beginning of the project. This was particularly a challenge with people in charge of different business processes: “...*They [participants from business units] don't understand what their role would be in an IT project*” (Diary 3.3.2011). The participants doubted if they had the expertise and ability to contribute to the project. This lowered their motivation and the level of participation.

### ***Roles and responsibilities***

Identifying appropriate roles and finding people to these roles were seen as important factors in establishing MDM: “*There are many different solutions for managing it and, therefore, the know-how is dispersed. Information management processes are not defined, and everything is done now in a decentralized manner. [There are] Ambiguous situations concerning the use [of data] and the decision making*” (Project group memo 12.5.2011). In general, it was seen important that the people are made accountable for the data quality. Yet the concern was that the responsibilities would then be handed to people without studying their workload and available resources, adding the MDM responsibilities as extra task. This prevented the initiative to be put into action. It also highlights that the MDM tasks and responsibilities were seen as an extra function.

Also switching the responsibilities from one person to another was seen problematic. When new tasks were planned it was noticed that people historically in charge of the task and activity would not be allowed to manage the data anymore. This was seen as a power issue, reflecting negative connotations to the MDM initiative.

### ***Unified terms and concepts***

The lack of unified terms was clearly an issue. The key terms “customer”, “product” and “service” were not defined. Consequently different participants had different connotations of what those terms meant. For example, for the term “customer” the units had their own definitions: “*When there isn't a shared understanding of master data, the dreams of "knowledge management" can easily be buried. The concept "service" in Process 1 is defined differently than in Process 2, and the definition for the concept "client" is different in System 1 than in System 2, making the aggregation of their data sets almost impossible*” (Presentation to the executives 17.8.2011). Unified definitions, which would cover and be used in the whole organizations, did not exist. Even though the issue was discussed regularly in the project group meetings and in the steering group meetings, a solution could not be agreed upon. This also implies that the terms should be clearly defined before continuing with these data types.

### ***The level of granularity for defining data sets***

The identification of master data sets also necessitates decisions about their level of granularity: “*Too*

*high level of practical applicability is nil. For example, "Human" is too broad [data set]. There is no "human" master data maintenance process. When taken too low, the field is diffused. Thus the wisdom relies somewhere in the middle, but it is a very thin line*" (Project steering group 24.5.2011). There the term "human data set" includes all kinds of humans e.g., customers, employees and patients. Yet their management and attributes are very different, currently distributed across the organization and its numerous processes. Due to these reasons, the appropriate level of granularity was difficult to find.

### ***MDM concept owner***

Ownership issues had various impacts on establishing MDM. In addition to data ownership issues, the MDM concept ownership was seen as a challenge. This role was seen as the responsible party for the whole MDM concept and is also accountable for developing the area.

From the beginning, there was no clear place for the MDM concept ownership: "*Challenges with master data are related to the responsibility and ownership of the data management concept: the core data bridging the processes, systems and organizational boundaries, there isn't an obvious home for them in the organization... such liability does not arise, for example, from the data warehouse project: it focuses on the existing assembly, not on existing infrastructures. Neither it challenges the construction. Similarly, enterprise architecture won't be able to solve the information content and process-related problems...*" (Presentation for the executives 17.8.2011). Few units in the organization were proposed and deliberated to act as MDM concept owners. Nevertheless, there was a common understanding about a need for neutral concept owner: "*Management model should be owned by a neutral party, not by the purchaser nor the producer*" (Project group memo 12.5.2011). The concern was finding a party that would look at the issues so that the whole organization is considered, not just its segments.

Generally, the concern for who would first adopt and then own MDM was about resources and capabilities. This role was regarded as very significant. The fear was that the chosen owner would not get appropriate resources. This would harm the future plans of the whole organization. The role of MDM concept owner included both a sponsor from the management and an operational leader that would actively take the initiative forward after the project. Management level sponsor was very difficult to identify as there were no obvious candidates due to the organizational structure and the uncleanness of the desired level of management.

### ***Data ownership***

In addition to concept ownership, also data ownerships related challenges were observed. Data ownership involves the responsibility of developing and maintaining a single data set. Process owners and the owners of different master data sets were discussed: "*[Data ownerships] should be clearly and unanimously defined and their responsibilities set ...*" (Closing seminar 1.11.2011). Setting these ownerships was considered very difficult, and ownerships of only a few data sets were clarified during the project. As the IT unit had the ownership of the major IS, it was suggested that they should also own the data: "*Ownership is a difficult concept because it easily gives an idea that information system ownership also refers to the ownership of data contents and process ownership...*" (Notes 15.8.2011). Master data was thus perceived to be bound with information systems. Yet no consensus was achieved despite of several discussions taking place both in project group and steering group meetings, and in the closing seminar.

### ***Organizational changes***

Over the years the organization had developed several practices for updating the data or creating new data entries. These activities were done differently by different business units and functions. The information systems administrators updated the data by the requests from different business units. Yet no explicit process was defined. This resulted many problems. For example it was not clear who could establish a new location or site, what information would be needed there and what would be its right format.

Thus the data about locations and sites was not accurate or uniform: “*We cannot manage the key pieces of knowledge by our current practices...*” (Presentation to the executives 17.8.2011).

As this was a customary way of working, it was very difficult to change. Although the persons involved with the MDM project were ready to change their practices, there was a lot of debate that establishing the changes into the organization would be very difficult. However, it was agreed that the change is needed: “[maintenance] processes and their follow-up should be a part of everyday activities” (Closing seminar 1.11.2011).

### **Communication**

The MDM concept was ambiguous to the organization. Different connotations originated from inconsistent definitions both in the literature and in the practices. This made communication and marketing MDM very difficult, especially at the beginning of the project when incorrect interpretations had to be first discarded, and because the non-existence of the unified definitions to replace them. This was severe as MDM was needed to be communicated widely across the organization. The people involved in the project felt that they had to justify the importance of data quality to their management and other stakeholders. The level of abstractions in the messages was seen important, but it turned out to be difficult in practice. For example: “*Communicating MDM to the executives should be very concrete. How that could be done?*” (Diary 15.8.2011). Different ways of communicating were discussed and argued. One effective way was the use of narratives. Also tailoring the message according to its recipients was challenging because of the heterogeneity of the employees. All what was wanted was to provide a basic understanding of MDM to the whole organization.

Communication within the project group was also important. People using the data on a daily basis are vital in achieving desired results of the MDM initiatives: ”...With spatial data, the main problem is that the people do not communicate. In other words, people [managing the data] need to tell what information is available [to data users]...” (Notes 5.7.2011). It is indeed important that people in the organization feel that their needs are considered or they will not support the MDM initiative [29].

### **Legislation driven challenges**

As being a public administration organization, legislation had its impact also on the MDM initiative. During the research project, a new law concerning information management in public administrations, Act on Information Management Governance in Public Administration, came into force. This obliged the organization into certain measures, e.g., with their information architecture. This fine-tuned the MDM project objectives abruptly a little as then “*this project attempts to make the changes that the Act [...] obligates us to do...*” (Notes 5.7.2011).

## **DISCUSSION**

Many of the issues identified in the literature were also identified from our case study. However, five new issues were discovered. A summary of the issues is presented in Table 2.

<b>ISSUE</b>	<b>ONLY IDENTIFIED IN THE LITERATURE</b>
Lack of management support	Prior research listed the lack of management support as a challenge. The case study did not explicitly emphasize this even though the issue was recognized as important.

<b>ISSUE</b>	<b>IDENTIFIED BOTH IN THE LITERATURE AND CASE STUDY</b>
Communication and marketing <ul style="list-style-type: none"> <li>- to management</li> <li>- to data owners</li> <li>- to data administrations</li> <li>- to general communication across the organization</li> </ul>	The importance was recognized both in the literature and in the case study. The target groups were also identical.
Data owners	Identifying the data owners was noted as one of the most crucial challenges.
Engaging organization to the project	It was seen difficult to engage people and business units to the project. Literature argued this being mainly related to the engagement of the idea of MDM, while the case study brought in also an issue concerning the commitment to the actual project.
Organizational changes	This was seen as a greater challenge in the case study than in the literature.
Responsibilities and roles <ul style="list-style-type: none"> <li>- New responsibilities</li> <li>- Changes in responsibilities</li> </ul>	This is linked with the identification of the data owners. It also plays an important part in establishing MDM.
Unified terms and concepts	A common understanding of terms and concepts is a major issue in identifying and managing master data. This was considered as one of the most fundamental factors.
MDM concept owner <ul style="list-style-type: none"> <li>- sponsor from the management</li> <li>- operational leader</li> </ul>	Literature identified the need for a high level coordination in order to control the parties involved. This was also recognized in the case study, and specified as a need for MDM concept owner. The need for operational leader was mentioned in the case study even though it is not evident in the literature.
<b>ISSUE</b>	<b>ONLY IDENTIFIED IN THE CASE STUDY</b>
Related concepts	Knowledge management was a topic in the organization. Yet it was found difficult to distinguish it from MDM because of deficiencies in the definition.
Consensus about the objectives	There was no commonly agreed consensus about expectations from MDM among the different business units in the organization.
Identifying essential parties	Prior research acknowledges that many different processes and functions of the organization are involved in MDM. However, it does not point out the challenges in identifying those parties.
Legislation driven challenges	Legislation was identified as an organizational issue as the case organization had to obey legal issues concerning its functions and information contents.
The level of granularity for defining master data sets	A unified level of granularity for the data sets has to be accomplished, because large amount of master data in the organization.

**Table 2 Summary of identified issues**

The lack of unified terms and concepts was identified both in the literature and in the case study. To help communication, some common examples (customer, product, service) were used. However, they were comprehended differently making the communication difficult. This can explain many of the problems of the MDM initiative. This finding of undefined concepts parallels with [4] study on another new concept –

service-oriented architectures. Some other issues from the literature were also confirmed by the case study. These include, for instance, the challenges with responsibilities and roles, identifying data owners, difficulties in engaging people and organizations in the project, and preparing and adapting to the organizational changes.

The case study emphasized the importance to cope with organizational change much more than the literature. The type and size of the case organizations and its management may have an explanation. The case organization is a large public sector organization, where the employees are hired for very long periods of time. Therefore some of the practices and efforts have become customary and personified. Under the circumstances all attempts to change the situation can easily be perceived as negative. This makes it difficult to define and implement new responsibilities and roles for MDM. This is emphasized especially when the issue or its terms are not understood, or when the new responsibilities are seen as extra work and not as activities to improve processes and data quality. The case organization had many of its functions in silos, delimiting the development of a common culture, shared by the whole organization. Organization-wide processes for ensuring the data quality will be difficult to achieve in this kind of situations.

The issues related to legislation were a challenge to the case organization, even if the literature did not identify them. This can be explained by the type of the case organization, being driven and guided by the laws, acts and other forms of legislation much more than an average enterprise. Earlier research has focused more on a private sector.

Consensus about the objectives was a great challenge. MDM was usually seen as an enabler for data warehouse, and nothing more, by the business units. Also, as the case study was conducted in a longitudinal manner from the beginning of the project to its end, objectives were more of an issue at the beginning. This differs from the literature where the consensus of the objectives has already been achieved and when the term are, at least to some extent, less unambiguous for the organization.

Another issue from the beginning of the MDM project was the identification of the parties needed to be involved in the MDM initiatives. Engaging people and getting them committed in the MDM implementation was difficult. As MDM initiatives should comprise different processes and functions of the organization, identifying the parties from all related areas and business units was a great challenge. This can be two reasons for this: the lack of identified real data owners, and a narrow understanding of MDM. The size of the organizations may also have had its impact, even though we believe the real reason is the large number of seemingly similar master data sets that are different in details. New master data sets actually emerged and were identified during the project, as their existence was not known at the beginning of the project when parties got together.

Large amount of master data in the organization could also provide an explanation for the problems with the levels of granularity. A unified and reasonable level of granularity had to be set in order to keep the master data models manageable.

The lack of high level coordination was identified as a challenge in the literature. This issue also came up in the case study, but more as a challenge for the MDM concept owner. During the project, IT unit acted as a high level coordinator. This was beneficial as the unit was considered as a neutral party, neither data nor process owner. However, because MDM is strongly associated with IS and technologies, it remains to be seen whether they are actually “neutral enough” in a long run.

The lack of management support was not perceived as a challenge in the case study. This might be due to its careful consideration both before and throughout the project. This is evident from early stage documentation: “*Creating a management model is basically about change management where the management's commitment is exceedingly important*” (Project plan 23.2.2001). This careful preparation can be a reason why it was not seen as a challenge. Management support was ensured even before starting the introduction of MDM to the organization. Management is also one of the four target groups of the MDM initiative. This was found from both literature and case study, where it was also seen important to educate them about the impacts of MDM, for instance impacts to the data quality. It was also evident that the management wanted to hear more about the impacts of MDM instead of the MDM activities itself.

The issue of management support is closely related to communication. Communication, and particularly its absence, was identified in the literature and in the case study. In addition to communication to the management, also communication to the whole organization was considered essential because almost all functions and processes use data.

This kind research had its limitations. First, the study was done in one organization. Thus, even though our list of issues complements the literature, it might not be complete. Also, we do not claim that the list of issues is prioritized – even though some issues seemed to be more important than the others – but that might be case specific – or not. Second, ethnographical study surely has an impact on the issues identified as some might be emphasized by the researcher’s personal interests. We have tried to minimize this by separating the data collection and analysis phases, and by relying also on other materials. Ethnography provides unique opportunities to understand profound reasons and causes, not just superficial and most obvious findings. Taking these criticisms into account, we still believe the list of issues identified provides a fruitful starting point for the future research – in other types of organizations and by a variety of research methods.

## CONCLUSIONS

In this paper, we gained new understanding of the challenges in establishing MDM function in an organization. Although organizational issues are considered as key factors in succeeding in a MDM initiative, still only limited research has been done to identify them. Through qualitative case study and ethnographical observations from one organization, 12 issues were found and compared to the prior research. Several issues, such as communicating the essence of MDM for different groups, established common terms and concepts, committing people in the initiative, preparing for organizational changes, needing high level coordination, setting organizational responsibilities and roles, and missing data owners were verified.

Several new issues were found from the case study. These were: accomplishing mutual understanding of the objectives, identifying the needed entities that should be involved in the MDM initiative, defining the level of granularity for defining organizations’ master data sets, the problems with related concepts, and considering legislation driven challenges. The case study also emphasized some issues more than the literature. For example unidentified data owners popped up through the MDM project and were seen as critical issues for the project progression. Also common terms and concepts and clear responsibilities and roles were underlined. These three issues were recognized compulsory and inevitable for a successful implementation of MDM – at least in our case. Yet it remains to be seen whether they are as important in the other settings.

It seems that the organizational features, environment, and context have an effect on the encountered challenges. Organization’s maturity on, e.g., knowledge management can actuate to encountered issues.

Our organization is a large public sector organization that has business functions in many different areas. This is also the reason for the organization's multiple and siloed master data sets. The legislation driven issues were clearly due to the fact that we were dealing with the public sector. Nevertheless, with this exception it seems that discovered issues were not bound to the public sector.

The issues discovered in the research shed light on the complexity of MDM. Organizational issues of MDM have not been studied earlier. Consequently our results may assist the researchers in their endeavour in understanding the organizational aspects in MDM, and in building theoretical models, frameworks, practices, and explanations. These results are also useful for professionals both in public administrations and in the enterprises when they are planning to introduce MDM, or if their projects are already progressing. Hence, the list of organizational issues provides a skeleton for future work even though beef around the bones is desperately needed.

## REFERENCES

- [1] Ambler, S.W. Agile Strategies for Master Data Management. *Cutter IT journal*, 20 (9). 2007. pp. 18-22.
- [2] Andreescu, A., Mircea, M. Combining Actual Trends in Software Systems for Business Management. International conference on computer systems and technologies. 2008. v.9-1--v.9.6.
- [3] Andriole, S. J. *Technology Due Diligence: Best Practices for Chief Information Officers, Venture Capitalists, and Technology Vendors*. Indiana University Press, Bloomington. 2008.
- [4] Antikainen, J., Pekkola, S. Factors Influencing The Alignment of SOA Development With Business Objectives. In Proceedings of the 17th European Conference on Information Systems (ECIS'09). 2009.
- [5] Berson, A., Dubov, L. *Master Data Management and Customer Data Integration for a Global Enterprise*. McGraw-Hill, New York. 2007.
- [6] Blomberg, J., Giacomi, J., Mosher, A., Swenton-Wall, P. *Ethnographic Field Methods and Their Relation to Design*. In Schuler, D. & Namioka, A. (Eds.) *Participatory Design: Principles and Practices*, Erlbaum, Hillsdale, NJ 1993. pp.123–155.
- [7] Dayton, M. Strategic MDM: The Foundation of Enterprise Performance Management. *Cutter IT Journal*, 20 (9). 2007. pp. 13-17.
- [8] Dreibelbis, A., Hechler, E., Milman, I., Oberhofer, M., van Run, P., Wolfson, D. *Enterprise Master Data Management: An SOA Approach to Managing Core Information*. IBM Press, Boston. 2008.
- [9] Fung-A-Fat, M. Why Is Consistency Is Inconsistent? The Problem of Master Data Management. *Cutter IT journal*, 20 (9). 2007. pp. 23-29.
- [10] Germain, C.P. *Ethnography: The Method*. In Munhall, P. & Boyd, C. (Eds.) *Nursing Research: A Qualitative Perspective*. National league for Nursing, New York, 1993. pp. 237–267.
- [11] Haug, A., Arljbjørn, J.S. Barriers to Master Data Quality. *Journal of Enterprise Information Management*, 24 (3), 2011. pp. 288–303.
- [12] Joshi, A. MDM Governance: A Unified Team Approach. *Cutter IT journal*, 20 (9), 2007. pp.30-35.
- [13] Kemmis, S. and McTaggart, R. *Participatory Action Research: Communicative Action and Public Sphere*. In N. Denzin and Y. Lincoln, ed 2005. *Handbook of Qualitative Research*. 3rd ed. Thousand Oaks: Sage, pp. 559-604.
- [14] Knolmayer, G. and Röthlin, M. Quality of Material Master Data and Its Effect on the Usefulness of Distributed ERP Systems. *Lecture Notes in Computer Science*, Vol. 4231. 2006. pp. 362-71.
- [15] Kokemüller, J., Weisbecker, A. Master Data Management: Products and Research. In proceedings of 14<sup>th</sup> International Conference on Information Quality (ICIQ). 2009.
- [16] Lee, Y., Pipino, I., Funk, J., Wang, R. *Journey to Data Quality*. The MIT Press, Cambridge. 2006.
- [17] Loshin, D. *Master Data Management*. Morgan Kaufman, Burlington. 2009.
- [18] Loshin, D. *The Practitioner's Guide to Data Quality Improvement*. Morgan Kaufman, Burlington. 2011.
- [19] Loser, C., Legner, C., Giganis, D. Master Data Management for Collaborative Service Processes. In Proceedings of the International Conference on Service Systems and Service Management 2004. Beijing.
- [20] Lucas, A. Corporate Quality Management. Iberian Conference on Information Systems and Technologies 2010. pp. 524-548.
- [21] Maedche, A. An ERP-centric master data management approach. AMCIS 2010. Paper 384.
- [22] McKnight, W. Master Data Management and the Elephant. *Information Management Magazine*, Nov/Dec

2009.

- [23] Moss, L. T. Critical Success Factors for Master Data Management. *Cutter IT journal*, 20 (9), 2007. pp. 7-12.
- [24] Myers, M.D. *Qualitative Research In Business & Management*. Sage Publications, London, 2009.
- [25] Oracle. Building the business case for master data management in the public sector. An Oracle white paper. April 2011.
- [26] Otto, B., Reichert, A. Organizing Master Data Management: Findings From an Expert Survey. In Proceedings of the 2010 ACM Symposium on Applied Computing.
- [27] Poolet, M. Master Data Management. A Method for Reconciling Disparate Data Sources. *SQL Server magazine*, January 2007.
- [28] Radcliffe, J. The Seven Building Blocks Of MDM: A Framework for Success. Gartner research, Paper G00151496. 2007.
- [29] Redman, T. *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business School Press, 2008.
- [30] Samaranayake, P. Enhanced Data Models For Master And Transactional Data in ERP Systems – Unitary structuring approach. International Multiconference of Engineers & Computer Scientists (IMECS). 2008. pp. 1607-1614.
- [31] Shankar, R. Master Data Management Strategies to Start Small and Grow Big. *Business Intelligence Journal*, 13 (3). 2008. pp. 37-47.
- [32] Silvola, R., Jaaskelainen O., Kropsu-Vehkaperä, H., Haapasalo, H. Managing One Master Data – Challenges and Preconditions. *Industrial Management & Data Systems*, 111 (1). 2011. pp. 146-162.
- [33] Smith, H.A., McKeen, J.D. Master data management: Salvation or Snake Oil?. *Communications of the Association for Information Systems*, 23 (4). 2008. pp. 63-72.
- [34] Snow, C. Embrace the Role and Value of Master Data. *Manufacturing Business Technology*, 26 (2). 2008. pp. 38-40.
- [35] Tuck, S. Is MDM the Route to the Holy Grail? *Journal of Database Marketing & Customer Strategy Management*, 15 (4). 2008. pp. 218-220.
- [36] Vilminko-Heikkinen, R., Dahlberg, T. Heikkilä, J., Heikkilä, M., Pekkola, S. Framework and Research Agenda For Master Data: A Literature Review on an Emerging Phenomenon. Submitted to *Communications of the Association for Information Systems Manager*. (In progress).

# **THE STATE OF INFORMATION AND DATA QUALITY EFFORTS IN TODAY'S ORGANIZATIONS**

(Practice-Oriented-Paper)

**Elizabeth Pierce**

University of Arkansas at Little Rock, USA

[expierce@ualr.edu](mailto:expierce@ualr.edu)

**C. Lwanga Yonke**

International Association for Information and Data Quality (IAIDQ)

[lwanga.yonke@iaidq.org](mailto:lwanga.yonke@iaidq.org)

**Piyush Malik**

IBM, USA

[piyush.malik@ua.ibm.com](mailto:piyush.malik@ua.ibm.com)

**Chitra Kagathur Nargaraj**

University of Arkansas at Little Rock, USA

[cxkagathurn@ualr.edu](mailto:cxkagathurn@ualr.edu)

**Abstract:** This report presents the findings of a survey jointly conducted by the International Association for Information and Data Quality (IAIDQ) and the Information Quality Program at the University of Arkansas at Little Rock (UALR-IQ) between March 19 and April 20, 2012. The purpose of the survey was to better understand the current state of information and data quality programs and practices in organizations around the world. The goal was to provide valuable insights for information/data quality practitioners, job seekers, employers, and the academic community in evaluating existing conditions and to aid in setting the agenda for future growth of the discipline. This ICIQ paper is a condensed version extracted from the full industry report that IAIDQ will publish in Fall 2012.

**Key Words:** Data Quality, Information Quality

## **BACKGROUND**

In early 2012, a team of UALR-IQ researchers and IAIDQ members developed a questionnaire to gather insights about information and data quality programs and practices in today's organizations. The survey was officially launched on March 19, 2012. IAIDQ sent several invitations via e-mail to individuals on its mailing list, asking them to complete the web-based survey. Invitations were also distributed via several data quality web sites and social networking groups. The survey closed on April 20, 2012.

Once the data collection period ended, the raw survey data were checked to eliminate any duplicates or abandoned survey responses (i.e., surveys where individuals exited the survey before completing any IDQ-related questions). A total of 296 participants started the survey. After duplicates and abandoned survey responses were eliminated, 270 participant responses remained. These 270 participants who completed our survey represented a diverse set of organizations from around the world. A summary of our participants' demographics is included in the Appendix of this paper. This work is a condensed version of the full industry report that IAIDQ will publish in Fall 2012 [1].

This paper summarizes our findings in four areas

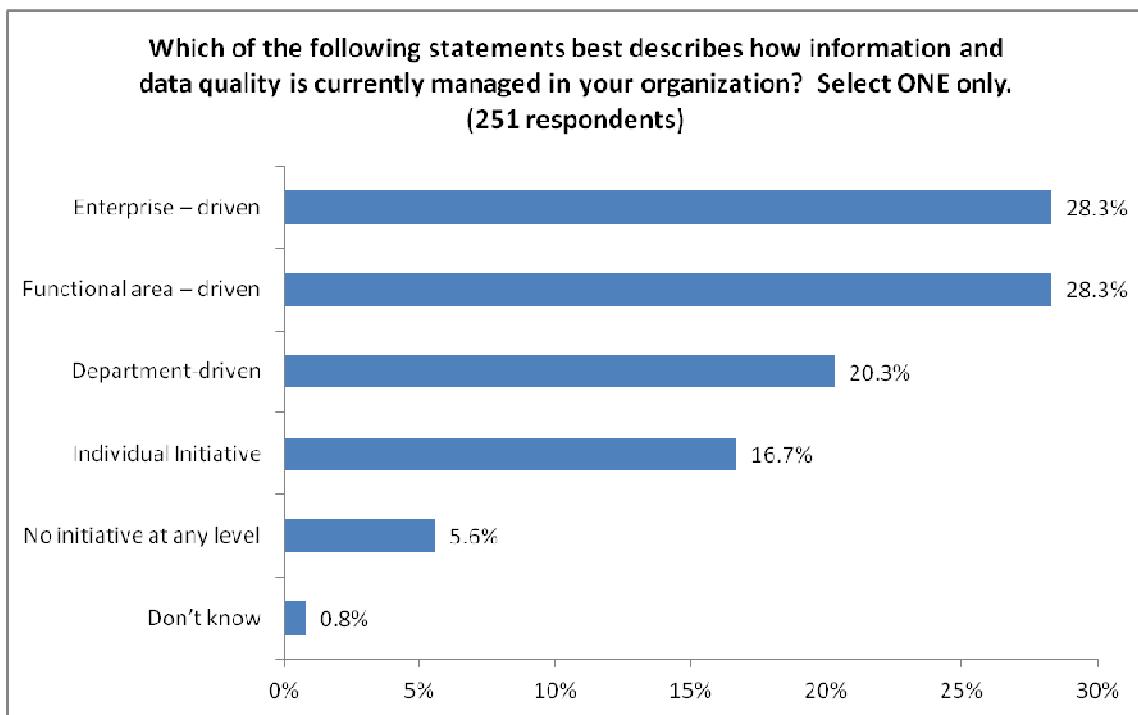
- Organization of Information & Data Quality (IDQ) Efforts
- Information and Data Quality (IDQ) Processes
- Information & Data Quality (IDQ) Maturity
- Information & Data Quality (IDQ) Tools

## ORGANIZATION OF INFORMATION & DATA QUALITY (IDQ) EFFORTS

The questions in this section of the survey focused on how organizations are structuring their information and data quality improvement efforts.

### ***How is IDQ managed in organizations?***

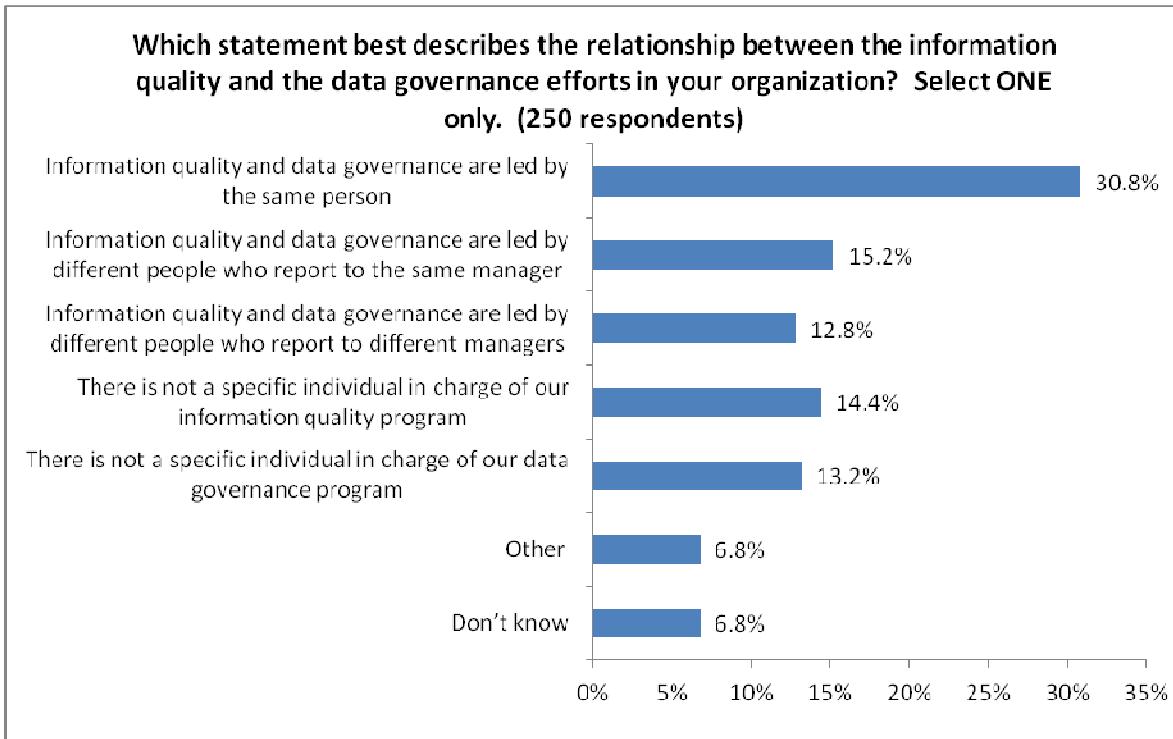
According to our survey results, most IDQ management efforts are driven at either the enterprise or functional areas. 28.3% of participants said that IDQ efforts are enterprise-driven, meaning senior leadership is involved with managing the quality of key information and data assets across the organization with involvement by various functional areas and departments. Another 28.3% of participants indicated that IDQ efforts are driven by the functional areas that are responsible for managing the quality of their information and data assets with participation from the departments that report to those areas. About 20% of participants said that departments are responsible for managing the quality of their organizations' information and data assets. 16.7% of participants reported that in their organizations information and data quality management is left to individuals to pursue on their own initiative while 5.6% of participants reported no information and data quality management at any level in their organizations.



### ***What is the relationship in organizations between IDQ efforts and Data Governance?***

We loosely define Data Governance as the collective set of decision-making processes for the use and value-maximization of an organization's data assets throughout its lifecycle [1]. Because issues surrounding the quality, integrity, or usability of information sometimes fall under the scope of an organization's data governance initiatives, we asked participants to share with us the relationship between IDQ efforts and data governance efforts in their organizations. While nearly a third (30.8%) of participants said that

in their organizations information quality and data governance are led by the same person, the rest pointed out that a wide range of relationships exist. In addition to the options listed in the survey question, a few individuals noted that in their organizations IDQ initiatives report directly into their Data Governance Group. Furthermore about 5% of participants wrote in comments explaining that no relationship exists because their organizations either did not have a Data Governance program or were still in the very early stages of developing a Data Governance program.



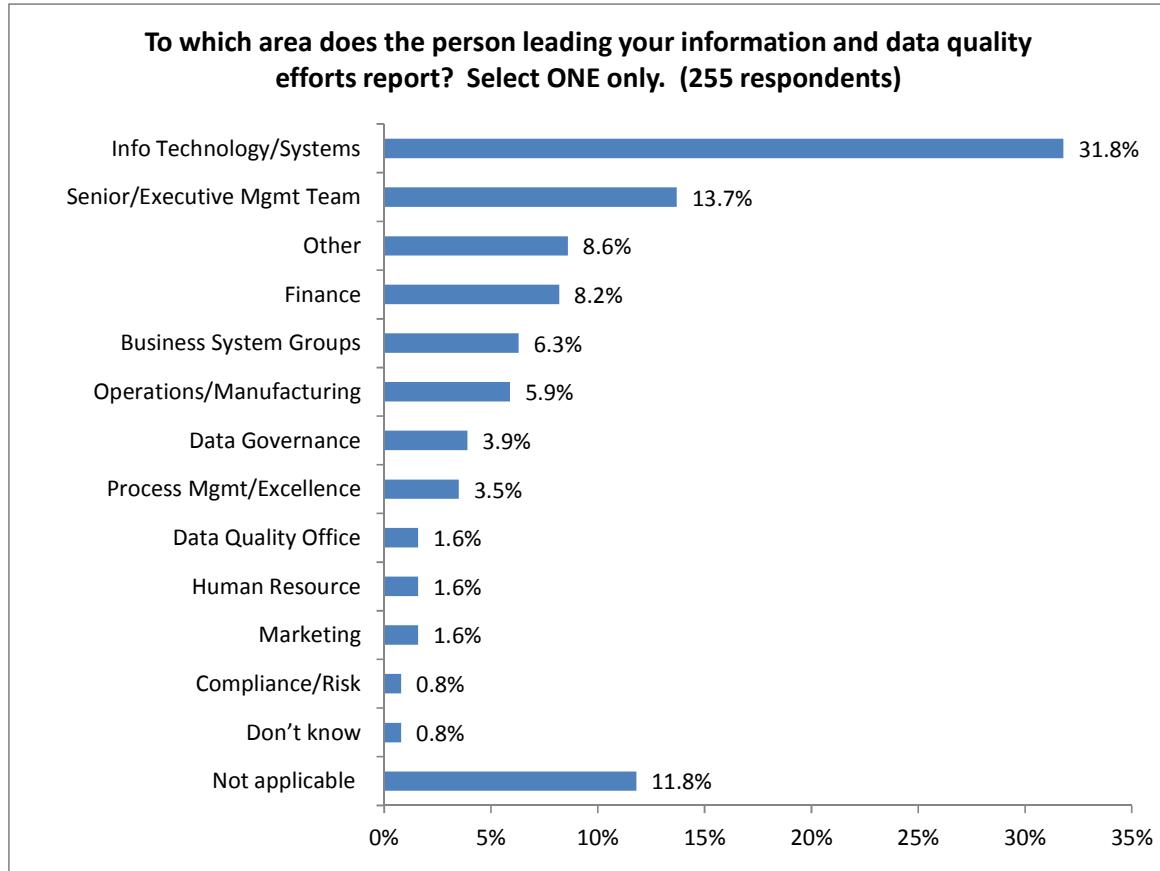
Note: The terms “information” and “data” can be used interchangeably for this question.

### **Where does the person leading IDQ efforts report? What is their level?**

According to our survey, Information Technology/Information Systems is the most common reporting area (31.8%) for individuals leading an organization’s IDQ efforts. This in turn suggests that 68.2 % of IDQ efforts are led by people who report outside of IT/IS. This would be a very encouraging fact, given the need to establish business ownership of IDQ efforts. According to the survey responses, the most common non-IT/IS area is the Senior/Executive Management Team (13.7%). It is apparent from the rest of the selected choices that individuals leading their organizations’ IDQ efforts report to a variety of functional areas. In addition to the choices listed in the question, participants contributed other areas where their IDQ leader reports such as Business Intelligence, Supply Chain Management, Internal Audit, Research, Medical Affairs, Asset Management, and Data/Information Management Groups separate from IT. About 12% of participants indicated “Not applicable”, most probably because their organization does not have a specific individual leading their organization’s IDQ efforts.

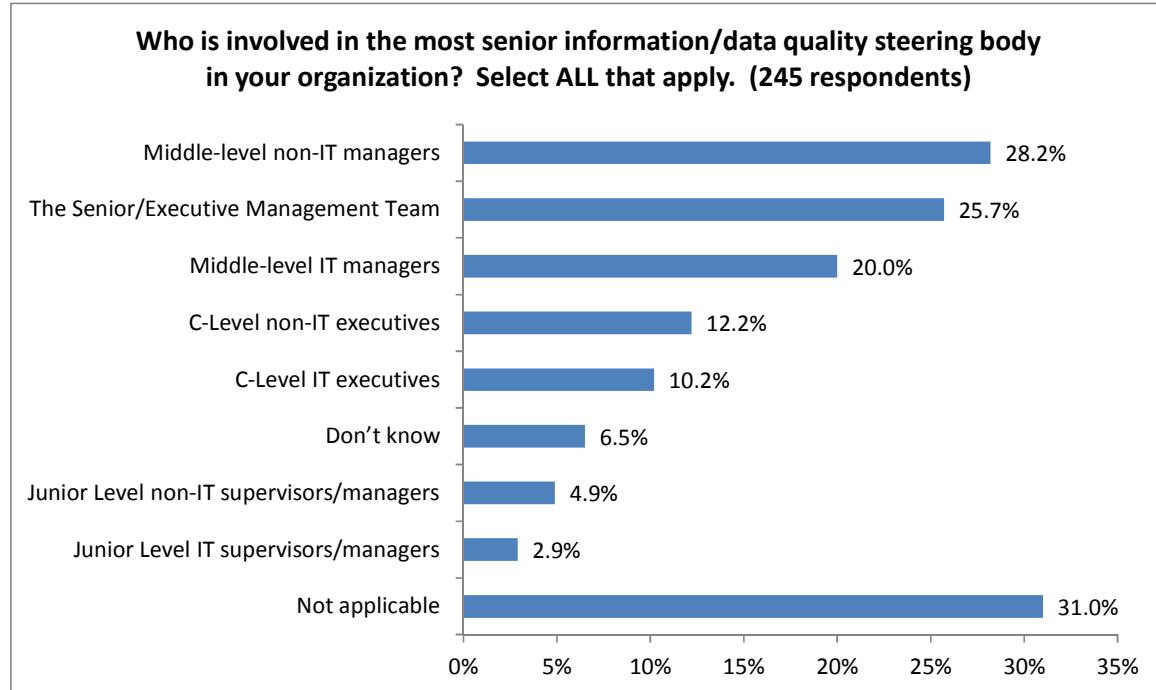
As a follow up to this question for those survey participants who indicated that their organizations had an IDQ leader, we wanted to know how high up in the organization this position was located. A third (33.0%) said that three levels separate the most senior leader of their organization and the person most directly in charge of their IDQ efforts with roughly another third (33.9%) reporting less than three levels and the remainder reporting more than three levels or unsure. See IAIDQ for the full report with these

charts [1].



### ***Who in the organization is involved in the most senior IDQ steering body?***

In many organizations, information and data quality initiatives and processes are guided by one or more bodies such as a Data Council, Steering Committee or the equivalent. According to our survey responses, middle-level business managers (i.e., non-IT) (28.2%), members of the senior/executive management team (25.7%), and middle-level IT managers (20.0%) were the ones most frequently cited as being involved in their organizations' IDQ steering body. Nearly a third (31.0%) selected "Not applicable" meaning those participants came from organizations that do not have a senior IDQ steering body.



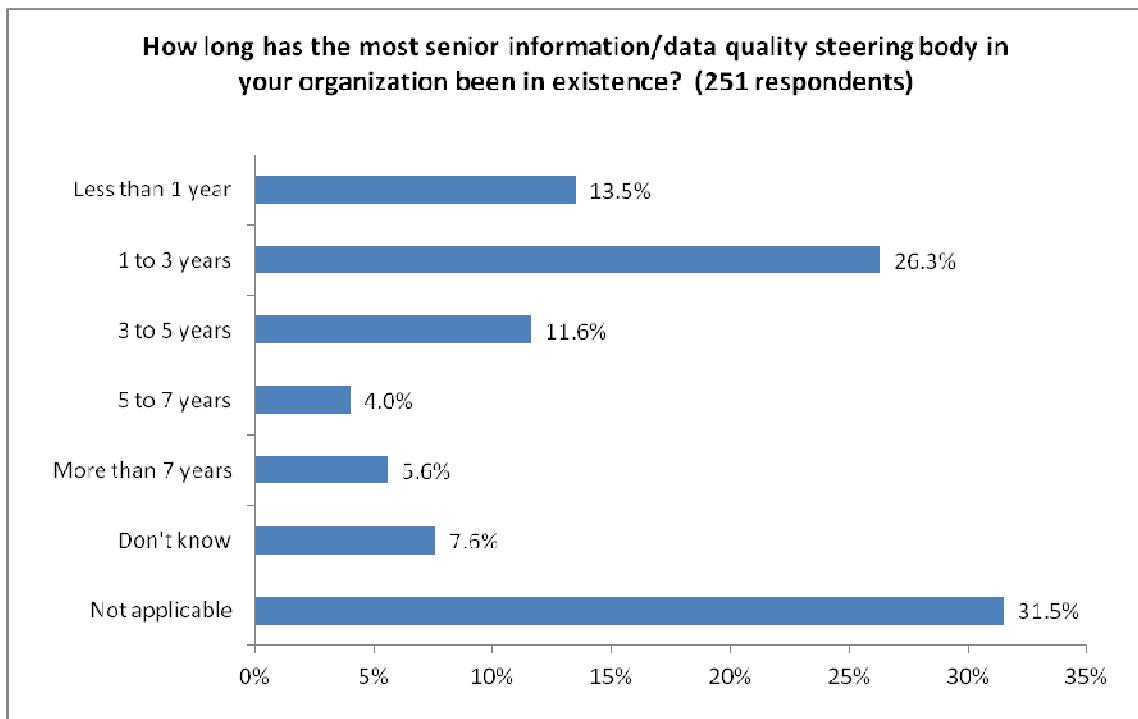
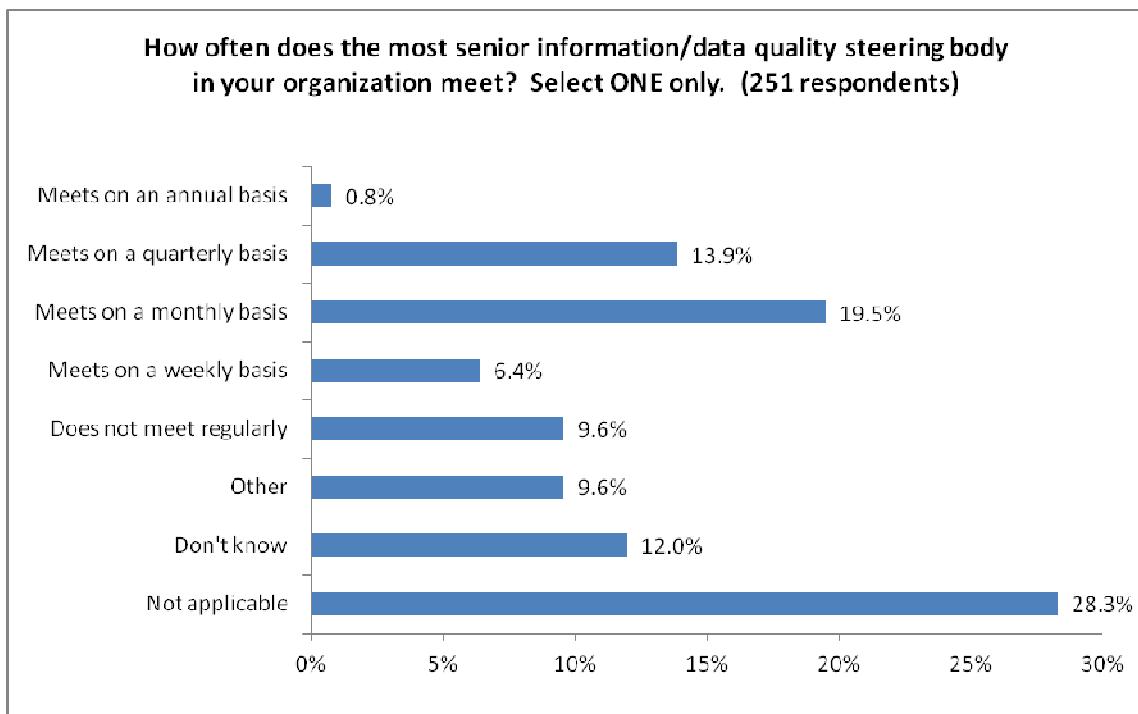
Note: “non-IT” means lines of business other than Information Technology / Information Systems.

**How often does the most senior IDQ steering body in organizations meet?**

For those participants whose organizations have a senior IDQ steering body, meeting monthly (19.5%) was selected most frequently followed by meeting quarterly (13.9%). In addition to the choices listed, several participants wrote in alternative meeting schedules for their most senior IDQ steering body such as every two months, every two weeks, and twice a year.

**How long has the most senior IDQ steering body existed in organizations?**

Participants whose organizations have a senior IDQ steering body reported that this body is fairly new with the majority reporting that their senior IDQ steering body was less than three years old.

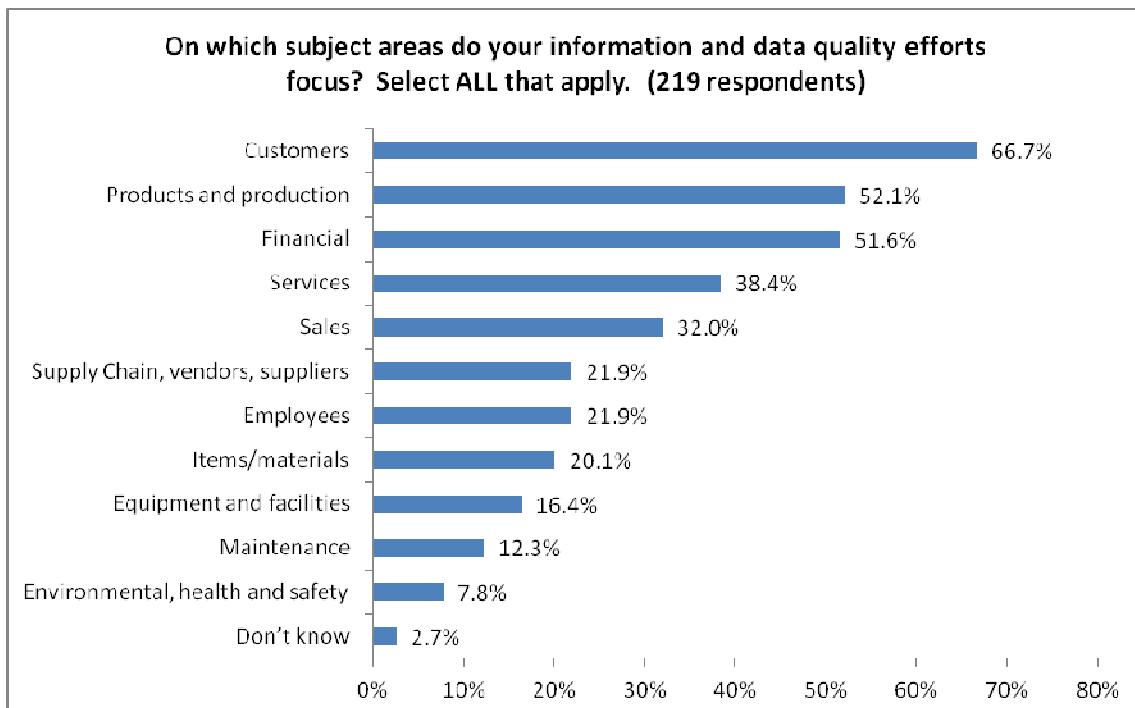


## INFORMATION AND DATA QUALITY (IDQ) PROCESSES

To learn more about what processes organizations are following for their IDQ efforts, we asked survey participants to tell us what their organizations are doing in regards to their IDQ activities.

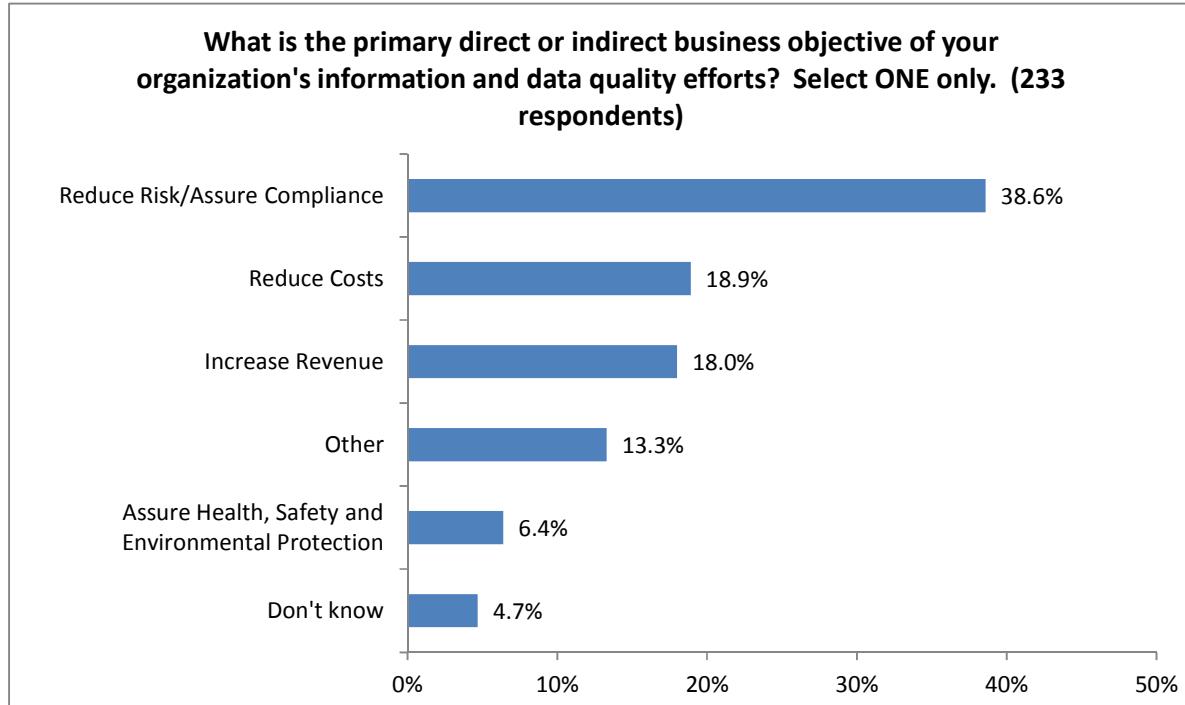
### ***What are the key subject areas for IDQ?***

According to our survey results, the top three domain areas on which organizations are focusing their information and data quality efforts are Customers (66.7%), Products and production (52.1%), and Financial (51.6%). In addition to the subject areas that we listed, participants provided several more information areas such as Property/Real Estate, Education, Research/Scientific, Health Care/Patients, Regulatory Compliance, and Road/Transportation.



### ***What are the primary business objectives for IDQ?***

Linking information and data quality activities to business needs is essential. Participants told us that the primary direct or indirect business objective of their organization's information and data quality efforts is to "Reduce Risk and Assurance Compliance" (38.6%). This area was cited twice as much as the objective to "Reduce Costs" (18.9%) or the objective to "Increase Revenue" (18.0%). In addition to the objectives listed in the survey, people mentioned other objectives such as better customer satisfaction/service, improved business decision making, reliable reporting, and accurate information/data-based products.



**What are the main motivators behind IDQ efforts?**

Participants named the general desire to improve the quality of their data as the main driver or catalyst behind their information and data quality efforts (68.4%). Other motivations chosen included Data Warehousing/Business Intelligence (47.2%), Compliance/Risk/Fraud/Legal Requirements (39.8%), and Master Data Management Projects (39.4%). In terms of other motivators, people referenced the increasing complexity of the business, cost/asset management, process improvement, long term archival requirements, patient safety, profitability measurement and reporting requirements.

<b>Which of the following are the main drivers, motivations, or catalysts behind your information and data quality efforts? Select ALL that apply. (231 respondents)</b>	
General desire to improve the quality of our data	68.4%
Data Warehousing / Business Intelligence	47.2%
Compliance / Risk / Fraud / Legal Requirements	39.8%
Master Data Management (MDM) project	39.4%
Suffered major negative impact from bad data quality	30.7%
Business Process Automation	28.1%
Customer Data Integration (CDI)	28.1%
Applications / Systems Integration	26.8%
Customer Relationship Management (CRM) project	26.4%
Enterprise Architecture	25.1%
Information Security / Privacy	18.6%
Enterprise Resource Planning (ERP) project	15.6%
Unstructured Data	13.9%
Database Marketing	11.7%
Reaction to competitors' activity	10.4%
Product Information Management (PIM) project	10.0%
Sales Force Automation	9.5%
Big Data	9.1%
Service-Oriented Architecture (SOA) project	8.7%
Merger & Acquisition planning or implementation	8.2%
Cloud Computing	2.2%
Don't know	2.2%

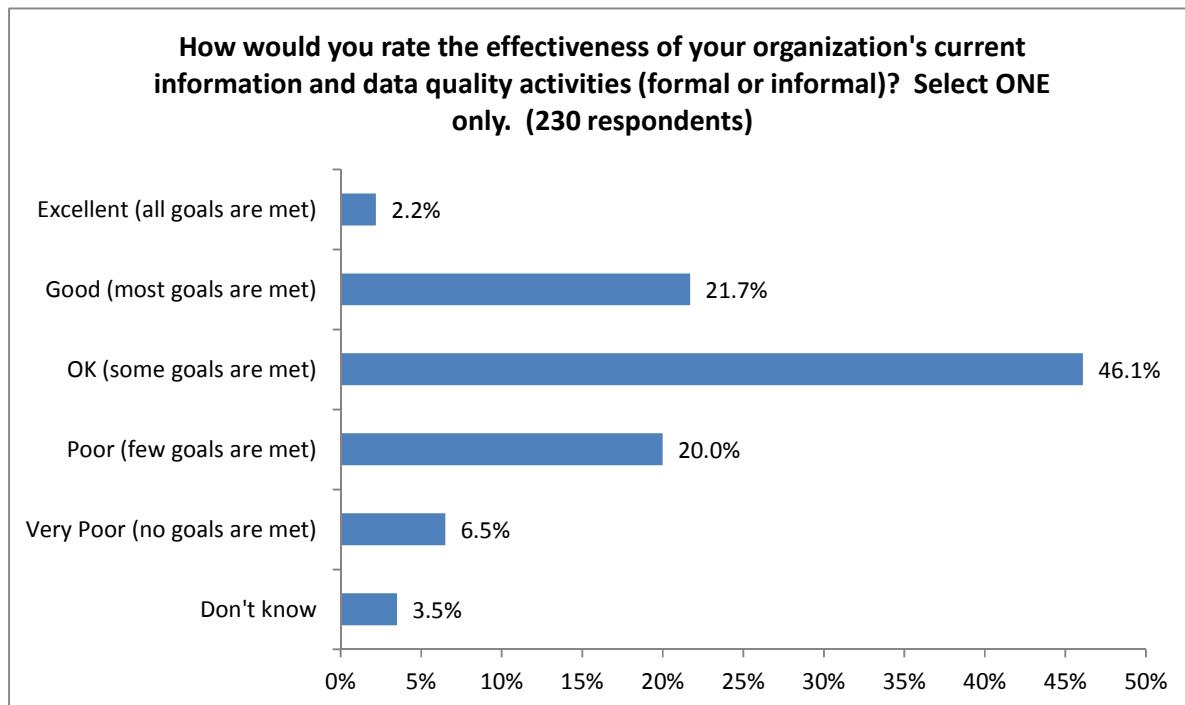
### **What is the current emphasis on IDQ activities?**

We asked participants to rate the amount of effort their organization spends on a variety of IDQ activities. We ordered the list by those activities that had the largest percentage of responses in the Moderate to Large-Scale efforts categories. Based on this, our survey indicates that the top 6 information and data quality activities that their organizations spend the most effort on are as follows:

- Data cleansing/remediation
- Propose, select or charter data quality improvement projects
- Data Quality monitoring
- Standardize data definitions across the organization
- Data Quality assessment
- Define and standardize common business rules across the organization

### **How effective are IDQ activities?**

Participants indicated that the effectiveness of their organization's current information and data quality activities (formal or informal) is mostly Okay (i.e., some goals are met) (46.1%). On the positive side, 21.7% of participants selected Good (i.e., most goals are met) and 2.2% selected Excellent (i.e., all goals are met). On the negative side, 20.0% of participants selected Poor (i.e., few goals are met) and 6.5% selected Very Poor (i.e., no goals are met).



## **INFORMATION & DATA QUALITY (IDQ) TOOLS**

The data quality tools market has been growing rapidly over the past several years, increasing organizations' ability to assure data quality. What tools are organizations currently using for their IDQ efforts? Here is the feedback we received from our survey participants.

### **What types of tools are being used in IDQ efforts?**

According to our participants, the top five categories of data quality tools being used by organizations are (1) Data profiling and quality assessment, (2) Data quality monitoring, (3) Data remediation / cleansing, (4) Data matching and reconciliation, and (5) Extract-Transform-Load. In addition to the ones listed, participants wrote in several other categories including "Statistical Analysis", "Microsoft Excel", "SQL scripts", "Quality Assurance/Quality Control" and "Program Management."

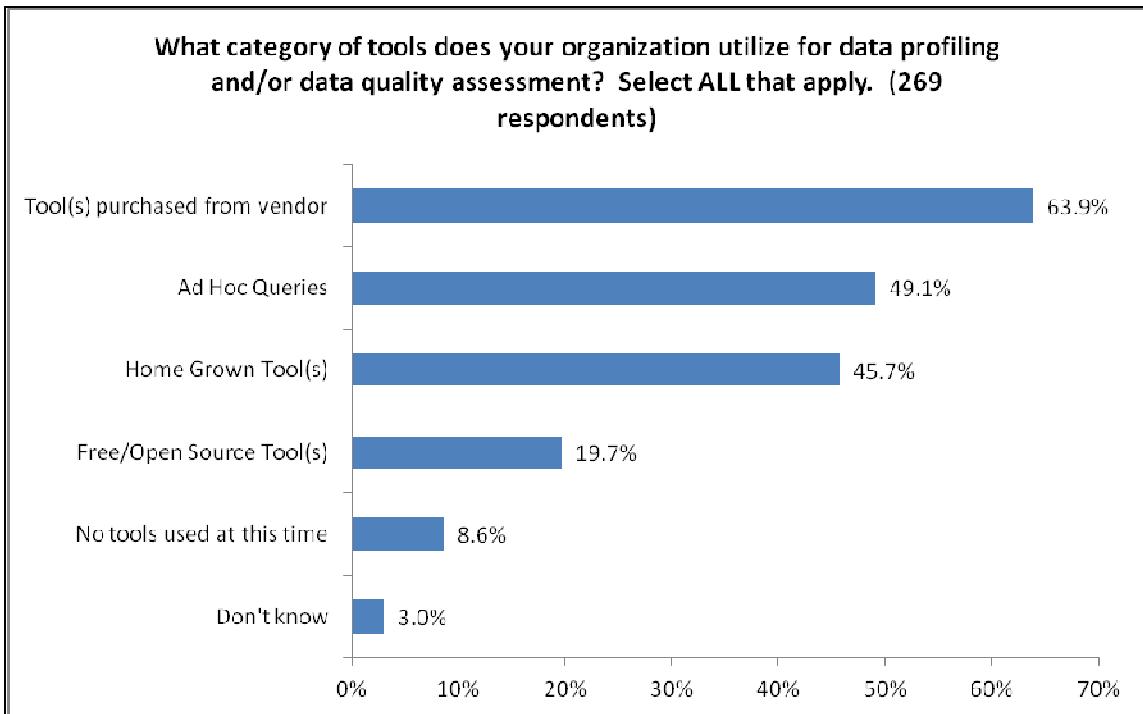
<b>Identify the categories of data quality tools currently used in your organization. Select ALL that apply. (221 respondents)</b>	
Data profiling and quality assessment	65.2%
Data quality monitoring	61.1%
Data remediation / cleansing	57.0%
Data matching and reconciliation (data de-duplication)	52.9%
Extract-transform-load (ETL) and other data integration tools	51.1%
Data modeling (computer-aided software engineering)	41.2%
Data parsing and standardization	40.3%
Metadata management tools (Business and Technical)	37.6%
Master data management (MDM)	35.3%
Data enrichment	30.3%
Business Process Management / Workflow	30.3%
Data discovery (relationship and mappings)	29.0%
Data visualization	25.8%
Data lineage	24.9%
Business rules engine	24.0%
Customer data integration (CDI)	20.8%
Rules discovery	15.4%
Collaboration tools (for data governance and stewardship workflows)	14.9%
Text Mining / Semantic	12.2%
Ontology and hierarchy building	12.2%
Product Information Management (PIM)	7.7%

### ***How important are tools to IDQ efforts?***

Not surprisingly, the top five tools listed in the previous question are also the five tools that our participants considered most important to their information data quality efforts: (1) Extract-Transform-Load, (2) Data quality monitoring, (3) Data remediation / cleansing, (4) Data matching and reconciliation, (5) Data profiling and quality assessment. One interesting item to note is that the order between the top five are very similar with the following exception. While Data Profiling and quality assessment tools were cited as the most used, Extract-Transform-Load and other data integration tools were considered the most important to the organization's information and data quality efforts by our participants.

### ***Where do organizations get their tools for data profiling and assessment?***

It appears that most organizations are using tools purchased from vendors to conduct data profiling and assessments in conjunction with Ad Hoc Queries and Home Grown Tools.



## INFORMATION & DATA QUALITY (IDQ) MATURITY

To discover where organizations rate in terms of IDQ maturity, we asked survey participants several questions based on attributes of the COBIT 4.1 Maturity Model [3]. COBIT was originally developed for IT governance. We chose those attributes that we felt were especially relevant to the management of information and data quality and modified the wording of the maturity levels accordingly.

### ***Responsibility and Accountability***

Nearly half of participants (47.7%) indicated their organizations had reached the Defined Level or above. At the Defined level organizations have defined IDQ responsibility and accountability roles with individuals assigned to carry out those duties; however, issues regarding authority still remain. Over half of the participants (52.3%) reported their organizations had not yet reached the Defined level. Fewer than 11% of participants felt their organizations had reached the higher stages of maturity when it comes to ensuring that an effective system is in place for defining, staffing, and empowering IDQ roles.

Which of the following statements best describes the <u>responsibility and accountability</u> for information and data quality among employees in your organization? Select ONE only. (220 respondents)			
Maturity Level	Description	Response Percent	Cumulative Percent
<b>5 – Optimized</b>	Information / data stewards and others with information/data quality roles are empowered to make information / data quality decisions and to take action. The acceptance of responsibility has been cascaded down throughout the organization in a consistent fashion. An effective governance structure has been established.	3.6%	3.6%
<b>4 – Managed</b>	Information and data quality responsibility and accountability are accepted and working in a way that enables information/data stewards and others with information/data quality roles to fully discharge their responsibilities. An appropriate reward structure is in place.	7.3%	10.9%
<b>3 – Defined</b>	Information and data quality responsibility and accountability are defined and information/data stewards have been identified. Occasionally, the information/data stewards and others with information and data quality roles may lack the full authority to exercise their responsibilities.	36.8%	47.7%
<b>2 – Repeatable</b>	One or more individuals have assumed responsibility for information quality and are usually held accountable, even if this is not formally agreed. There is often confusion and blame about responsibility when information and data quality problems occur.	25.5%	73.2%
<b>1 - Ad-hoc</b>	There is no clear definition of accountability or responsibility for information and data quality issues. People take ownership of information/data quality issues based on their own initiative as problems arise.	26.8%	100%

### **Policies, Plans, and Procedures**

When it comes to IDQ processes, about a third of participants (37%) felt their organizations had reached the Defined Level or above. At the Defined level organizations have defined and documented IDQ processes and policies along with more formal and structured practices for communicating these plans. The majority of the participants (63%) reported their organizations had not yet reached the Defined level. Less than 8% of participants said their organizations had reached the higher stages of maturity when it comes to ensuring that an effective system is in place for defining and following IDQ best practices.

Which of the following statements best describes the status of information and data quality <u>policies, plans, and procedures</u> in your organization? Select ONE only. (222 respondents)			
Maturity Level	Description	Response Percent	Cumulative Percent
<b>5 - Optimized</b>	Benchmarking against external best practices and standards for information/data quality are applied. The effectiveness of information and data quality processes and policies are continually being improved. Management is engaged in proactive and ongoing communication of these practices.	1.8%	1.8%
<b>4 – Managed</b>	All aspects of information and data quality processes and policies are documented and repeatable. Policies have been approved and signed off on by management. Standards for managing and improving the quality of information and data quality processes and policies are adopted and followed. Management is communicating on these practices on a frequent and widespread basis.	5.9%	7.7%

<b>3 – Defined</b>	Information and data quality processes and policies are defined and documented for all the subject areas the organization is focusing on. Management is becoming more formal and structured in its communication of these practices	29.3%	37.0%
<b>2 – Repeatable</b>	Some documentation and/or understanding of common information and data quality processes and policies are emerging, but are largely intuitive because of individual expertise. Management is communicating on some of these practices.	35.1%	72.1%
<b>1 - Ad-hoc</b>	Information and data quality processes and policies are largely undefined. Several ad hoc processes and policies exist, but management communication about these practices is sporadic.	27.9%	100%

### ***Goal Setting and Measurement***

A similar situation exists when it comes to IDQ performance measurements. About 43% of participants indicated their organizations had reached the Defined Level or above. At the Defined level organizations have set some IDQ goals and metrics, but there are consistency issues in applying these performance measures which often lack a clear link with strategic goals in the organization. In addition communication about these IDQ goals and metrics are not widespread. Nearly half of the participants (47.5%) felt their organizations had not yet reached the Defined level. About 15% of participants believed their organizations were at the higher stages of maturity when it comes to ensuring that an effective system is in place for defining, measuring, and monitoring IDQ performance.

Which of the following statements best describes the status of information and data quality <u>goal setting and measurement</u> in your organization? Select ONE only. (221 respondents)			
Maturity Level	Description	Response Percent	Cumulative Percent
<b>5 - Optimized</b>	An organization-wide integrated information and data quality performance measurement system is in place. It links information/data goals to organizational strategic goals. Goals are routinely met. Deviations are consistently noted by management and root-cause analysis is applied. Continuous improvement of information and data quality processes is ongoing.	2.3%	2.3%
<b>4 – Managed</b>	Efficiency and effectiveness goals are set, communicated, measured, and linked to organization's strategic goals. Continuous improvement of information and data quality processes is emerging.	12.7%	15.0%
<b>3 – Defined</b>	Some information and data quality effectiveness goals and measures are set, but may not be widely communicated. There is no clear link to strategic organizational goals. Measurement processes for these goals are emerging but are not consistently applied.	27.6%	42.6%
<b>2 - Repeatable</b>	Some information and data quality goal setting occurs. Measurement of success against these goals is inconsistent and typically limited to a few areas.	28.1%	70.7%
<b>1 - Ad-hoc</b>	Information and data quality goals are not clear and no measurement exists.	29.4%	100%

## **CHALLENGES AHEAD FOR INFORMATION & DATA QUALITY (IDQ)**

Perhaps the most critical finding of this survey was the response to our question regarding the obstacles that people perceived as most inhibiting data quality improvement in their organizations. IDQ professionals listed numerous obstacles that they face on a regular basis in their organizations.

- Lack of accountability and responsibility for data quality
- Too many information silos
- Lack of awareness or communication of the magnitude of data quality problems
- Lack of common understanding of what data quality means
- Lack of awareness or communication of the opportunities associated with high quality data
- Lack of senior leadership in tackling data quality issues
- Lack of data quality policies, plans, and procedures
- Perception that data quality is an IT issue only rather than an organization wide issue (and in some organizations there may be a reverse perception that data quality is a business issue only and cannot be helped with IT support)
- Lack of data quality goal setting and measurement
- Lack of data quality skills and expertise
- Lack of data quality tools and automation
- Lack of resources including limited staff to manage data issues and promote data quality, cost to build a good data quality program, time to get proper tools and automation in place.
- Out of date policies, plans, and procedures.
- Lack of grass roots development of data quality as a strategic vision
- Lack of data quality rules that are customer focus
- Lack of understanding by data collectors of their impact on quality
- Lack of awareness of impact of frequent organizational changes on contextual meaning and usability of data assets

If Information and Data Quality is to make progress as a discipline, these obstacles must be alleviated. IAIDQ and other IDQ leaders must continue to work together to raise awareness across the diverse stakeholders groups. It will also be important for IAIDQ and others to expand their efforts to promote the development and exchange of the IDQ knowledge base, and to provide support and strategies for those trying to establish and grow an IDQ culture in their organization.

## **REFERENCES**

- [1] Pierce, E., Yonke C. L., Malik, P., Nargaraj, C. K. (est. July 2012), *The State of Information and Data Quality: Understanding How Organizations Manage the Quality of their Information and Data Assets*, IAIDQ.
- [2] Pierce, E.; Dismute, W. S., Yonke, C.L.(2008) *The State of Information and Data Governance: Understanding How Organizations Govern their Information and Data Assets*, IAIDQ.  
<http://iaidq.org/publications/pierce-2008-04.shtml>
- [3] ISACA. *COBIT 4.1: Framework for IT Governance and Control*, accessed on February 15, 2012, [www.isaca.org/Knowledge-Center/cobit/Pages/Overview.aspx](http://www.isaca.org/Knowledge-Center/cobit/Pages/Overview.aspx). Questions based on content from COBIT 4.1, which is used by permission of the IT Governance Institute (ITGI). Copyright IT Governance Institute. All rights reserved.

## **APPENDIX - SURVEY PARTICIPANTS AND THEIR ORGANIZATIONS**

Here is a brief summary of the characteristics of the 270 participants who completed our survey.

---

### **Roles that survey participants assume in their organizations:**

- Supervisor / Manager -- 42.2%
  - Staff / Faculty / Individual Contributor – 31.9%
  - Senior Level Executive – 13.7%
  - Other – 8.1%
  - Owner – 4.1%
- 

### **Functional areas that best describe the work of survey participants:**

- Information Technology / Information Systems – 37.7%
  - Business line or other non-IT/IS – 36.6%
  - Consultant – 11.2%
  - Other – 9.3%
  - Academia -- 3.4%
  - Software vendor – 1.9%
- 

### **Top 12 countries where survey participants work:**

- United States – 49.3%
  - Australia – 7.5%
  - India – 6.0%
  - United Kingdom – 5.6%
  - Canada – 4.5%
  - South Africa – 3.0%
  - Ireland -- 2.6%
  - Netherlands – 2.2%
  - Belgium – 1.9%
  - Columbia – 1.9%
  - Philippines – 1.9%
  - China – 1.5%
  - Other – 12.3%
- 

### **Part of organization that survey participants had in mind when answering questions:**

- The entire organization – 49.3%
  - A functional area – 24.8%
  - A department – 15.2%
  - A subsidiary of the organization – 10.7%
- 

### **Work force size of survey participants' organizations:**

- More than 10,000 employees – 31.5%
  - 2,500 to 10,000 – 23.3%
  - 500 to 2,500 employees – 16.3%
  - 100 to 500 employees -- 13.0%
  - 50 to 100 employees – 6.7%
  - Fewer than 50 employees – 8.1%
  - Unsure of work force size – 1.1%
-

**Annual revenue size of survey participants' organizations (USD):**

- More than ten billion – 19.3%
  - One billion to ten billion -- 21.1%
  - 100 million to one billion – 15.9%
  - Ten million to 100 million – 9.6%
  - One million to ten million – 8.9%
  - Less than one million – 3.7%
  - Unsure of annual revenue – 21.5%
- 

**Type of organizations for whom survey participants work:**

- Private company – 39.0%
  - Public company – 34.9%
  - Non-profit – 7.4%
  - College/University – 4.8%
  - Federal government – 5.6%
  - State government -- 5.2%
  - Local government – 1.5%
  - Military – 1.5%
- 

**Top 15 industries associated with survey participants' organizations:**

- Financial Services – 13.3%
  - Energy / Oil & Gas – 12.2%
  - Healthcare – 8.5%
  - Consulting / Professional Services – 7.8%
  - Insurance – 7.4%
  - Government (Federal/National/State/Local)– 6.3%
  - Software / Internet – 5.6%
  - Education -- 5.6%
  - Government: Federal/National - 4.4%
  - Manufacturing (non-computers) – 3.7%
  - Telecommunications / Communications – 3.7%
  - Pharmaceuticals – 2.6%
  - Retail / Wholesale Distributions – 2.6%
  - Manufacturing (computers, technology) – 2.6%
  - Utilities – 2.2%
  - Other – 17.8%
- 

**Market scope of customers that survey participants' organizations serve:**

- International – 49.6%
  - National – 33.3%
  - Regional (state or province) – 13.0%
  - Local (e.g., metropolitan area) – 3.3%
  - Unsure -- 0.7%
-

# DESIGNING BUSINESS PROCESSES ABLE TO SATISFY DATA QUALITY REQUIREMENTS

(Research-in-Progress)

**Angélica Caro, Alfonso Rodríguez**

Department of Computer Science and information Technologies  
University of Bío-Bío, Chillán, Chile  
{mcaro, alfonso}@ubiobio.cl

**Cinzia Cappiello**

Dipartimento di Elettronica e Informazione – Politecnico di Milano  
Piazza Leonardo da Vinci 32, 20133 Milano, Italy  
cappiell@elet.polimi.it

**Ismael Caballero**

Alarcos Research Group-Instituto de Tecnología y Sistemas de la Información,  
University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain  
Ismael.Caballero@uclm.es

**Abstract:** Nowadays, data quality is a fundamental issue to be considered in order to avoid inefficiencies and to fully exploit all the benefits of adopting sophisticated information technology platforms that can support essential activities for business such as decision making, business intelligence and customer services. Business efficiency and effectiveness also depend on the way in which business processes are modeled. A sound modeling of the business processes is becoming a higher priority for business managers and analysts since documenting and understanding business processes support them in the optimization and improvement of the business functions. In this paper we propose a methodology (named BPiDQ) to consider data quality issues in the business process modeling phase to support the design of data quality-aware business processes.

**Key Words:** Data Quality, Business Process Model, BPMN, Data Quality Requirements.

## 1. INTRODUCTION

Modern organizations use different strategies to achieve success, sustainability and competitiveness. Most of them concentrate their efforts in adopting sophisticated information technology platforms that can support essential activities such as making decision, business intelligence, and costumer services, among others. However, these platforms per se are not useful if the core business relies on inefficient processes. For this reason, some organizations have focused their efforts on the definition and management of suitable Business Processes (BP) that optimize the procedures, the use of information technology, and the involvement of the human resources.

On the other hand, in order to avoid inefficiencies and to achieve all the benefits of the adoption of advanced information management solutions, high quality data is also needed [1]. Thus, achieving adequate levels of Data Quality (DQ) could be a strategic approach to consider as part of the business process management.

Formally, DQ is often defined as “*fitness for use*”, i.e., the ability of a data collection to meet users’ requirements [2]. DQ is a multidimensional and subjective concept since it is usually evaluated by means of different criteria, namely DQ dimensions, whose selection of those that better describe users’ DQ requirements and the corresponding evaluation largely depends on the context of use.

Guaranteeing high levels of DQ for the data used in tasks at hands is an important issue especially in information-intensive organizations. In general, poor data quality exposes organizations to non-

depreciable risks especially when a business process relies on incorrect, incomplete or out-of-date data. Such data quality issues might also imply the complete or partial failure of the business process: e.g., the use of a wrong address for a product delivery, or the delay in communicating the needed information in a process with strict temporal constraints. These consequences can be avoided or at least alleviated by adopting suitable strategies to early tackle the DQ necessities as a proactive attitude facing the occurrence of data quality problems when the BP is executed.

A BP model represents the flow of physical items or informational artifacts through a sequence of tasks and sub-processes that operate on them [3]. In business process modeling, the main objective is to produce a description of the business work in order to better understand the process, and eventually, improve it: for example, the way in which a commercial transaction is carried out. Our idea is to model business processes considering in addition the data quality issues, and consequently, including activities able to minimize the risk associated with data-related errors. To this aim, it is also important to have a suitable notation for modeling the essence of the business as clearly as possible. Among all possible choices, a recent study shows that BPMN (Business Process Model and Notation) is one of the most important and popular standard to modeling business process [4]. Unfortunately, BPMN lacks of the mechanisms to represent data quality concerns. In addition, there is no guide either that allows business people to incorporate data quality requirements into the representation of the business model when this is done by means of BPMN. So, as part of our research-in-progress work, and as the main contribution of this paper, we introduce a methodology named “Business Process including Data Quality view point” (BPiDQ), that aims to provision a methodological approach for the modeling and design of data quality-aware business processes as well as the generation of the corresponding DQ requirements for the software development that support the business processes. This contribution extends our previous work [5], which consists of an extension of BPMN 2.0 that allow business people, in a simply way, to identify the critical points where the DQ is crucial for the success of the BP. BPiDQ aims to support the workers (business analyst/designer, DQ expert and System analyst) to improve the BP by means of some changes or by introducing new activities that guarantee the satisfaction of the DQ requirements and derive DQ use case for the software development. In addition, other necessary artifacts that complement the methodology are introduced.

The rest of the paper is organized as follow. Section 2 discusses the related works. The BPiDQ methodology and its main components are introduced in Section 3. To illustrate the use of the BPiDQ an example is developed in Section 4. Finally, Section 5 gives our conclusions and future works.

## **2. RELATED WORKS**

DQ management has been widely recognized as a relevant aspect that deserves to be considered in order to globally improve the effectiveness of organization’s performance [6]. Thus, it is important that business people are aware of DQ requirements from the earliest stages of the design of a business process, i.e., business process modeling. The most used languages to model business processes, namely BPMN and UML [4], do not allow process designers to fully specify DQ requirements at a high level.

To the best of our knowledge, at the present time, there is only one specific notation to represent DQ issues in business process, allowing the depiction of what its authors named information products maps (IP-MAP) [7]. It permits the specification of business processes by means of a conceptual map and a sort of activity diagrams, in which the efforts corresponding to data quality management are properly addressed by means of some specific constructs [7]. Indeed, BPMN is widely recognized as de facto standard to model business processes [4, 8]. Its expressiveness can be and has been already extended, to support some other concerns of interests. For example, it has been extended to support customer needs related with quality requirements such as time, cost, and reliability [9], to submit/response-style user interaction [10], to specify non-functional properties such as performance and reliability oriented to a characterization of the business process [11], to include Business Activity Monitoring (BAM) relevant con-

cepts in BP models [12], to capture the temporal perspective of business processes [13], to include information coming from sensors and smart devices [14], to model security requirements [15], to represent explicitly legal constraints directly by specific artifacts [16], or to analyze business processes performance [17], to name a few.

However, DQ concerns are not new to BP research area: some existing contributions highlight the need of addressing DQ in the business process modeling during the design time. So, for instance, in [18], Soffer explores the inaccuracies of data, the situation where the information system does not truly reflect the state of a domain where a process takes place. The potential negative consequences of data inaccuracy are discussed. The work provides the bases to support the design of robust processes and avoid problems related to data inaccuracy. Bringel et al. in [19] propose a business process pattern to ensure data quality in an organization. The pattern consists in a business process model that can be reused through adaptation in specific organizational scenarios. For this, they define DQ attributes associated with information entities having different meanings depending on the business view and the different organizational dimensions. The Data Excellence Framework is proposed in [1]. This framework describes the methodology, processes and roles required to generate the maximum business value while improving business processes using data quality and business rules. In this approach, DQ requirements are specified as business rules. The set of business rules supporting data quality grows over time as part of the process of continuous improvement. Bagchi et al. in [3] introduced a business process modeling framework for quantitative estimation and management of data quality in information systems. Based on this framework, they propose to exploit the structure provided by the business process flows to estimate errors arising in transaction data and the impact of their propagation to the key performance indicators.

Also, Heravizadeh et al. in [20] proposed the QoBP framework for capturing the quality dimensions of a process. The framework helps modelers in identifying quality attributes in four quality dimensions: quality of functions, quality of input and output objects, quality of non-human resources and quality of human resources. In particular, they specify eleven DQ attributes for the input and output information objects.

Finally, the work presented in [21] introduces some concerns focused on the concept of compliance. Compliance essentially means ensuring that business processes, operations and practices are in accordance with a prescribed and/or agreed set of previously defined norms. Lu et al. consider that a sustainable approach for achieving compliance should fundamentally have a preventative focus, thus achieving compliance by design [21]. Their proposal consists in incorporating compliance issues within business process design methodology to assist process designers. Specifically they propose to model a set of control objectives in the BP that will allow process designers to comparatively assess the compliance degree of their design as well as be better informed on the cost of non-compliance. A DQ aspect considered in these control objectives is the data integrity.

The cited studies consider different DQ dimensions, which are summarized in Table 1.

**Table 1. Data Quality Attributes identified in BP modelling.**

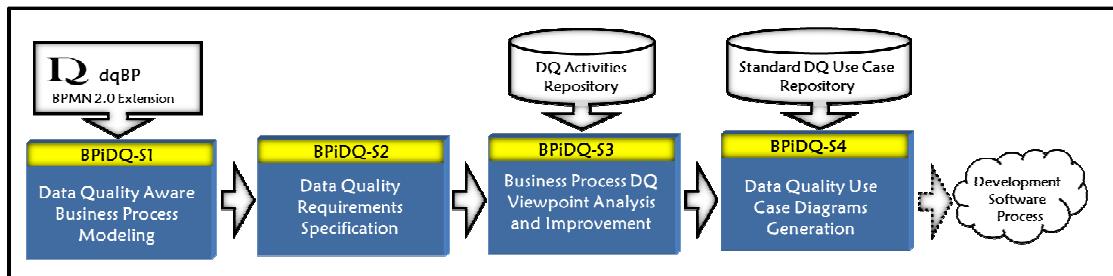
DQ ► Dimension	Integrity	Accuracy	Uniqueness	Completeness	Non-Obsolescence	Consistency	Timeliness	Objectivity	Believability	Reputation	Accessibility	Security	Relevancy	Value-added	Amount of Data	Interpretability	Understandability	Concise Rep.	Consistent Rep.	Easy of Manip.
Work ▼																				
Lu et al. (2000)	x																			
Soffer (2010)		x																		
Bringel et al. (2004)	x		x			x	x	x	x	x	x	x	x	x	x	x	x	x	x	
el Abed (2011)	x	x	x	x	x	x														
Heravizadeh et al (2008)	x		x			x	x	x	x	x	x	x	x	x	x					

### 3. BPiDQ: A METHODOLOGY TO DESIGN DQ-AWARE BUSINESS PROCESSES

A methodology is generally defined as a guideline for solving a problem, with specific components such as phases, tasks, methods, techniques and tools [22]. We propose BPiDQ, a methodology to support the modeling and design of data quality-aware business processes and the generation of DQ requirements for the software development.

BPiDQ uses BPMN as BP modeling language and works with different models in the two first out of three levels of abstraction of BPMN [23]. Such levels are: (i) the *Descriptive level*, which uses the basic set of shapes and symbols that are adequate for the needs of business people seeking to document a process; (ii) the *Analytical level*, in which the full set of shapes and symbols can be used to deal with events and exception handling showing the complexity and depth of the process; (iii) the *Executable level*, which deals with the XML language underneath the shapes.

As shown in Figure 1, BPiDQ is composed of four stages. The first stage (BPiDQ-S1) in the BPMN Descriptive level, starts by introducing high-level DQ requirements into the BP model. In our work we have defined as high-level DQ requirement a mark included into a shape of a BPMN element to highlight a point where the DQ is necessary for the BP success. In the second stage (BPiDQ-S2) the high-level DQ requirements will be refined in order to generate low-level DQ requirements. In our work a low-level requirement is a detailed specification that included among others: the data involved and a set of relevant DQ dimensions. In the third stage, (BPiDQ-S3) in the BPMN Analytic level, the DQ requirements will guide the data quality-aware BP improvement that will imply the addition of new activities or the modification of the process flow. Finally, the fourth stage (BPiDQ-S4) supports the generation of use case diagrams to specify DQ software requirements.

**Figure 1. Methodology to design data quality-aware business processes**

Also, Figure 1 shows that BPiDQ uses three basic components to support the stages: (a) dqBP, a BPMN 2.0 extension to include high-level DQ requirements in a BP model, (b) a repository of DQ activities to improve the BP from DQ point of view, and (c) a repository of standard DQ use cases to specify DQ software requirements. Such components will be described in the following sections together with a detailed description of the BPiDQ's stages.

### 3.1 Components to support BPiDQ

In the following, three components used in BPiDQ to model and design data quality-aware business process and to generate DQ requirements for the software development are introduced.

#### 3.1.1 dqBP: A BPMN 2.0 extension to support BPiDQ

Various elements of BPMN are used for data representation (e.g., Data object or Message). However, aspects related to data quality cannot be included in this kind of elements using the BPMN language. Thus, to support the first stage of BPiDQ and to fill this gap, we have introduced an extension of BPMN 2.0, named dqBP that enriches the BP modeling with DQ requirements [5]. The high level DQ requirements will be modeled in a BP model by means of a set of flags, named DQ Flags. The DQ Flags may be associated with the BPMN data-related elements (Data Objects, Message, Message flow, Conversation, Data Store, and Activity) to mark that they are susceptible to be linked to special data quality requirements. We have also defined the symbol **DQ**, coming from merging letters D and Q, to perform the marking of these BPMN elements. Consequently, such symbol must be included into the shape of the BPMN data-related element in order to show that the quality of data in that specific point of the process is crucial for the business. Table 2 shows a description of these BPMN elements and their graphical representation.

**Table 2: Representation of BP data-related elements enriched with the DQ flags**

Data-related BPMN element	Graphical Representation	Intended use of the Graphical Representation
<b>Message:</b> Content of a communication between two participants. May be data structured or unstructured.		It represents that data contained in the message might satisfy some DQ requirements for the sake of the business success, e.g. Completeness and Consistency in a drug prescription from the doctor to a patient.
<b>Message flow:</b> It shows the flow of Messages (explicit with a Message or implicit without the Message) between two Participants		It represents that data implicitly contained in the message (the message does not appear in the flow) might satisfy some DQ requirements to develop successfully the BP, e.g. Timeliness for a credit card authorization from the bank.
<b>Conversation:</b> Logical grouping of Message exchanges (Message Flows) that can share a Correlation. Conversation has the data contents in the messages included on it.		It represents that data in some messages contained in the conversation might satisfy some DQ requirements for the sake of the success of the business process, e.g., Security and Accuracy of the data interchanged between a customer and an airline Web application during the flight booking process.
<b>Data Object:</b> Primary construct for modeling data within the Process flow in BPMN. It can represent a singular object or a collection of objects, input data or output data.		It represents that data in the data object might satisfy some DQ requirements to successfully achieve the goals of the business process, e.g. Completeness, Consistency and/or Accuracy of the data required to successfully deliver and ordered package to a customer.
<b>Data Store:</b> It provides the necessary mechanisms for Activities to retrieve or update stored information that will persist beyond the scope of the Process.		It represents that data contained in a data store might satisfy some DQ requirements for the sake of the success of the business process, e.g. Checking the completeness of the data updated about product sale.
<b>Activity:</b> Work that is performed within a Business Process. The activity's work may be the generation/processing of data.		It represents that used/produced data in the activity might satisfy some DQ requirements to the business success, e.g. Checking the Precision and Accuracy of the budget generated as the output of one activity.

Figure 2 shows graphically the extension in BPMN 2.0 and the metamodel proposed to support the specifications of the DQ requirements for each DQ Flag in a BP model. In white color, Figure 2 illustrates some classes from BPMN 2.0: (a) the extension metamodel classes (Definition and Extension) from where is derived our proposal, and (b) a set of BPMN classes related with dqFlag class (our extension). In the same figure, in grey color, the metamodel that will support the derivation of DQ requirements from DQ Flags in the BPMN model is showed. More details about the extension can be found in [5].

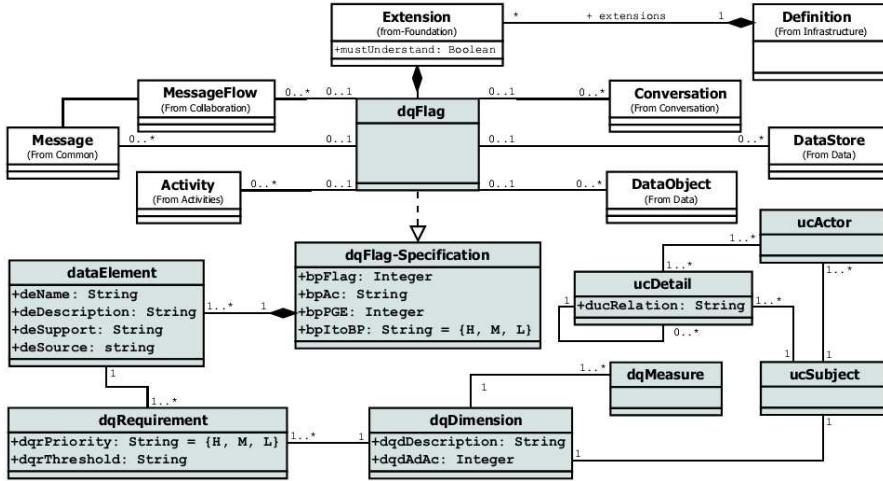


Figure 2. BPMN 2.0 extension and metamodel to obtain of DQ requirements from a BPMN model.

### 3.1.2 A Repository of DQ Activities to tackle DQ Requirements

In BPiDQ, the DQ requirements are expressed by means of some DQ dimensions. The DQ field provides different DQ models (generic set of DQ Dimensions) in order to address DQ concerns in different contexts [2]. Anyway, these DQ models have to be interpreted, and adapted to better fit in specific context. In our work we have decided to consider the most referenced DQ dimensions in the BP literature (see Table 1) to define a repository with DQ activities to tackle each one of them. Thereby, BPiDQ aims to enrich the BP model including a set of DQ activities, obtained from the repository, to tackle the DQ requirements. Table 3 shows some commonly used DQ dimensions and some DQ activities to be performed in order to guarantee that DQ requirements are satisfied within the considered Business Process. Note that these activities have been defined in a generic way and they need to be customized on the basis of the analyzed process and its corresponding context.

DQ Dimension	Definition	Improvement Activities	Examples
Accuracy	The extent to which data reflects a real-world view within a context and a specific process [1, 18, 20].	<ul style="list-style-type: none"> <li>- Determine the data set, which requires accuracy.</li> <li>- Verify data provided against the right domain.</li> <li>- Verify data coming from alternatives sources.</li> <li>- Clean database to achieve the required level of accuracy.</li> </ul>	<ul style="list-style-type: none"> <li>- The price received by the client for a booking hotel must be accurate.</li> <li>- In a medical prescription, the name of the medicines can be confronted with the Vademecum.</li> <li>- The weight of a package to be delivered must be contained within a specific range of values.</li> </ul>
Timeliness	The extent to which data are sufficiently updated for the context and a specific process [1, 19, 20].	<ul style="list-style-type: none"> <li>- Verify if data have the required age for the task.</li> <li>- From different sources, select the one providing data with the age required for the process.</li> <li>- Check if data are delivered within the required time.</li> </ul>	<ul style="list-style-type: none"> <li>- Check if the same data are in different company's source and if it is closer to the right age required, and then take values from this source.</li> <li>- Bank's response to check a credit card must be lower than 5 seconds.</li> </ul>
Completeness	The extent to which data have all values necessary for a successful execution of a process in a specific domain and context [1, 19, 20].	<ul style="list-style-type: none"> <li>- Specify which data are mandatory</li> <li>- Verify/Ensure whether all mandatory items of data have values.</li> <li>- Complete data provided with other sources of data.</li> <li>- Use a procedure to force the delivery of all mandatory data.</li> </ul>	<ul style="list-style-type: none"> <li>- Check if the same data are in different company's and then complete the golden register</li> <li>- To deliver a package, all data about the address and customer identification must be complete.</li> </ul>

Table 3. Example of improvement Activities associated with DQ dimensions.

### 3.1.3 A Repository of Standard DQ Use Cases to tackle the DQ requirements

Taking into account that the BP will be supported by an information system, BPiDQ supports the generation of a set of use cases to represent the DQ requirements for the application to develop. Due to this reason, we have introduced as the third component of BPiDQ a repository that contains a set of standard use cases for each DQ dimension. The use cases have been customized and defined by considering: (a) the definition of each DQ dimension, (b) the set of DQ activities to address each DQ dimension (the previous component), and (c) knowledge previously extracted from existing literature contributions or/and from software developers experience. The idea is that based on these standard use cases the workers could specify a final use case version according to the BP modeled as it is explained in the following subsection. The standard DQ use cases do not consider specific associated actors because they must be specified in the final use case diagram (that will represent the requirements of the application that will support the BP) as «include» use cases for the use cases that will have interaction with the real actors (system's users). As an example, Table 4 shows some standard DQ use cases for the DQ dimensions for accuracy (part a) and for completeness (part b).

<p><b>Accuracy</b></p> <ul style="list-style-type: none"> <li>Verify data with reference values from a database</li> <li>Verify data with other sources</li> <li>Verify accuracy of data by means of procedure</li> <li>Check data with the valid domain</li> <li>Transform data using a master data</li> </ul> <p>(a) DQ dimension of Accuracy</p>	<p><b>Completeness</b></p> <ul style="list-style-type: none"> <li>Verify which all mandatory data have values</li> <li>Verify missing values in other sources</li> <li>Complete data with values in other sources</li> </ul> <p>(b) DQ dimension of Completeness</p>
---	--

Table 4. Some standard use cases for Accuracy and Completeness DQ dimensions

## 3.2 BPiDQ's Stages

In this section the four stages of BPiDQ, detailing the workers involved, component used, input and outputs for each one of them will be described.

### 3.2.1 BPiDQ-S1: Data Quality-Aware Business Process Modeling

This stage is devoted to capture *high level DQ requirements* at a BPMN Descriptive Level [23]. Such requirements are specified by Business People/Analysts and are graphically expressed by means of a specific mark called DQ Flag. Figure 3 shows graphically this stage highlighting the involved workers, inputs and outputs.

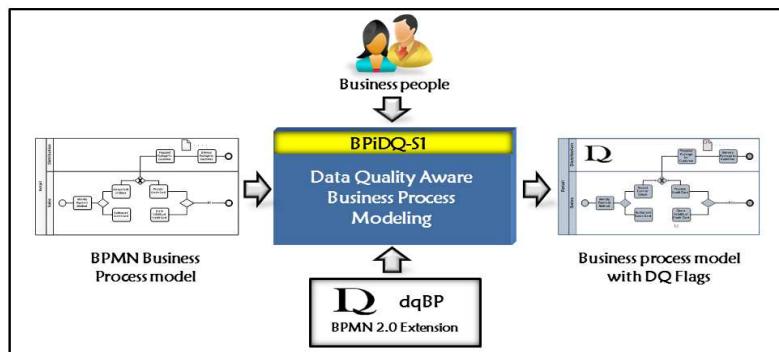


Figure 3. BPiDQ-S1: Data Quality-Aware Business Process Modeling

- As showed in Figure 3, the workers start modeling a new BP or analyzing an existing one. The result of this stage is the BP enriched with a set of DQ Flags that highlight the points of the BP where the DQ is considered essential for the business success. The main activities involved in this stage are:
- **BPiDQ-S1.1. Enrichment of BP model with DQ Flags.** Workers model a BP in the traditional way or start analyzing a BP model created previously. Using the dqBP extension, workers place DQ Flags for some data-related BPMN elements where they think that some DQ management activities are necessary to warranty the BP success.
- **BPiDQ-S1.2. Registration of additional information about the BP and DQ Flags.** Some additional information must be registered by means of text annotations. In particular: (a) business people must include the identification of each data element contained in the data-related BPMN elements marked with a DQ Flag, and (b) an estimation of the level of influence of each DQ Flag in the overall success of business process ranged as {"Low", "Medium", or "High"}.

It is important to note that this stage is supposed to be performed by business people. Generally speaking, they are not expert in technical issues, but they are expert in their own business processes. Thus, in this stage, our aim is to provide adequate mechanisms to express in a simple way the DQ necessities for a specific BP.

### 3.2.2 BPiDQ-S2: Data Quality Requirement Specification

In the second stage, the involvement of Business Analyst/Designer and also the DQ Expert is required. These workers should work together to analyze the modeled BP from a DQ point of view. Figure 4 shows graphically the involved workers, inputs and outputs of this stage.

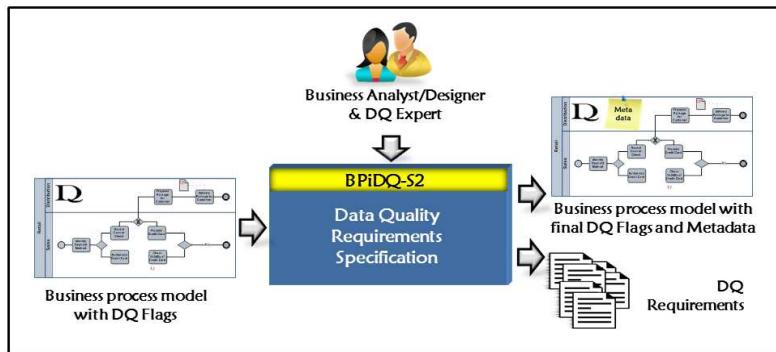


Figure 4. BPiDQ-S2: Data Quality Requirement Specification

Taking as input product the artifact generated in the previous stage (a BP model enriched with DQ Flags), this stage is dedicated to specify *low level DQ*. Workers must review and analyze each DQ Flag in order to make a more refined and complete specification of the DQ requirements related with each one of the DQ Flags. The main activities involved in this stage are:

- **BPiDQ-S2.1. Collection and registration of metadata about the BP and the DQ Flags included on it.** This metadata can be provided by experts and/or by software applications (when the BP is actually implemented and working). We have defined three types of metadata which should be collected:
  - *Metadata about the BP flow* provide information related to the BP control flow, and thus some metadata about the execution of certain activities in a process. For example, it should be useful to know the execution probability of each path on the BP. The range for the probability value is greater or equal than 0 and less or equal than 1. This probability could be estimated by the business analyst from previous executions of the BP, or it may be calculated taking into account the alternative paths drawn by the gateways in the BP.
  - *Metadata about the BP Performance* refer to the performance conditions or constraints within

process flows. This metadata can be defined either at the process or at the task level. In both cases, the metadata store data about temporal conditions (e.g., maximum time that may be needed to respond to a request).

- *Metadata about Data* provide information regarding the data used throughout a process. For example, for each DQ Flag, and for each data element on it, the corresponding metadata must be registered: BPMN element associated with the DQ Flag, DQ Flag's path, previous and posterior activities associated to the DQ Flag, data element description, support (electronic/manual, etc.), source (internal/external), actions of use (use, creation, modification, etc.), the data volatility (i.e., permanent or transient information), and some other points of the BP where the same data is used, to name a few.
- **BPiDQ-S2.2. Specification of the DQ requirements for each data element on a DQ Flag.** For each data element, a set of DQ dimensions along with their corresponding level of importance ranged typically as {"Low", "Medium", "High"} must be identified. The DQ experts must study dependencies between the DQ dimensions associated with each data element to decide if any of them can be eliminated (e.g., to be incompatible with other). This decision must also consider the importance given to the DQ dimensions.
- **BPiDQ-S2.3. Refining the set of DQ Flags in the BP.** Taking into account some of the metadata described previously, the specified DQ requirement, and the cost to satisfy the DQ requirements, the workers must decide the final set of DQ Flags. For this decision the following information is needed:
  - The level of influence of each DQ Flag on the overall success of the BP (registered in the first stage).
  - Probability of execution of the path in which each DQ Flag is placed (obtained from the metadata about the flow).
  - DQ Flag overhead, defined as the ratio between the number of new activities that has been added to tackle with the new DQ requirements (one activity for each DQ dimension) and the total number of activities in the BP. This factor shows the overhead relative of each DQ Flag in the BP.
  - Business constraints (obtained from metadata about performance).

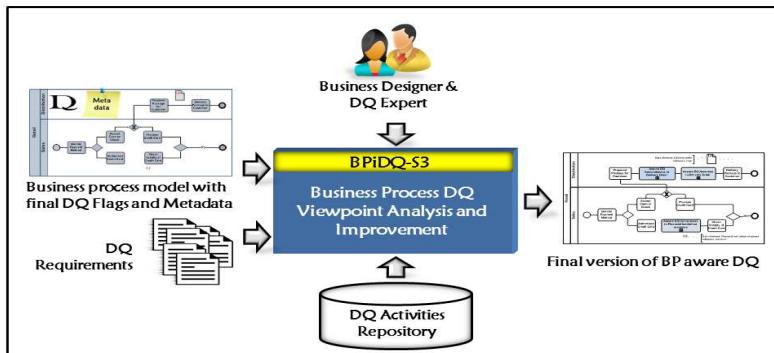
For example, if a DQ Flag has (i) a grade of influence higher than another one, (ii) a higher probability to be executed than another DQ Flag, and (iii) a medium overhead in the BP, then, the first DQ Flag is considered more important and could have more probabilities to be addressed.

Finally, the dependencies between the data elements in the same BP branch have to be studied, (for example to eliminate some redundant DQ Flags).

As a result of this stage, the final configuration of DQ Flags for the BP model should be released. Also, the documentation about all DQ Flags (data-related BPMN elements and data elements associated), and the specification of DQ requirements for it (in low level) should be generated.

### 3.2.3 BPiDQ-S3: Business Process DQ viewpoint Analysis and Improvement

The third stage is devoted to analyze and decide the most suitable way to improve the BPMN model from the DQ point of view. This stage is executed at the BPMN analytic level [23], and the workers involved are the Business Designer and DQ Expert. Figure 5 shows graphically the workers, inputs and outputs of this stage.



**Figure 5. BPiDQ-S3: Business Process DQ viewpoint Analysis and Improvement**

As showed in Figure 5, in this stage the workers generate the final version of the BP model. This version will include the modifications needed to address the DQ requirements and defined taking into account the metadata associate with the BP to decide the best solution. In this stage, the following activities are executed:

- **BPiDQ-S3.1. Selection of the improvement actions for satisfying the low-level DQRequirements.** Considering the DQ requirements in each DQ Flag and some metadata, a set of new activities should be selected to assure the adequate level of DQ for the data in the BP. Thus, the BP model will be enriched with the inclusion of new activities in order to avoid some DQ problems and consequently to minimize the risk due to poor data quality. The activities will be provided from the DQ activities Repository (component of BPiDQ previously explained) that contains a set of DQ activities for each DQ dimension and considering the use of data (creation, use, modification).
- **BPiDQ-S3.2. Improvement of the BP model to satisfy the DQ requirements.** Taking into account the DQ requirements, the flow of the BP and the set of new DQ activities selected to be included in the BP model, workers must study how to change the BP model in order to assure the most appropriate configuration to satisfy the DQ requirements. This means:
  - o Generate alternative BP models, which integrate the DQ activities to satisfy the DQ requirements in the BP flow. For example, alternative models depending on the actions to develop where a DQ problem raises may be generated, one of them to abort the execution or another one to develop some actions and follow the execution.
  - o Study the BP flow to decide whether it is necessary a redefinition of it in order to satisfy some DQ requirements. For example, if two sequential activities are independent between them, then, they can be executed in parallel in order to improve the time-related data quality dimensions.
  - o Evaluate the proposed alternatives and select the most suitable one that better satisfy both data quality requirements and the business objectives. A cost-benefit analysis must be conducted, considering the costs of the implementation, the user satisfaction, the success of the BP, etc.

In this stage, the final version of the BP will be released. This stage works with the BP model at the BPMN analytic level what allows modeling the BP with more details than at the BPMN descriptive level. Thus, considering the granularity of the activities in this level, we have decided to generate the model with two levels of details. The first one will include, for each DQ Flag, a set of collapsed sup-processes. Each one of these sub-processes represents a DQ dimension that must be assured for the data element involved in the DQ Flag. The second, for each one of these DQ collapsed sub-processes, in a lower level of detail, will include an expanded Sub-Process that contains all the activities selected to assure the corresponding DQ dimension.

### 3.2.4 BPiDQ-S4: Data Quality Use Case Diagram Generation

The common next step for the business process modeling, considering for example an MDA approach, is the development of software to support it. Thus, the fourth stage of BPiDQ represents a first approach toward the definition of requirements for developing applications able to satisfy the specified DQ requirements. For doing so, we provide the support by means of the generation of a set of use cases that represent the requirements related with the activities in the BP that tackle the DQ expressed like DQ Flags. Figure 6 shows graphically the workers, inputs and outputs of this stage.

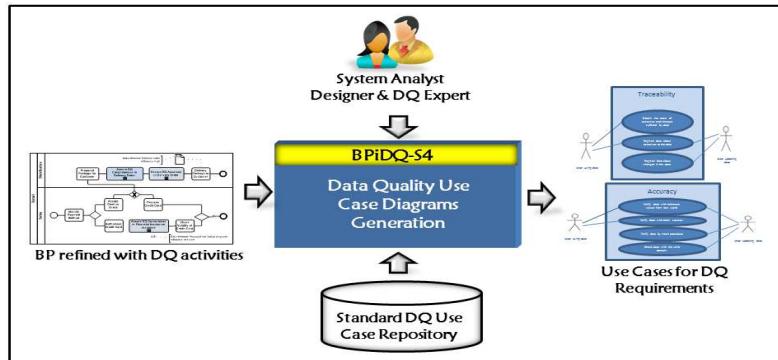


Figure 6. BPiDQ-S4: Data Quality Use Case Diagram Generation

Figure 6 introduces the involved workers: System Analyst and DQ Expert. Based on the BP refined with DQ activities and a set of standard use cases (from the repository explained in section 4), these workers will instantiate a set of use case diagrams customized for the specific BP. The following activities are to be executed in this stage:

- **BPiDQ-S4.1. Generation of use case diagrams based on the DQ requirements.** For each DQ collapsed sub-process incorporated in BP, the workers will select the appropriate use cases from the standard use case repository. Indeed, they must select the use case based on the DQ dimension related, and the activities contained in the expanded Sub-Process.
- **BPiDQ-S4.2. Customization of the use cases with the specific BP.** The workers will refine the DQ use cases diagrams generated, customizing the use cases with the BP. From the swimlanes, they can identify the actors. From metadata, they can also identify the data elements to be manipulated, the action developed with the data (creation, elimination or use), the sources of data, etc.

Thus, in this stage a set of use cases diagram available of input for the software development will be produced. Our intention is that the developers can be aware as early as possible of the DQ requirements for the BP, and they have a set of seminal use cases to implement the software considering DQ issues.

The following section illustrates the use of BPiDQ by means of an example.

#### 4. AN EXAMPLE OF THE APPLICATION OF BPiDQ

Let us consider the process of payment and delivery of the ordered products. The description of the BP of this example starts with the payment phase. The payment can be processed in two different ways: by credit card or by cash (or check). If payment is made by credit card, it is necessary to ask for card authorization to the «Financial Institution». If the credit card payment is not authorized, then, the process finishes. If the payment is performed by cash (or check), no controls are needed. When the payment is complete, the Distribution Department prepares the package and delivers it to the customer, and after this, the process ends. The remainder of this section is devoted to demonstrate how the proposed methodology can be applied.

In the first stage (“*BPiDQ-S1.Data Quality Aware Business Process Modeling*”), business people must identify which data-related BPMN elements could be susceptible to be linked to DQ Flags. In our example, two DQ Flags are defined. The first one, named DQFlag1, is associated with the Data Object needed

as input in the “*Delivery package to customer*” activity (see Fig. 7). This Data Object contains the *Delivery Order* with the customer information necessary to deliver the package (identification, address). The second DQ flag, named *DQFlag2*, is associated with the Message Flow coming from the Financial Institution pool to Sales lane. This Message Flows contains the *Financial Institution response* a message with the authorization or the rejection to process the payment with customer credit card. As output of this stage, the business process model shown in Figure 7 is generated and enriched with symbol for DQ Flag. In this figure it is possible to see how the data-related BPMN elements have been marked with the special symbol  $\text{DQ}$ . Also, workers registered the additional information about DQ Flags, data-element identified and their influence in the BP success.

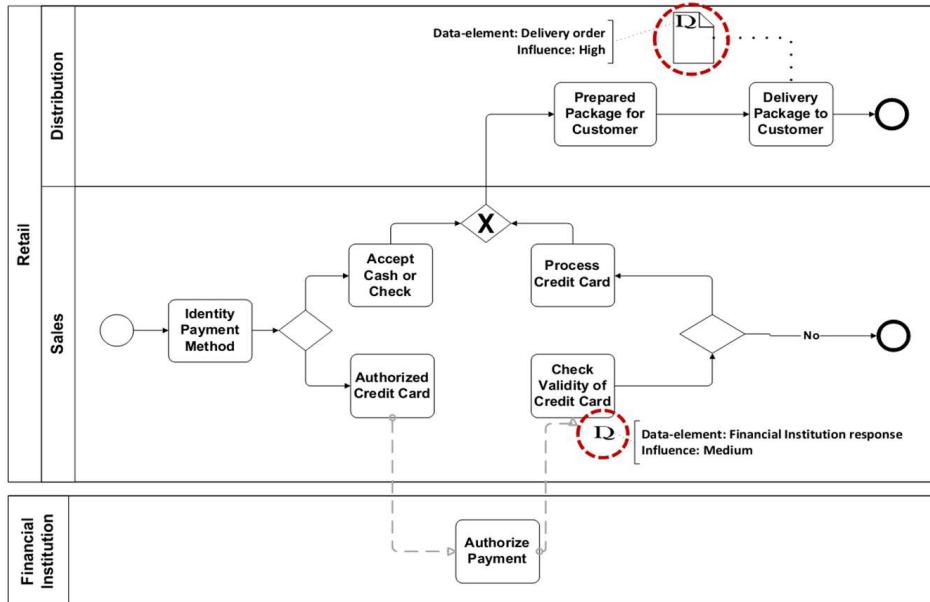


Figure 7. Illustrative example: BPMN model with DQ Flags.

In the second stage (“*BPiDQ-S2. Data Quality Requirements Specification*”), the workers (Business Analyst/Designer and DQ Expert) register metadata about the BP and DQ Flags. They also reviewed each one of the DQ Flags to specify the corresponding low level DQ requirements. For performing the realization of the *DQFlag1*, the definition of a *DQFlagSpecification1* is required. Therefore, they must define the DQ dimensions, and their importance. DQ requirements for *Delivery Order* involve two DQ dimensions, which are considered as relevant for the BP: Accuracy and Completeness. On the other hand, for the *DQFlag2*, the *DQFlagSpecification2* is defined and DQ Requirements for “*Financial Institution Response*” consider the DQ Dimension Currentness. In addition, for the two DQ Flags the probability of execution and overhead (and some other information) are obtained and/or calculated. Most important details about both DQ flags specifications are shown in Table 5. Taking into account the available information, the workers must decide the definitive set of DQ dimensions for the data elements in each DQ Flag. Besides, they must decide the final set of DQ Flags. In our example, *DQFlag1* has a *High* impact on the success of the business process. Even if they have not any initial knowledge on the process execution, the estimated probability of execution of the delivery action is 75% because the BP flow shows (taking into account the exclusive gateways) that in some cases the activity related with the DQ Flag may be not executed.

Table 5. DQ Flags specifications

DQFlagSpecification1		DQFlagSpecification2	
BPMN data element	Data Object	BPMN data element	MessageFlow
Influence	High	Influence	Medium
Probability Exec.	75%	Probability Exec.	50%
Overhead	2/8*100=25%	Overhead	1/8 * 100 = 12.5%
<b>Data Quality Requirement</b>	<b>Data Elements</b>	<b>Data Elements</b>	
Name	Delivery Order	Name	Financial Institution response
Description	Delivery order (customer information)	Description	Message from the Financial Institution
Support	Electronic	Support	Electronic
Source	Internal	Source	Internal
<b>DQ Requirements</b>		<b>DQ Requirements</b>	
Accuracy	High	Currentness	High
Completeness	Medium		

(a)

(b)

The overhead associated with this DQ Flag is 25% because in order to tackle with the DQ requirements, two new activities must be included in the process (see in grey colour, the new activities in the left side of Figure 8).

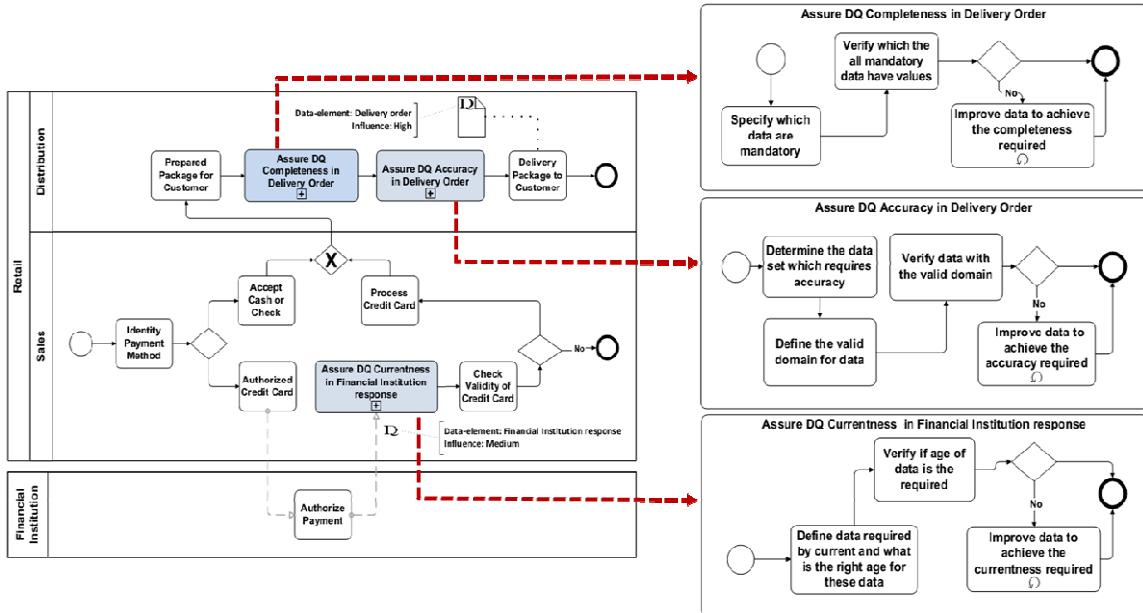


Figure 8. BP model improved.

*DQFlag2* has a *Medium* impact on the success of the business process. The probability of requesting the payment authorization is 50% because when the payment is not performed by credit card the activity related with the DQ Flag is not executed. The overhead associate to this DQ Flag is 12.5% because to tackle the DQ requirements must be included a new activity in the process (see, in grey colour, the new activity in Figure 8 (left side)). Finally, since the data elements associated with each DQ Flag are crucial for the business process success, the workers decided to implement the improvement actions for both DQ Flags. Note that in this stage the BP is modified including a new activity (collapsed sub-process) for each DQ dimension in each DQ Flag point (BPMN Descriptive Level).

In the third stage (“*BPiDQ-S3. Business Process DQ Viewpoint Analysis and Improvement*”), the business process designer and DQ Expert must decide which specific DQ improvement activities should be adopted. First, and considering each DQ dimension to engage, the use of the data elements and the necessary information recollected, they must select from the repository the most suitable activities. After this, workers must evaluate the possible alternatives to integrate these new activities in the BP. In our example, the activities selected and their flow is showed in Figure 8 (right side). Note that in this stage the

collapsed sub-process are replaced for expanded sub-process, considering a more detailed level in the BP model (BPMN Analytic Level). Finally, in the fourth stage (“*BPiDQ-S4. Data Quality Use Case Diagrams Generation*”) the Use Case diagrams which specify the DQ requirements for the software that will implement the improved BP model must be generated. Thus, in our example, the standard use cases for each DQ Flag and the corresponding DQ requirements were firstly selected. After this, the workers customized the use cases the BP modelling. Figure 9 shows the use case diagram generated for the requirements related with the DQFlag1.

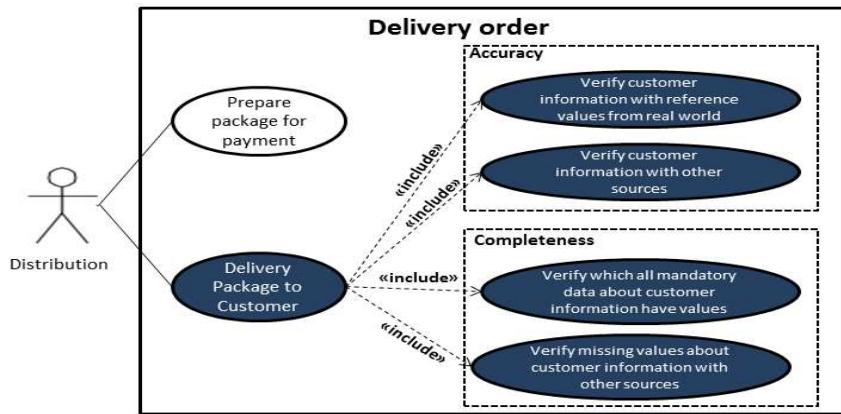


Figure 9. Use case diagram for the BP model.

The use case diagrams generated in this stage are generic, and they constitute a very first approach towards the software development. In our opinion, although in a simple way, this example demonstrates that our methodology is useful to involve the business people since the earliest definition of the DQ requirements of their business process.

## 5. CONCLUSIONS AND FUTURE WORK

Poor data quality has severe impacts on the performance of an organization. Most of the organizations are aware about data quality issues, but frequently, they do not have a proactive attitude to address the DQ problems before their apparition. To this aim, in this paper we have presented the BPiDQ methodology that is oriented to support the modelling and design of data quality-aware business process and the generation of DQ requirements for the software development. BPiDQ allows business people to include DQ needs in business process modeling using DQ Flags. Then, for each one of these DQ Flags, BPiDQ allows workers to specify DQ requirements that will drive improvements over the original BP model in order to guarantee the DQ level required. Furthermore, the methodology supports the specification of use cases for the data quality-aware software development. Our future work will focus on three different goals: (a) Conduct some more case studies to obtain the opinion and feedback of the different workers involved, (b) Build a tool to support the methodology allowing the automatic development of some activities, and (c) Refine the methodology stages in order to better support the process improvement.

## Acknowledgments

This research is part of the following projects: MECESUP (UBB0704) and IQMNET (TIN2010-09809-E).

## REFERENCES

- [1] el Abed, W., *Data Governance: A Business Value-Driven Approach*. 2009.
- [2] Wang, R. and D. Strong, *Beyond accuracy: What data quality means to data consumers*. Journal of

- Management Information Systems; Armonk; Spring. 12(4). 1996, pp. 5-33.
- [3] Bagchi, S., X. Bai, and J. Kalagnanam. (2006). *Data quality management using business process modeling*. pp. 398-405.
  - [4] Harmon, P. and C. Wolf (2011) *Business Process Modeling Survey*. Business Process Trends (<http://www.bptrends.com/>).
  - [5] Rodriguez, A., A. Caro, C. Cappiello, and I. Caballero. (2012). *A BPMN extension for including data quality requirements in business process modeling*. In *4th International Workshop on the Business Process Model and Notation*.
  - [6] Redman, T., *Data Driven*. 2008: Harvard Business School Press.
  - [7] Shankaranarayanan, G., R.Y. Wang, and M. Ziad. (2000). *Ip-map: Representing the manufacture of an information product*. In *Fifth International Conf. on Information Quality (ICIQ'2000)*. pp. 1-16.
  - [8] Recker, J., *Opportunities and constraints: the current struggle with BPMN*. Business Process Management Journal. 16(1). 2010, pp. 181-201.
  - [9] Saeedi, K., L. Zhao, and P.R. Falcone Sampaio. (2010). *Extending BPMN for Supporting Customer-Facing Service Quality Requirements*. In *Proceedings of the 2010 IEEE International Conference on Web Services* pp. 616-623.
  - [10] Auer, D., V. Geist, and D. Draheim. (2009). *Extending BPMN with Submit/Response-Style User Interaction Modeling*. In *IEEE Conference on Commerce and Enterprise Computing*. pp. 368-374.
  - [11] Bocciarelli, P. and A. D'Ambrogio. (2011). *A BPMN extension for modeling non functional properties of business processes*. In *Proceedings of the 2011 Symposium on Theory of Modeling & Simulation: DEVS Integrative M&S Symposium*. pp. 160-168.
  - [12] Friedenstab, J.-P., C. Janiesch, M. Matzner, and O. Müller. (2012). *Extending BPMN for Business Activity Monitoring*. In *Proceedings of the 45th Hawaii International Conference on System Sciences* pp. 4158-4167.
  - [13] Gagne, D. and A. Trudel. (2009). *Time-BPMN*. In *IEEE Conference on Commerce and Enterprise Computing*. pp. 361 - 367.
  - [14] Gao, F., M. Zaremba, S. Bhiri, and W. Derguerch. (2011). *Extending BPMN 2.0 with Sensor and Smart Device Business Functions*. In *IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises*. pp. 297 - 302
  - [15] Rodríguez, A., E. Fernández-Medina, and M. Piattini, *A bpmn extension for the modeling of security requirements in business processes*. IEICE transactions on information and systems. 90(4). 2007, pp. 745-752.
  - [16] Goldner, S. and A. Papproth. (2011). *Extending the BPMN Syntax for Requirements Management*. In *Business Process Model and Notation*. pp. 142-147.
  - [17] Lodhi, A., K. Veit, and G. Saake, *An Extension of BPMN Meta-model for Evaluation of Business Processes*. J. Riga Technical University. 432011, pp. 27-34.
  - [18] Soffer, P., *Mirror, mirror on the wall, can i count on you at all? exploring data inaccuracy in business processes*. Enterprise, Business-Process and Information Systems Modeling. 2010, pp. 14-25.
  - [19] Bringel, H., A. Caetano, and J. Tribolet. (2004). *Business Process Modeling Towards Data Quality Assurance*. In *6th International Conference on Enterprise Information Systems*. pp. 565-568.
  - [20] Heravizadeh, M., J. Mendling, and M. Rosemann. (2009). *Dimensions of business processes quality (QoBP)*. pp. 80-91.
  - [21] Lu, R., S. Sadiq, and G. Governatori, *On managing business processes variants*. Data & Knowledge Engineering. 68(7). 2009, pp. 642-664.
  - [22] Klein, H.K. and R. Hirschheim, *Choosing Between Competing Design Ideals in Information Systems Development* Information Systems Frontiers. 3(1). 2001, pp. 75-90.
  - [23] Silver, B., *BPMN Method & Style: A levels-based methodology for BPM process modeling and improvement using BPMN 2.0*. 2009: Cody-Cassidy Press.

# **KNOWLEDGE ACQUISITION FROM AND SEMANTIC VARIABILITY IN SCHIZOPHRENIA CLINICAL TRIAL DATA**

(Research Paper, IQ Assessment, Policies, and Standards)

**Meredith Nahm, PhD**

Duke University Center for Health Informatics, Durham, NC

[meredith.nahm@duke.edu](mailto:meredith.nahm@duke.edu)

**Abstract:** Recent federal requirements in the United States mandate sharing of research data, meaningful use of health information technology, and data standardization for regulatory review of marketed therapeutics. These requirements are predicated on the assumption that both healthcare organizations and the public will benefit from the enhanced secondary use of healthcare data. Because necessary standards are lacking across most clinical therapeutic areas, large-scale efforts are underway to create authoritative, consensus-based, and publically available standard data element sets. Knowledge acquisition is a key component of such efforts to improve information quality through decreasing semantic and syntactic variability in clinical data, i.e., data standardization. The extent and impact of semantic variability has not previously been rigorously assessed in clinical research. Such a characterization informs data standardization efforts and provides metrics to support data governance efforts. This article reports 1) evaluative data describing a potentially more scalable process for the knowledge acquisition, synthesis and definitional aspects of data element standardization and 2) characterizes the semantic variability component of information quality in data from pivotal clinical trials in schizophrenia.

Semantic variability in clinical trials for Schizophrenia compounds recently reviewed for marketing authorization was substantial, implicating semantic variability as a key information quality problem in secondary use of clinical research data. Based on the relatively high proportion of data elements that the synthesis and clinical review process marked for deletion, an appreciable amount of the semantic variability was unnecessary. The form-based knowledge acquisition method used achieved 95% domain coverage as adjudicated by clinical experts and outperformed knowledge acquisition from experts. Within mental health, form-based knowledge acquisition appears to provide a feasible production scale for data element standardization.

**Key Words:** Data Quality, Information Quality, Data Standards, Data Elements, Data Governance, Knowledge Acquisition, Clinical Research

## INTRODUCTION

One of the most fundamental questions in biomedical informatics is how to represent data, information, and knowledge in ways that allow them to be exchanged and used by computers and humans [1-3] — in other words sharing unambiguous meaning, or *semantic interoperability*. Current data standards used in clinical research are insufficient to support such exchange and reuse of data. [4-10] Although data may be pooled across clinical studies and exchanged among organizations, semantic and syntactic variations in these data often necessitate extensive and burdensome manual procedures to ensure usefulness. A 2001 study conservatively estimated that data transfers cost the clinical trials industry \$156 million per year, a significant proportion of which is attributable to lack of semantic interoperability. [11] Further, handling data variations often result in difficulty reusing data for research, [4, 12-13] and may cause degradation and loss of information. [14]

In therapeutic development, as in other industries, the data element is the fundamental unit of exchange. As such, the data element is the level at which standardization and metadata governance should occur. A brief summary of data element approaches to information representation, documentation, and exchange in health care has previously been published. [15] Although historically deemed impossible, standardization of data elements to support patient care and clinical decision-making is increasingly considered a part of the solution to the problems of lack of semantic interoperability and poor information quality in health care. The real challenge associated with a data element approach lies not in the usefulness of data elements as such; indeed, most meaningful data exchange and reuse today is based on data elements. Rather, the challenge lies in 1) the large number of data elements in need of standardization, and 2) the investment required standardizing and maintaining them. The latter includes not just representational aspects such as scale, enumeration, data type, and units, but also the time required to obtain authoritative agreement on semantics, including paring potential data elements down to atomic concept(s), identification of semantic matches (e.g., synonymy, or different words with equivalent meaning) and semantically similar terms, including differentiation of the latter. Further, in health care, adoption of standard data elements depends on their value proposition, i.e., their 1) authority, 2) consistency with existing or required data, and 3) benefit to the organization and individual responsible for collecting the data.

Until recently, efforts aimed at standardizing data elements in health care have focused on a specific use; e.g., a research study or disease registry. [15] Two United States National Institutes of Health (NIH)-funded initiatives sought to change this paradigm by including primary and secondary data use stakeholders when defining standard data elements for the purpose of supporting both primary and secondary data uses. These initiatives, conducted in the fields of cardiology (acute coronary syndromes [ACS]) and infectious diseases (tuberculosis [TB]) convened clinical thought leaders from medical specialty societies in each area and worked with international clinical thought leaders and medical specialty societies to identify or create authoritative definitions for data elements, to represent them in a computable format, and to standardize them through an American National Standards Institute (ANSI) accredited and international standards development organization, Health Level Seven (HL7).

Both projects, as well as several others since then, have achieved balloted international standards, but with elapsed times ranging from 1-3 years and productivity ranging from 21 to 139 data elements with up to 300 associated valid values. Given a context of more than 100 medical specialties, thousands of disease areas, and a rapidly developing clinical science enterprise, this is a slow pace. Further, with a cost of such efforts in the neighborhood of \$150,000 per year, [16] most clinical professional societies have not yet sponsored their own efforts. When we consider the tipping point represented by United States federal incentives for “meaningful use” of health information technology that include use of health information to increase the proportions of patients that receive guideline-recommended care, medical specialties have an increasingly compelling reason to pursue data standardization.

In addition, the promise of widely available healthcare data gives secondary data use stakeholders (public health, research, regulatory reporting, therapeutic development, etc.) significant motivation to engage. More scalable processes for identification, synthesis, and ultimately standardization of clinical data elements would benefit all stakeholders, and indeed, public health.

## BACKGROUND

In December of 2010, the Center for Drug Evaluation and Research (CDER) of the United States Food and Drug Administration (FDA) published the initial version of their Data Standards Plan, [13] the purpose of which is to support and promote development of data standards for all key data needed to guide regulatory decision-making. Following publication of the plan, CDER posted 55 (now 58) priority disease/domain areas for data element standardization, [14] and an R24 program announcement entitled *Data Concepts and Terminology Standards for Clinical Research and Drug Development*, which essentially called for standardization of data elements in disease/domain areas of high impact for regulatory decision-making. [16] Work is underway in eight of the priority areas. [14] The FDA has distinguished its approach from predecessors by embracing a Single Source philosophy, also known as collect-once-reuse-many: “Ideally, data requirements for multiple use cases (e.g. healthcare, clinical research, public health reporting, regulatory review) are used to create a “superset” data standard that supports multiple uses of the data,” under the rationale that harmonization between healthcare and secondary data uses can overcome information silos that hamper assessments across a medical product’s lifecycle. [14] While this program is sponsored by a United States FDA, it has far-reaching international impact because a significant portion of patients enrolled in clinical trials submitted for marketing authorization in the United States are enrolled in countries around the world. When participating in clinical trials to be submitted for regulatory review in the United States, international clinical investigational sites are subject to United States regulations. Further, while available resources and tests used to diagnose disease and disorders may vary internationally, much of the information generated and used in patient care remains the same.

Data standards currently used in clinical research are insufficient to support the exchange and reuse of data necessary for regulatory decision-making. Although data can be pooled across clinical studies and exchanged between organizations, variations in meaning, measurement, recording, formatting, and coding systems usually necessitate manual and point-to-point mapping. This in turn can render data inaccessible or altogether unusable, or may lengthen regulatory review processes. For the current data standard for regulatory submission of clinical trial data, the Clinical Data Interchange Standards Consortium (CDISC) Submission Data Tabulation Model (SDTM), two chief problems exist:

- 1) The SDTM today does not provide for unique mapping of some data into the model; i.e., more than one mapping of fields from a data collection form to the SDTM can be conformant, [12] meaning that for some data the current version of the SDTM (v1.2) is underspecified. [12]
- 2) The SDTM v1.2 primarily addresses data that are common across therapeutic areas; e.g., adverse events, demography, vital signs, and physical exams. The SDTM standard today lacks coverage of clinical domain-specific data such as efficacy data that are critical to drug evaluation and regulatory decision-making.

In noting these two problems, we emphasize that this critique should not be interpreted as undervaluing the strong body of work represented by the SDTM. Rather, the problem statement reflects the complex reality of semantic interoperability in health care and clinical research, and provides an indication of the work that lies ahead of us.

Unfortunately, in health care, the needed degree of authority for clinical definitions can only be conferred by the authoritative clinical specialty society, e.g., a working group of experts convened by the authorita-

tive clinical professional society or societies. The most scalable process will likely be one that optimizes the use of such highly skilled resources. Because national or international standards are usually required for widespread adoption, expertise in information and knowledge modeling is also required, as is the open and consensus process of an accredited Standards Development Organization (SDO). While the latter are established, e.g., through Health Level Seven (HL7) the ANSI accredited SDO for healthcare, acquisition of authoritative clinical expert knowledge remains a major challenge in standardization of clinical data elements.

In therapeutic product development, a significant source of expert knowledge exists in the data collection forms used for the clinical trials submitted for marketing authorization; leveraging the knowledge encoded in these forms may decrease the time required from individual experts and clinical professional societies and optimize their involvement in such efforts. However, because the source originates from one particular secondary data use, clinical research, it may or may not reflect the focus on health care that we primarily desire. The amount of semantic overlap between a clinical trial and standard of care for different trial phases is driven by the research goal and design. Briefly, in late-phase research the mechanism of action, efficacy, and gross safety have been established by preceding studies; thus, late-phase research typically involves large, simple trials and observational studies that assess how the new therapy performs “in the wild.” Alternatively, in early-phase research, special data may be collected to confirm mechanistic action, e.g., bronchial biopsy or washings for an experimental asthma drug that are not collected during routine asthma management encounters. Thus, early phase research should be expected to have substantially less semantic overlap than late phase research.

Phase III clinical trials are those upon which marketing authorization decisions are based; the actual trials submitted as the basis or a marketing application are called Phase III pivotal clinical trials. As such, we expect significant overlap with data generated and used in standard care in the therapeutic area. Thus, use of data collection forms from phase III pivotal trials may be a good, but not complete source of candidate data elements for “single source” data standards. Their use for such requires some involvement of therapeutic area experts.

## **METHODOLOGY**

We report empirical observational data from a modified process for knowledge acquisition, synthesis, and definitional aspects of data element standardization. The modified process uses data collection forms from phase III pivotal clinical trials as the source of knowledge. Data elements are abstracted from clinical trial data collection forms and semantically equivalent data elements are synthesized to obtain a set of semantically distinct data elements. Definitions are drafted from form context and the available literature. The data element set is ultimately subjected to international ballot according to HL7 process. Measures of time required, number of data elements defined, and healthcare content coverage were collected from the data element abstraction, synthesis and clinical expert review process. These measures were compared with those from the prevailing knowledge acquisition method, knowledge acquisition from clinical experts, to evaluate the modified process. Importantly, this work provided the opportunity to characterize semantic variability in Schizophrenia clinical trial data.

### ***Data Element Knowledge Acquisition***

New drug approvals for compounds with a schizophrenia indication since 2006 were identified from the FDA New Drug Approvals Database. [17] The data collection forms for pivotal clinical trials used to guide regulatory decision-making regarding efficacy and safety of these compounds were reviewed. Data elements specific to schizophrenia were abstracted by the author using systematic document analysis techniques. Data elements included on validated questionnaires were explicitly excluded because by virtue of inclusion on such an instrument, these data elements are already semantically standardized. Where

questionnaire data elements were used outside of the context of the questionnaire, they were included because such use invalidates the data elements as questionnaire items and is often subject to user modification. For each data element, the data collection form page, form module name, prompt (item) number, prompt text, data format, valid values, and representational notes were listed in a spreadsheet. These included essential attributes of a data element as defined by the ISO 11179 metadata registry standard. [18] For the initial five trials, the entire data collection form was abstracted to assure the categorization of data elements as schizophrenia-specific *versus* not schizophrenia-specific was reproducible; afterward, only new schizophrenia specific data elements were abstracted. The data element acquisition process was systematized so that no decision-making was applied other than categorization of a data element as schizophrenia-specific *versus* not.

### **Data Element Synthesis**

For each trial abstracted, a new set of columns was added to the spreadsheet, semantically equivalent data elements were listed on the same row, semantically similar and semantically distinct data elements were listed on separate rows. In this way, semantic matching was performed during abstraction. Data elements were grouped in adjacent rows on the spreadsheet according to semantic similarity and topicality. Thus, semantically equivalent data elements from different trials were listed on the same row and in a different set of columns. For example, data elements from two different forms that represented the same concept but used different valid values were listed on the same row. In this way, the information content in the form context was preserved, and the abstracted information was condensed into a set of distinct concepts. After all data elements were abstracted, the data elements were reviewed to identify any possible semantic matches that were missed. A draft information model diagram was created as a visual representation of the data elements.

### **Data Element Definition**

Data element definitions were drafted based on the form context, the clinical literature, and National Institutes of Health knowledge sources, e.g., the Medical Subject Headings vocabulary. Using Chisholm's classification of definition types, [19] essential, distinctive, and genetic definitions were preferred over nominal, ostensive, causal, accidental, and stipulative definitions. In mental health, authoritative definitions of psychiatric disorders exist in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition text revision (DSM-IV-TR). Data element definitions explicitly relied upon or referred to the DSM-IV-TR for definitions of disorders and diagnostic criteria. Areas of definitional uncertainty and remaining questions about semantic similarity or the scope of concepts were flagged for review by the Clinical Expert Review Committee (CERC) convened for the schizophrenia data standards development effort. The draft data element set with definitions, data format, and valid values was provided to the CERC for review.

### **Data Collection for Metrics Evaluation and Semantic Variability Characterization**

Time spent on data element abstraction and synthesis was recorded during the abstraction. Following the abstraction, the spreadsheet described in the preceding sections was analyzed to obtain the percentage of therapeutic area specific data elements and the amount of new semantic content added by each trial and form abstracted, as well as reasons for the new semantic content. The concept "new semantic content" was operationalized by the occurrence of a new and semantically distinct data element. This was used to calculate the new semantic content contributed by sequentially abstracted forms. Each occurrence of new semantic content was reviewed and labeled with a statement of the apparent purpose for the new semantic content. Because clinical trials are research experiments, the purpose for variables is usually evident in the variable, e.g., outcome characterization, population characterization, protocol adherence. The statements were coded with categories arising from the data resulting in seven categories.

The data element set was reviewed by a Clinical Expert Review Committee (CERC) of nineteen clinical thought leaders in schizophrenia pharmacotherapeutic development (one representative per organization). Comments were also received from the FDA review division for a total of 20 CERC members. The CERC was convened for the purpose of assuring that the data element set was complete for data generated and used in the diagnosis and treatment of schizophrenia applicable for regulatory decision-making, that the data element set included data elements that were reasonably necessary for regulatory decision-making regarding pharmacotherapeutics with a schizophrenia indication, and that the definitions were clinically accurate. The CERC was provided the list of 86 (expandable to 204) draft data elements as well as a visual representation (a static information model in the style of a Unified Modeling Language [UML] class diagram was used) and instructions for review. The CERC was asked to review the data element definitions and to consider the following questions:

1. Are there any data elements in the list/on the model that are not relevant for regulatory decision making in Schizophrenia?
2. What relevant data elements are missing?
3. Are the valid values for each data element: at the appropriate level of detail, exhaustive and mutually exclusive?
4. Are the relationships (grouping in boxes and associations shown by lines) in the model accurate according to how you relate the data elements with each other; are there any missing?

CERC members recorded their comments on a copy of the data element spreadsheet. CERC comments were collected after the review and each comment was logged. A comment disposition (persuasive, not persuasive, and no change indicated) and disposition action, e.g., update the definition, delete the data element, was recorded for each comment. In this way, metrics on the number persuasive, supportive and not persuasive comments were captured, as was the overall impact on the draft data elements based on the review, e.g., number of data elements added, deleted, and modified.

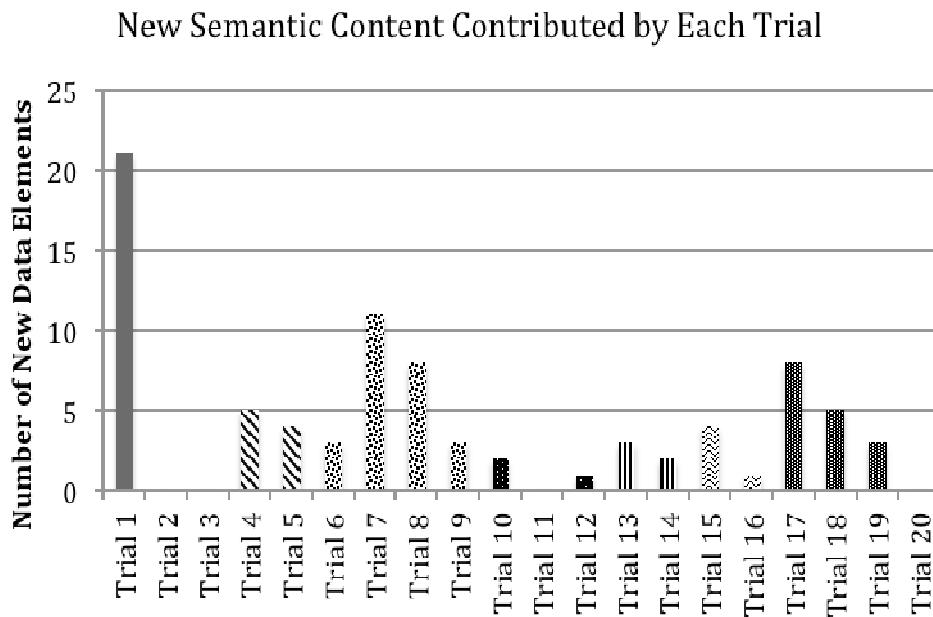
## RESULTS

Data collection forms for seven New Drug Applications (NDAs) approved between 2006 and 2010 were reviewed and abstracted. A total of 20 data collection forms with an estimated 550 semantically unique data elements per form were reviewed, for a total of 11,000 data elements.

### *New semantic content*

The data element synthesis resulted in 86 schizophrenia-specific distinct semantic concepts. The draft set included an additional 118 candidate data elements that were suggested to the CERC for exclusion. These were relevant to schizophrenia, but were either thought to be too detailed or were elsewhere standardized. After synthesis, the 86 core data elements were analyzed to explore the amount of new semantic content added for each additional compound and trial abstracted. As expected, the first trial abstracted contributed the highest number of data elements—twenty-one. Four trials contributed no new data elements. At the onset of the project, an asymptotic effect was expected, wherein each subsequent trial abstracted would add fewer new data elements; however, our actual results contradict this. Figure 1 shows the number of new data elements contributed per sequential trial abstracted; bars with the same fill pattern are from the same compound (from left to right, compounds A-G). One data element was accidentally not associated with a trial and could not be counted for this analysis. The results do show a trend (six out of seven compounds, Compound C) toward diminution of new semantic content from sequentially abstracted trials within the same compound. When we examined the compound that did not conform to this trend, we found that the trials were similar in design to those for the other compounds, i.e., all short-term studies in hospitalized patients. The last trial to be abstracted included a broader set of patients (i.e., ones with either of two mutually exclusive diagnoses) but only two of the added data elements were attributable to the broader patient population.

The difference in semantic content on the data collection forms for Compound C (trials 6-9) is attributed to the four abstracted trials having been conducted by two different sponsor organizations.

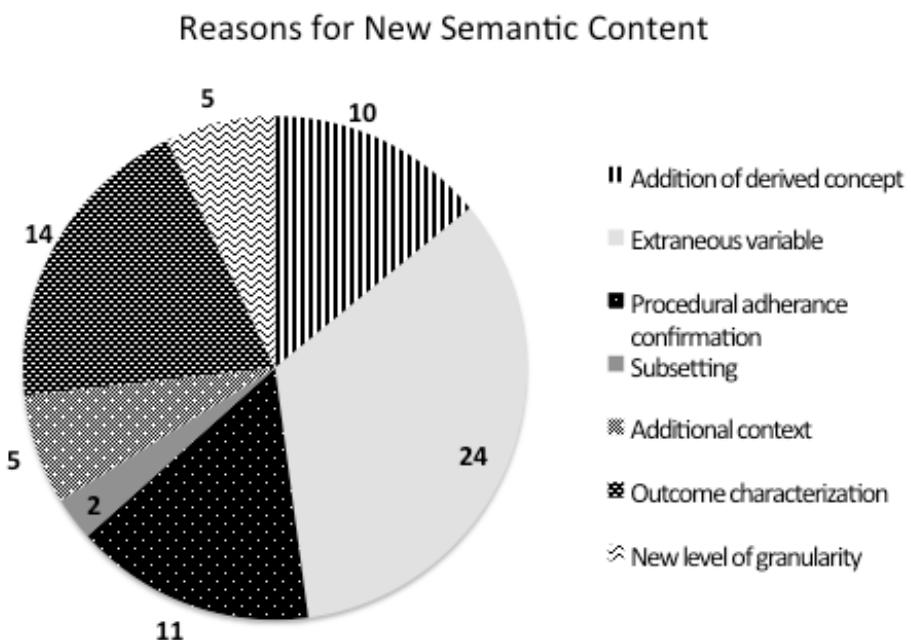


**Figure 1. New semantic content contributed from sequential abstraction of 20 trials**

To characterize the variability in semantic content, each of the 64 data elements contributed from Trials 2 through 20 were reviewed and labeled with a statement of the purpose for the new semantic content. Seven categories were synthesized from the statements. Ten instances of “addition of derived concept”, i.e., rather than collecting the raw data from which the was derived, were noted when a data element whose value could be deterministically calculated from existing data elements from previously abstracted trials was added; for example, the addition of a data element for age at diagnosis when diagnosis date and patient’s date of birth were already collected. The 24 instances of “addition of extraneous variable” comprised cases where a semantically new data element was added to further characterize the disorder (17 elements, including severity, frequency, or acuity, etc.); to characterize the treatment (five data elements); or predisposing risk factors such as family history of disease (two data elements) and the data element would plausibly be used to explain treatment effects or lack thereof.

“Procedural adherence”, eleven instances, although also a possible extraneous variable, included data elements collected to confirm completion of study procedures (for example, a check box to indicate that the schizophrenia diagnosis was made based on the research protocol-required instrument). The two “subsetting” data elements were those used to separate the data for analysis; e.g., a data element to label the DSM-IV diagnoses as Axis I vs. Axis II, or a data element to label the relative in which a predisposing family history was reported. The five instances of “additional context” included cases where the new data element provided additional information, further qualified, or added specificity for an existing data element (for example, the data element “diagnosis date” in the case where date of diagnosis was not previously collected).

The category “outcome characterization” (14 instances) included new data elements that collected information about a new outcome, or the previously unspecified extent, frequency, or severity of an outcome, such as rehospitalization, reason for readmission, or continuation of study-required hospitalization. Subgroups of data elements within the “outcome characterization” category included three safety data elements and eleven efficacy outcome data elements. Finally, the five instances of “new level of granularity” was used for new data elements pertaining to the same concept as an existing data element, but at a higher or lower level of abstraction: for example, “number of prior hospitalizations for schizophrenia” as an existing data element and a new data element collecting “number of prior psychiatric hospitalizations.” In this case, psychiatric hospitalizations may include an admission for major depressive disorder, whereas number of prior hospitalizations for schizophrenia is more specific.



**Figure 2. Categorization of reasons for new semantic content**

The number of reasons for semantic additions contributed from Trials 2-20 is shown in Figure 2 (categories are not mutually exclusive). There were 71 total reasons applied, with seven data elements having two categories applied. Importantly, the analysis concentrated on semantics only. Although representational differences such as collection of the same concept with two different scales or expressions of valid values often imply important differences in information content, [20] representational differences were explicitly excluded from the analysis.

#### **Volume and time metrics**

In addition to new semantic content, *volume*, *elapsed time*, and *time on task* were collected to further evaluate the form-based knowledge acquisition method. The data form review and abstraction was completed in 103 hours comprising three on-site sessions. The elapsed time for the abstraction was 2 months, with an additional 2 months of preparation during which the forms to be abstracted were identified from publicly available medical review reports, versions of the questionnaires referenced in the reports were obtained and abstracted, and the required background and security checks occurred. The total time on-task averaged 0.56 minutes per data element reviewed, and 1.48 minutes per data element

abstracted. Both metrics are given because the latter will vary with the percentage of reviewed data elements abstracted. In this case, after the first five trials, only new schizophrenia-specific data elements were abstracted for the second and subsequent forms from a new compound. Similarly, the number of minutes per schizophrenia-specific data element abstracted was calculated to range from 29-72 minutes. The range is provided because there were 118 data elements in addition to the core 86 that while they were schizophrenia relevant based on content, the data elements were either too detailed or defined elsewhere in more general standards, e.g., data elements regarding administration of antipsychotics are standardized in the concomitant medications domain of the CDISC SDTM. These 118 “possible candidates” were described and their inclusion or exclusion was determined by the Clinical Expert Review Committee.

The CERC review period for the set of 86 draft data elements lasted 1 month and generated 395 total comments from eleven of the twenty members. Of the 395 total CERC comments, 157 were persuasive (resulted in the addition, deletion or update of a data element), 94 not persuasive (did not prompt addition, deletion or change in a data element), and 144 were supportive (stated agreement with the data element, its definition or representation). The comments resulted in the deletion of thirty seven data elements and exclusion of 117 of the 118 aforementioned “possible candidates”.

Seventeen data elements were added based on the CERC comments. Five of the seventeen added data elements were the result of resolved conflations, i.e., suggested to the CERC instead of the conflation and thus not new content. Three added data elements were alternate operationalizations of existing concepts. One added data element was a qualifier. One added data element were promoted from the “possible candidate” list, and four others from concepts suggested to the CERC as possible missing data elements. One added data elements was an adherence variable created to encompass three different existing operationalizations. And two data elements were added as semantically new concepts, i.e., not identified from the data collection forms.

Thirteen data elements were modified based on the CERC comments (fourteen total modifications because one data element had more than one type of modification): eight modifications were to indicate precision, specifically the data element definition was modified to encompass both self reported information as well as information obtained from clinical records. We anticipate adding contextual data qualifiers to these data elements to allow indication of the information source (patient or proxy, clinical records or Both). One modification changed a data element name to more accurately represent the intended scope, two modifications were semantic changes, two modifications consisted of definitional clarification without semantic change (the reporting period was constrained from lifetime to past 24 months), and one modification was to a valid value set. Having only two added data elements containing new semantic content provided a clinical expert adjudicated domain coverage of better than 95%. Further, semantically, the subsequent HL7 ballot added two data elements to the set; these two were data elements identified in the knowledge acquisition phase and deleted by the CERC as less important than the others. The HL7 ballot comments also refined several definitions for clarity, but added no additional semantic content. This in addition to the consistency in the CERC comments suggests that form-based knowledge acquisition method provides reasonable completeness with respect to available expert knowledge.

## **DISCUSSION**

Data definition (semantic standardization) is not just important for organizational data governance, but is critical to our ability to use data for secondary analyses in healthcare. Many of the use cases for secondary data use in healthcare involve use of data by organizations other than those in which the data originated – thus semantic standardization must be broader than any one organization. Definitional

(semantic) variability has been cited as a barrier to secondary use, but this variability has not previously been rigorously analyzed in clinical research nor has its impact on data standardization. Ongoing data standardization efforts, using forms submitted in support of regulatory authorization for pharmacotherapeutics with an indication for schizophrenia, offered the opportunity for a much-needed characterization of semantic variability. Further, with national emphasis in the United States on meaningful secondary use of healthcare data, such qualitative and quantitative characterization of semantic variability and knowledge acquisition techniques that can be used to identify and synthesize data elements for standardization is urgently needed.

The initial hypothesis was that the semantic variability—as operationalized by new data elements identified from subsequent sequential abstraction from clinical trial data collection forms would be low because phase III clinical trials are a rather ideal case compared to healthcare settings. Such low semantic variability would imply that candidate data elements for standardization could be identified from a few randomly selected data collection forms. Our results, however, suggest that this will likely miss semantic content, leaving content to be identified by clinical expert review or during the ballot process. Appreciable numbers of new data elements were contributed from each compound and from each trial (form) within a compound, indicating that future form-based knowledge acquisition efforts will need to scale to cover large text-based knowledge sources. Although some steps in the identification and synthesis of data element candidates have been automated using natural-language processing techniques, these methods have yet to be evaluated for identification of candidate data elements. Such text extraction methods assume consumable text, whereas the majority of source forms used in our analysis were scanned PDF documents that required manual transcription. Further, this semantic variability comes in addition to representational variability. Although much of the latter can often be overcome by devising and applying data transformation routines, some representational variability results in the reduction of information content that similar to semantic variability often renders the information “not comparable” for many secondary data uses. Thus, semantic and representational variability present a significant information quality problem and an obstacle to secondary data use.

The majority of the suggested deletions were procedural adherence indicators, multiple operationalizations of the same concept, and data elements collecting conflated concepts. The latter were replaced with semantically resolved data elements. Multiple operationalizations included both data elements that were redundant with concepts covered in validated questionnaires and different approaches to operationalizing important concepts, e.g., diagnosis date versus date of first definitive symptoms to mark the onset of schizophrenia. With standardization of a data element set, the proportion of these variations should decrease, as will instances of varying levels of granularity that were resolved in the synthesis step. Based on the initial 204 schizophrenia specific data elements abstracted resulting in a post-review set of 67 data elements, a significant amount of the semantic variability was unnecessary, possibly attributable to lack of data element standardization. Further, the extent of semantic variability in the data collected for marketing authorization of schizophrenia compounds supports semantic variability as a significant information quality problem in secondary use of clinical research data, and further indicates that semantic variability should be analyzed as an early step in the standardization of clinical data elements.

The volume and time metrics reported are quite promising. The form-based method of knowledge acquisition, in the single clinical area examined here, appears to perform well with respect to domain coverage as adjudicated by clinical experts, and better with respect to elapsed time and productivity than previously reported methods based on knowledge acquisition from experts. [16] Additionally, the method reported here reduces as much as possible the amount of time and effort required by clinical experts.

## **LIMITATIONS AND SUGGESTIONS FOR FURTHER WORK**

This report provides a detailed characterization of semantic variability in schizophrenia data submitted

for regulatory review, caution should be used when applying the findings to other arenas. This work was observational by nature, and causal statements should not be made based on these results. Further, the types of data collected in this therapeutic setting—for instance, the large proportion of information derived from questionnaires or the extent of extant authoritative clinical definition of concepts —may not be typical of other clinical specialties. In addition, current standards of care play a role with regard to generalizability. Therapeutic areas differ with respect to the volatility of data generated and used in standard care; thus, therapeutic areas with differential rates of knowledge generation and translation into care may also have different semantic variability profiles. The work we present here, however, remains an important contribution, because it is the first to analyze semantic variability in a clinical specialty area while also examining a methodology for assessing the effects of semantic variability on data element definition and standardization. The use of one abstractor, described in the methodology section, and the use of that same abstractor to synthesize the abstracted data elements is a weakness of the work and represents potential bias. Due to the expense of the manual abstraction and the travel required to abstract the forms on site at the FDA, this weakness could not be avoided. Two steps were taken to mitigate the potential impact: 1) the abstraction was completely systematized, and the synthesis was systematized to the extent possible. To further mitigate the impact, the synthesized data elements are publicly available for reanalysis. Further research should be directed toward characterizing semantic variability in other therapeutic areas to assess generalizability of these findings. Other aspects that deserve consideration, given the accelerating growth of data-driven knowledge acquisition in clinical setting, are 1) automated identification and extraction of data elements from text-based knowledge sources, 2) the efforts required to maintain standard data element sets, and 3) the effectiveness of processes in handling new semantic content during the interval between concept identification and standardization. Further work is also needed to establish methods for integration of organizational data governance processes with national and international public metadata (data element) registries, and in health care, to streamline translation of new knowledge into clinical care guidelines and from care guidelines to performance measures and the standard data elements to support performance measurement.

## **CONCLUSIONS**

In the therapeutic area studied, significant semantic variability existed. Based on the presented results, we conclude that for schizophrenia, the number of forms abstracted could not have been reduced without sacrificing completeness of the data element set and relying more heavily on clinical experts for concept identification. Further, based on the small amount of semantic content added by the CERC, two data elements, it seems that the form-based knowledge acquisition method performs well, better than 95% domain completeness as adjudicated by clinical experts. Based on the initial 204 schizophrenia specific data elements resulting in a post-review set of 67 data elements, a significant amount of the semantic variability exhibited in the schizophrenia phase III pivotal clinical trials was unnecessary. Further, the semantic variability supports the implication of semantic variability as a significant information quality problem in secondary use of clinical research data. The form-based knowledge acquisition method performed well with respect to productivity and elapsed time, and may outperform knowledge acquisition from experts. At least within mental health, form based knowledge acquisition appears to provide a feasible production scale for data element standardization.

## **ACKNOWLEDGEMENTS**

This work was made possible by grant number 1R24FD004271-01 from the United States Food and Drug Administration (FDA), a component of the Department of Health and Human Services (HHS). Its contents are solely the responsibility of the author and do not necessarily represent the official view of the FDA.

## REFERENCES

- [1] Kalet IJ. Principles of Biomedical Informatics. London: Academic Press; 2009.
- [2] Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The Canon Group's effort: working toward a merged model. *J Am Med Inform Assoc.* Jan-Feb 1995;2(1):4-18.
- [3] Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS. Toward a medical-concept representation language. *The Canon Group. J Am Med Inform Assoc.* May-Jun 1994, 1(3):207-217.
- [4] Kush R, Alschuler L, Ruggeri R, et al. Implementing Single Source: the STARBRITE proof-of-concept study. *J Am Med Inform Assoc.* Sep-Oct 2007, 14(5):662-673.
- [5] Mead CN. Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? *J Health Inf Manag.* Winter 2006, 20(1):71-78.
- [6] Biondich PG, Downs SM, Carroll AE, Shiffman RN, McDonald CJ. Collaboration between the medical informatics community and guideline authors: fostering HIT standard development that matters. *AMIA Annu Symp Proc.* 2006:36-40.
- [7] Pawlson LG, Scholle SH, Powers A. Comparison of administrative-only versus administrative plus chart review data for reporting HEDIS hybrid measures. *Am J Manag Care.* Oct 2007, 13(10):553-558.
- [8] Meads S, Cooney JP. The medical record as a data source: use and abuse. *Top Health Rec Manage.* Jun 1982, 2(4):23-32.
- [9] Ewen EF, Zhao L, Kolm P, et al. Determining the in-hospital cost of bleeding in patients undergoing percutaneous coronary intervention. *J Interv Cardiol.* Jun 2009, 22(3):266-273.
- [10] Dick RS, Steen EB, Detmer DE, Eds. The computer-based patient record: an essential technology for healthcare. Second ed. Washington DC: National Academy Press; 1997.
- [11] Kush R. The Cost of Clinical Data Interchange in Clinical Trials: A CDISC White Paper. Austin, Tx: Clinical Data Standards Interchange Consortium (CDISC); August 2001.
- [12] Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc.* Nov-Dec 2007;14(6):687-696.
- [13] Tang PC. AMIA advocates national health information system in fight against national health threats. *J Am Med Inform Assoc.* Mar-Apr 2002;9(2):123-124.
- [14] Tcheng J, Nahm M, Fendt K. Data quality issues and the electronic health record. Drug Information Association (DIA) Global Forum. December 2010 2010;2(6):36-40.
- [15] Nahm M., Kush R., Richesson R., Standardizing Clinical Data Elements to Support Regulatory Review of Marketed Therapeutics. AMIA Clinical Research Informatics Summit, San Francisco, Ca. March 2012.
- [16] Nahm, M., Walden, A., McCourt, B., Pieper, K., Honeycutt, E., Hamilton, C.D., Harrington, R.A., Diefenbach, J., Kisler, B., Walker, M., Hammond, W.E., Standardizing Clinical Data Elements. *International Journal of Functional Informatics and Personalised Medicine (IJFIPM)* Special Issue on: "The Informatics of Meta-data, Questions, and Value Sets". Vol. 3, No. 4, 2010.
- [17] United States Food and Drug Administration New Drug Approvals Database. Accessed June 30, 2012, available from <http://www.fda.gov>.
- [18] ISO/IEC JTC1 SC32 WG2 Development/Maintenance standard. ISO/IEC 11179, Information Technology -- Metadata registries (MDR). <http://metadata-standards.org/11179/>. Accessed June 22, 2012.
- [19] Chisholm, M.D. Definitions in Information Management: A Guide to the Fundamental Semantic Metadata. San Francisco, CA: DesignMedia; 2010.
- [20] Nahm M, Zhang J, Operationalization of the UFuRT methodology for usability analysis in the clinical research data management domain. *Journal of Biomedical Informatics* 2009 42(2):327-33.

# **TOWARDS EXPERTISE MODELLING FOR ROUTING DATA CLEANING TASKS WITHIN A COMMUNITY OF KNOWLEDGE WORKERS**

(Research-in-Progress)  
(Data Scrubbing and Cleaning, Crowd Sourcing, Community Input)

**Umair ul Hassan**  
Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
umair.ul.hassan@deri.org

**Sean O'Riain**  
Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
sean.oriain@deri.org

**Edward Curry**  
Digital Enterprise Research Institute  
National University of Ireland  
Galway, Ireland  
ed.curry@deri.org

**ABSTRACT:** Applications consuming data have to deal with variety of data quality issues such as missing values, duplication, incorrect values, etc. Although automatic approaches can be utilized for data cleaning the results can remain uncertain. Therefore updates suggested by automatic data cleaning algorithms require further human verification. This paper presents an approach for generating tasks for uncertain updates and routing these tasks to appropriate workers based on their expertise. Specifically the paper tackles the problem of modelling the expertise of knowledge workers for the purpose of routing tasks within collaborative data quality management. The proposed expertise model represents the profile of a worker against a set of concepts describing the data. A simple routing algorithm is employed for leveraging the expertise profiles for matching data cleaning tasks with workers. The proposed approach is evaluated on a real world dataset using human workers. The results demonstrate the effectiveness of using concepts described the data for modelling expertise, in terms of likelihood of receiving responses to tasks routed to workers.

**Keywords:** data cleaning, crowd sourcing, web 2.0, linked data

## INTRODUCTION

The information systems of a business contain data on entities important to the business such as products, customers, suppliers, employees, etc. Entity information is spread across the organization, shared with partners, or even outside its boundaries of control, for example on the web. Maintaining a clean and consistent view of business critical entities is a core requirement of any knowledge based organization, as highlighted by a recent survey on the value of data analytics in organizations [1]. The study found that more than 30% executives considered *integration*, *consistency*, and *trustworthiness* their top most data priorities. Most of the information quality research has focused on the development of sophisticated data quality tools and approaches such as Master Data Management. However these tools and techniques necessitate high technical expertise for successful implementation. Consequently, one of the major obstacles to data quality are the high operational costs due to limited availability of a few experts, and changes to business rules and policies [2], [3]. To overcome his limitation automatic or semi-automatic data cleaning algorithms can be used to improve data quality. However, the output of these algorithms can still require human review to ensure trust for decision making.

Involving the community of users in data management activities has shown promising results for maintaining high quality data [4]. Recent developments in *crowdsourcing* [5] and *human computation* [6] have fuelled the interest in algorithmic access to human workers, within or outside organizations, for performing computationally difficult tasks. Most of the current approaches of human computation publish tasks on task markets such as Amazon Mechanical Turk<sup>1</sup>. Therefore leaving the choice of task selection to the unknown workers, through search and/or browse capabilities of the platform. As a result the quality of responses provided by the workers may suffer from lack of domain knowledge or expertise for the task at hand. However, if the knowledge of workers' expertise is understood, tasks can be assigned to appropriate workers in a crowd or community. This process is known as *task routing*.

In this paper we propose a approach for task routing that profiles knowledge workers according to their expertise of concepts related to data quality issues and then assigns data quality tasks to appropriate workers. The approach is implemented in the *CAMEE* (Collaborative Management of Enterprise Entities) system. Given a set of data cleaning updates, CAMEE automatically converts them to feedback tasks for further verification from the group of knowledge workers considering their individual expertise levels. We argue that the expertise level of workers can be effectively measured against concepts associated with data quality tasks, where concepts are extracted from source data.

In this paper, we address the problem of building expertise profiles of worker and leveraging these profiles for routing tasks to appropriate workers. The contributions of this paper are as follows:

- An approach for modelling and assessment of knowledge worker's expertise with concepts and a prototype implementation of the approach using SKOS<sup>2</sup> concepts
- A simple concept matching approach for routing data quality tasks to appropriate worker
- A preliminary evaluation of proposed system on real world dataset with real world workers to demonstrate its effectiveness

The rest of this paper is organized as follows. Next section motivates the research work with respect to data quality management. Then we provide an overview of the system architecture and related research challenges. The implementation section details the prototype system using SKOS concepts for modelling expertise, as well as two approaches of building expertise model for task routing. The section on evaluation presents the experimental details and discusses the results. Finally we provide the review of existing work in closely related research areas and summarize the paper afterwards.

---

<sup>1</sup> <http://www.mturk.com>

<sup>2</sup> <http://www.w3.org/2004/02/skos/>

## MOTIVATION

*Master Data Management* (MDM) [7] has become a popular approach for managing quality of enterprise data. The main benefit of a successful MDM implementation is readily available high quality data about entities in an enterprise. Although attractive, recent studies estimate that more than 80% data integration projects in enterprises either fail or overrun their budget [2], [8]. MDM is heavily centralized and labour intensive, where the cost and effort in terms of expertise can become prohibitively high. The main responsibility for data quality management lies with the MDM council in a *top-down* manner [9]. An MDM council usually includes members from senior management, business managers and data stewards.

The significant upfront costs in terms of development efforts and organizational changes make MDM difficult to implement successfully across large enterprises. The concentration of data management and stewardship between few highly skilled individuals, like developers and data experts, also proves to be a bottleneck. To this end, the lack of delegation of data management responsibilities is considered as one of most the significant barriers to data quality [2]. Due to the limited number of skilled human resources, only a small percentage of enterprise data comes under management. As a result, the scalability of MDM becomes a major issue when new sources of information are added over time. Not only are enterprises unable to cope with the scale of data generated within their boundaries. As the web data becomes important, there will be a need for enterprises to manage external data existing outside their boundaries within shared global information ecosystems [10].

Effectively involving a wider community of users within collaborative data cleaning and information management activities is attractive proposition. The *bottom-up* approach of involving crowds in creation and management of general knowledge has been demonstrated by projects like Freebase<sup>3</sup>, Wikipedia<sup>4</sup>, and DBpedia<sup>5</sup> [4]. Similarly data quality workload can be delegated to community of end-users by effectively guiding them towards specific tasks in *top-down* manner [11]. Sourcing data quality tasks to a community or crowd necessitates explicit control over the actions required from humans and their potential outcome.

*Human computation* [6] is a relatively recent field of research that focuses on the design of algorithms with operations or functions carried out by human workers. One of the major aspects of human computation is to understand the expertise of available humans and match them with the appropriate tasks. In this respect, systems using human computation need to overcome two challenges; 1) how to assess and model human expertise towards, and 2) how to effectively route tasks to appropriate workers. In this paper we outline a collaborative data quality management system that follows a human computation approach for involving end-users in the cleaning process. We introduce a concept based approach for modelling the expertise of human workers for task routing.

## CAMEE OVERVIEW

CAMEE follows a human computation approach that utilizes community participation to incrementally increase the quality of data. Using CAMEE, technical experts (e.g. developers, data stewards, and data analyst) define the data quality processes with the objective of routing tasks to human workers having relevant domain knowledge to complete the task. The worker may be employees of the organization or sourced from an online marketplace. The rest of this section describes the workflow of the system followed by discussion on challenges of expertise modelling and task routing.

### System Workflow

Figure 1 presents the high level workflow of the CAMEE system. The input to CAMEE is a dirty dataset that is assessed by *data cleaning algorithms* against pre-defined policies or rules, to identify data quality issues.

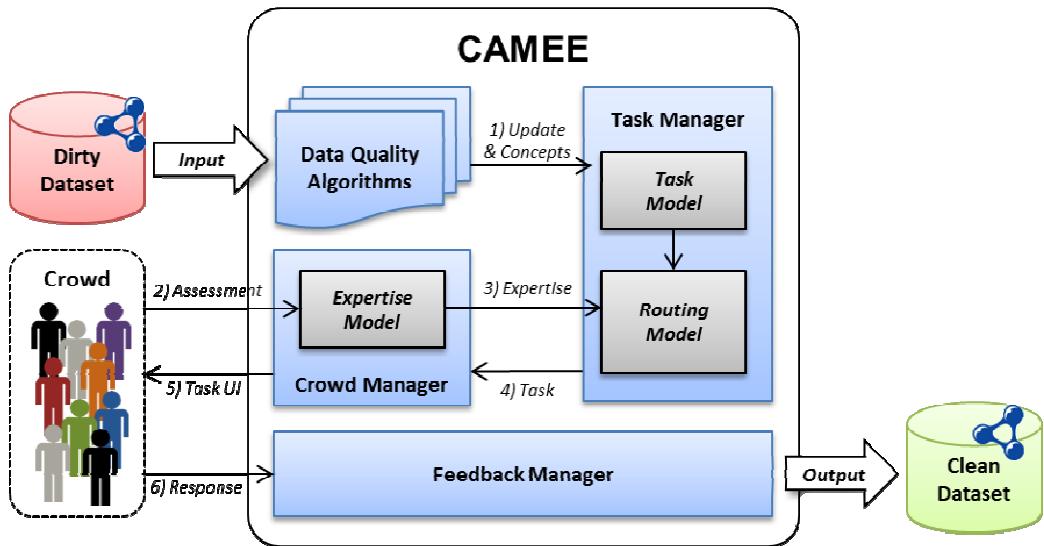
- 1) Data quality algorithms suggest *updates* to the dataset for each data quality issue. The *concepts* describing the dataset are extracted and associated with each update. The suggested updates are fed to the *task manager* component, which converts an update into a task.

---

<sup>3</sup> <http://www.freebase.com>

<sup>4</sup> <http://www.wikipedia.org>

<sup>5</sup> <http://www.dbpedia.org>



**Figure 1: An example workflow of CAMEE for cleaning dataset with crowdsourcing.**

- 2) The *crowd manager* component maintains an *expertise model* by either soliciting expertise level directly from workers, or by calculating indirectly through their performance for test tasks with known responses.
- 3) The *routing model* matches each task with the appropriate worker according to their expertise, and then;
- 4) Submits the task to the crowd manager for execution.
- 5) The crowd manager renders each task using an appropriate user interface.
- 6) The *feedback manager* captures the response to the tasks and generates a cleaned dataset as output of the system.

### **Expertise & Routing**

Human computation approaches rely on explicit control over routing of tasks to appropriate human workers. The tasks can be routed following a pull method by posting tasks on an online marketplace, such as Amazon Mechanical Turk. In pull method the decision of routing is delegated onto the humans themselves by allowing them to select tasks using search or browse features of the marketplace. On the other hand the push method of routing actively selects appropriate workers from a pool of available human resources. CAMEE follows push method of task routing that requires an understanding of the expertise of human workers for matching tasks to appropriate workers. The main challenges associated with push routing are

- How to represent domain knowledge of data quality task
- How to assess and represent expertise of workers for a particular domain of knowledge
- How to match domain of data quality task with expertise of workers

The expertise required to complete a data quality tasks not only depends on the type of task but also on the domain knowledge. In this paper we propose a concept based approach for addressing above mentioned challenges. We show that concepts extracted from the source data can be effectively used for modelling worker expertise and routing tasks. In next section we describe an example implementation of the approach within CAMEE that exploits concepts in source data as the common denominator for annotating data quality tasks, building worker expertise, and routing tasks.

## CONCEPT-BASED EXPERTISE MODELLING WITHIN CAMEE

In this section we provide details of the prototype implementation of concept-based expertise modelling within CAMEE. We illustrate by example the application of concepts based expertise modelling and task routing within data quality management.

### SKOS Concepts

The *Simple Knowledge Organization System* (SKOS) is a W3C recommended data model designed to represent knowledge organization systems and share them through the Web [12]. The organization systems can include thesauri, subject headings, classification schemes, taxonomies, glossaries and other structured controlled vocabularies. In SKOS the basic element is a concept, identified by URI<sup>6</sup>, which is considered to be ‘unit of thought’; ideas, meanings or objects. Furthermore, SKOS defines attributes for labelling concepts with lexical strings and providing additional textual information regarding the concept. Concepts can be grouped into concept schemes and linked with other concepts by using semantic relationship hierarchical or associative attributes in SKOS. The overall objective of SKOS is to provide a common data model for knowledge organization systems, to facilitate their interoperability, as well as to make them machine-readable through a web-based data format called *Resource Description Framework*<sup>7</sup> (RDF). The usability of SKOS has been demonstrated with use cases of knowledge organization systems from life sciences, agriculture, product lifecycle, and media [13]. In this paper, we use the case of DBpedia [14] which is a structured knowledge base constructed by extracting and linking entities from Wikipedia. Figure 2 shows properties and values of concept [American biographical films](#) in DBpedia.

About: American biographical films	
An Entity of Type : <a href="#">Concept</a> , from Named Graph : <a href="http://dbpedia.org">http://dbpedia.org</a> , within Data Space : <a href="http://dbpedia.org">dbpedia.org</a>	
	
Property	Value
<a href="#">rdf:type</a>	skos:Concept
<a href="#">rdfs:label</a>	American biographical films
<a href="#">owl:sameAs</a>	<ul style="list-style-type: none"> <li>▪ <a href="http://ru.dbpedia.org/resource/Категория:Фильмы-биографии_США">http://ru.dbpedia.org/resource/Категория:Фильмы-биографии_США</a></li> <li>▪ <a href="http://cs.dbpedia.org/resource/Kategorie_Americké_životopisné_film">http://cs.dbpedia.org/resource/Kategorie_Americké_životopisné_film</a></li> <li>▪ <a href="http://ko.dbpedia.org/resource/분류:미국의_전기_영화">http://ko.dbpedia.org/resource/분류:미국의_전기_영화</a></li> </ul>
<a href="#">skos:broader</a>	<ul style="list-style-type: none"> <li>▪ category:<a href="#">Biographical_films_by_country</a></li> <li>▪ category:<a href="#">American_films_by_genre</a></li> </ul>
<a href="#">skos:prefLabel</a>	American biographical films
<a 0;"="" href="http://www.w3.org/ns/prov#wasDerivedFrom&lt;/a&gt;&lt;/td&gt;&lt;td&gt; &lt;ul style=" list-style-type:="" none;="" padding-left:=""> <li>▪ <a href="http://en.wikipedia.org/wiki/Category:American_biographical_films?oldid=490787017">http://en.wikipedia.org/wiki/Category:American_biographical_films?oldid=490787017</a></li> </a>	
<a href="#">is dcterms:subject of</a>	<ul style="list-style-type: none"> <li>▪ <a href="#">dbpedia:The_Diving_Bell_and_the_Butterfly_(film)</a></li> <li>▪ <a href="#">dbpedia:Summer_of_‘42</a></li> <li>▪ <a href="#">dbpedia:Patton_(film)</a></li> </ul>

Figure 2: Screenshot of RDF data in DBpedia about the SKOS concept [American biographical films](#)

DBpedia converts Wikipedia articles to entities in RDF format through hand crafted mappings and natural language techniques. Similarly it converts concepts from Wikipedia category system to SKOS concepts. Figure 3 shows some attributes and concepts of the Wikipedia article for the movie “A Beautiful Mind” in RDF format.

<sup>6</sup> Uniform Resource Identifier

<sup>7</sup> <http://www.w3.org/RDF/>

**About: A Beautiful Mind (film)**

An Entity of Type : [work](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](http://dbpedia.org)

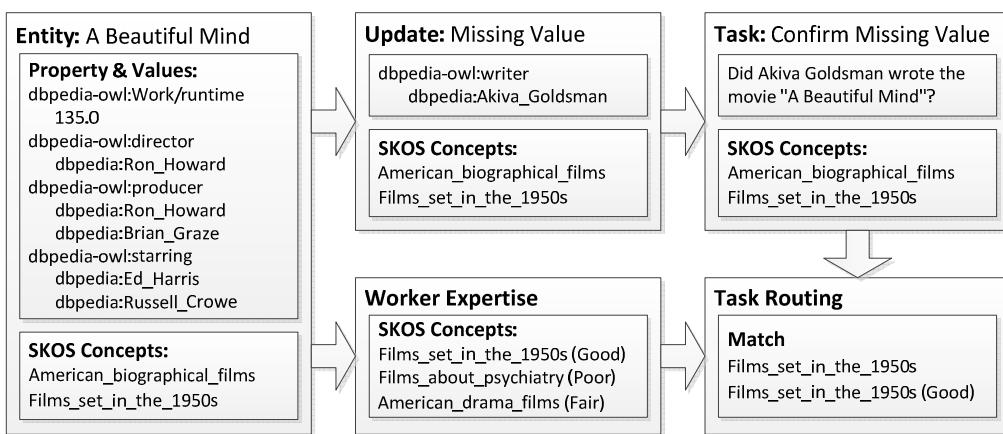


A Beautiful Mind is a 2001 American biographical drama film based on the life of John Nash, a Nobel Laureate in Economics. The film was directed by Ron Howard and written by Akiva Goldsman. It was inspired by a bestselling, Pulitzer Prize-nominated 1998 book of the same name by Sylvia Nasar. The film stars Russell Crowe, along with Ed Harris, Jennifer Connelly, Paul Bettany and Christopher Plummer. The story begins in the early years of a young prodigy named John Nash.

Property	Value
dbpedia-owl:Work/runtime	▪ 135.0
dbpedia-owl:budget	▪ 6.0E7
dbpedia-owl:cinematography	▪ dbpedia:Roger_Deakins
dbpedia-owl:director	▪ dbpedia:Ron_Howard
dbpedia-owl:producer	▪ dbpedia:Ron_Howard ▪ dbpedia:Brian_Grazer
dbpedia-owl:runtime	▪ 8100.000000 (xsd:double)
dbpedia-owl:starring	▪ dbpedia:Ed_Harris ▪ dbpedia:Russell_Crowe
dcterms:subject	▪ category:Best_Drama_Picture_Golden_Globe_winners ▪ category:American_biographical_films ▪ category:Films_set_in_the_1950s
rdf:type	▪ owl:Thing ▪ dbpedia-owl:Work ▪ dbpedia-owl:Film

**Figure 3: Screenshot of RDF data in DBpedia about the movie “A Beautiful Mind”**

In Figure 3 the *dbpedia-owl:starring* attribute have been extracted from the InfoBox of the Wikipedia article. The *dct:subject* attributes has been assigned the SKOS concept extracted from article’s categories box. For example, [http://dbpedia.org/resource/Category:American\\_biographical\\_films](http://dbpedia.org/resource/Category:American_biographical_films) represents the SKOS concept equivalent of Wikipedia category “American Biographical Films”. While the Wikipedia category system is collaboratively created and updated by editors, similar or even more sophisticated knowledge organization systems exists within large enterprises. There are tools<sup>8</sup> available for generation and management of SKOS concept schemes from existing taxonomies, vocabularies or knowledge organization systems. Figure 4 give an example use of SKOS concepts by CAMEE for representing domain of knowledge for data quality tasks, expertise of knowledge worker and task routing decisions.



**Figure 4: Example use of SKOS concepts for representing expertise and task routing in CAMEE**

<sup>8</sup> <http://www.w3.org/2001/sw/wiki/SKOS>

## Expertise Modelling

SKOS provides a language to design knowledge structures in as simple as possible way. We use SKOS concepts, from source data, for modelling expertise requirements of tasks and knowledge level of workers for CAMEE. Assuming that the entities in the dataset have been annotated with some simple SKOS concept scheme as highlighted in Figure 2, the task manager associates concepts with the data quality task. For example the data quality task for the movie entity [A Beautiful Mind \(film\)](#) has [American biographical films](#), [Best Drama Picture Golden Globe winners](#), and [Films\\_set\\_in\\_1950s](#) SKOS concepts associated with it. The crowd manager component builds worker profiles for the SKOS concepts according one of the following two approaches:

- *Self-Assessment (SA):* In this approach a worker is asked to rate their knowledge level among the list of all concepts in the dataset.
- *Test Assessment (TA):* A worker's knowledge expertise is based on her performance of data quality tasks with known answers, where each tasks has concepts associated with it.

For example, a worker can specify their knowledge level for [American biographical films](#) concepts as excellent for SA approach. However during the TA approach her responses for the test tasks associated with [American biographical films](#) can suggest a below average level of knowledge. Table 3 gives an example of expertise profiles for 3 workers on 4 concepts related to movies, where each value represents the knowledge level between the values of 0 and 1.

Concept	Worker 1	Worker 2	Worker 3
1990s_comedy-drama_films	0.6	0.2	0.2
Films_about_psychiatry	0.6	0.2	0.6
American_biographical_films	0.8	0.4	0.4
American_comedy-drama_films	0.8	0.6	0.6

**Table 3: Example of matrix of expert profiles for 3 workers and 4 movie concepts**

## Task Routing

The expertise model is exploited by the task routing model for matching tasks with appropriate knowledge workers. In this paper following matching strategies are employed for the purpose of routing

- *Random:* Sends a particular task to any randomly selected worker from the pool of all available workers. This routing strategy assumes unavailability of a worker's expertise model, thus serving as the baseline approach as well as fall back strategy.
- *Expertise Match:* This strategy ranks workers according to the weighted matching score between task concepts and the worker's expertise profile. The weights are based on the expertise model built earlier. The example task discussed would be routed to the worker with highest score for the [American biographical films](#), [Films\\_about\\_psychiatry](#), and [Films\\_based\\_on.biographies](#) concepts

## EVALUATION

We performed an empirical evaluation of task routing based on the proposed expertise model using the two approaches; self-assessment and task-assessment. The two objectives of the experiments are 1) to compare random routing without using workers' expertise models versus routing based on matching task concepts and worker exper-

tise, and 2) to investigate the best approach for building the worker expertise model. We evaluated if the concepts extracted from the dataset can be utilized effectively for representing the knowledge space of data quality tasks and worker expertise. In this regards we have explored the following proposition through empirical evaluation:

Data quality tasks routed using a concept-based expertise profiles have higher response rates if the expertise model is built using a task-assessment approach as compared to a self-assessment based approach.

## Experiments

In this section we provide the details of the experiment design employed for the purpose of evaluation. We have divided the experimental evaluation in two stage process.

- The *assessment stage* focused on building the expertise model of workers. During this stage workers were asked to complete one assessment for each of the expertise building approaches. A simple 5 points belief scale (i.e. *none*, *poor*, *fair*, *good*, and *excellent*) was used for the self-assessment of knowledge about concepts. The workers were asked to provide responses to task-assessments based on Likert scale<sup>9</sup> (i.e. *don't know*, *strongly disagree*, *disagree*, *neutral*, *agree*, *strongly agree*). *None* and *Don't Know* were the default selected options for belief scale and Likert scale, respectively.
- The *routing stage* used the generated expertise for routing data quality tasks to appropriate knowledge workers. These responses were used to calculate quality for final output dataset.

The response of workers for tasks routed to them is recorded against Likert scale with default response of “Don’t Know”. So for a particular approach a high percentage of workers providing “Don’t Know” responses indicate a low likeliness of getting data cleaned with help of workers. While a low percentage of “Don’t Know” responses indicated a high likeliness. In the rest of this section, we describe the datasets used for experiments, as well as the data quality tasks required to clean these datasets. Details of the population of knowledge workers and their characteristics are also discussed.

## Dataset Description

We have used a subset of DBpedia describing movies within the experimentation. A test dataset was created by selecting Academy Award and FilmFare Award winning movies, as well as the top 100 grossing movies from Hollywood and Bollywood. The DBpedia database provides variety of concept schemes for entities. However for the purpose of this experiment we selected 42 film genre concepts associated with movies. Detailed statistics of the dataset are listed in the Table 4.

Characteristic	Value
Number of entities (dbp:Film)	724
No. of concepts	42
No. of data quality tasks	230

**Table 4: Characteristics of dataset describing award winning and top 100 grossing movies from Hollywood and Bollywood in DBpedia**

## Data Quality Tasks

The original movie dataset a variety of data quality issues. Table 5 highlights three particular types of issues. Each of these data quality issues is converted to a human computation task, which can be routed to knowledge workers. The conversion process involved creating a short question for the DQ issue, by using available data for the entity.

<sup>9</sup> [http://en.wikipedia.org/wiki/Likert\\_scale](http://en.wikipedia.org/wiki/Likert_scale)

DQ Issue Type	Example question for DQ task
Identity Resolution	Does the following URIs represent the same entity? (Answer YES or NO) <a href="http://dbpedia.org/resource/Shanghai_(2010_film)">http://dbpedia.org/resource/Shanghai_(2010_film)</a> <a href="http://rdf.freebase.com/ns/m/047fjjr">http://rdf.freebase.com/ns/m/047fjjr</a>
Missing Value	Did the following actor starred in the movie “Titanic”? (Answer YES or NO) <a href="http://www.dbpedia.org/resource/bruce_willis">http://www.dbpedia.org/resource/bruce_willis</a>
Data Repair	Was the following movie released in 21-10-2011 or 21-10-2010? (Answer YES or NO) <a href="http://www.dbpedia.org/resource/the_iron_lady">http://www.dbpedia.org/resource/the_iron_lady</a>

**Table 5: Examples of questions for the human computation tasks associated with specific data quality issues**

The dataset was cleaned manually by an expert to serve as the gold standard. The data quality tasks were created by collecting correct and incorrect values for the “starring” attribute for movies. Figure 5 shows a screenshot of a human computation task.

The screenshot shows a web-based application interface for a human computation task. At the top, there is a navigation bar with links for Home, Assessments, Tasks, and Datasets. On the right side of the header, it says "Welcome [User Name] - you are logged in as Admin User (Logout)". Below the header, the main content area has a title "Film: Crash (2004 film)" and a subtitle "Due by: 10 Jul 12, 00:01". It also lists "Related concepts: Gang films". A section titled "Task Description" contains the text "MICHAEL PENA starred in film CRASH (2004 FILM)" in red. Below this, a question asks "Please indicate if above is true as per your knowledge:" followed by a row of radio buttons for responses: "Don't know", "Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree". At the bottom of the task area, there are three buttons: "Save and Previous", "Save", and "Save and Next".

**Figure 5: Screenshot of the CAMEE prototype system for crowd sourcing data quality tasks**

## Knowledge Workers

We recruited volunteer workers to perform the human computation tasks for data quality. The final community of workers contained people from 3 regions of worlds (Europe, South Asia, and Middle East) having varying knowledge about the movie dataset, as shown in Table 6.

Characteristic	Value
No. of Workers	11
Tasks for Assessment Stage	100
Tasks for Routing Stage	130

**Table 6: Characteristics of knowledge worker recruited for the experiments, as well as statistics of tasks assigned to them during test stage**

## Results

The following results show the distribution of responses for the *Random* routing as compared to *Expertise Match* based routing coupled with the expertise modelling approaches. As expected both matching based routing strategies outperform random routing of tasks. The data confirms that building expertise models based on performance on task-assessments is a better approach as compared just soliciting self-assessment of knowledge about concepts.

Expertise Approach	Random	Self-Assessment + Matching	Task Assessment + Matching
Don't know	73.85%	56.15%	36.92%
Strongly Disagree	6.92%	14.62%	16.15%
Disagree	6.15%	5.38%	13.08%
Neutral	0.00%	3.85%	7.69%
Agree	3.08%	5.38%	8.46%
Strongly Agree	10.00%	14.62%	17.69%

**Table 7: Distribution of responses during routing stage, for 3 task routing approaches. A high percentage of “Don’t Know” response indicates that the tasks has been routed to worker with no domain knowledge.**

## RELATED WORK

The crowdsourcing approaches for data management activities can be categories in three approaches; *algorithmic approaches*, *crowd-sourced databases* and *application platforms*.

*Algorithmic approaches* focus on the designing algorithms for reducing uncertainty of data management with human computed functions. In these approaches human attention is utilized to support data management system in different activities, such as schema matching [15], entity resolution [16] and data repair [17]. The objective of algorithmic approaches is to help increase utility of human attention through optimization of specific data management activities. Consequently the evaluation of these approaches focus on the measurement of incremental utility improvement after successive human interventions. Our work focuses on modelling expertise required for data quality tasks and building worker profiles to facilitate task routing.

*Crowd-sourced database* systems focus on providing programmatic access to human computation platforms for database operations such as joins, sorts, and inserts. This facilitates platform independence with respect to the details of access to human services. Typically existing query languages are extended to minimize the learning curve associated with programming human computation. For example, CrowdDB [18] extends *standard query language* to provide database services on top on crowd sourcing platforms. An initial list of information quality problems which can be solved with crowdsourcing have be identified in [19]. The application of human computation has been demonstrated for data management problems such as *data ranking* [20], *relevance assessment* [21] and *entity linking* [22]. These research efforts focus on improving the quality of crowd responses through various task aggregation techniques after execution. Instead we focus the step before execution of tasks; improving the routing of tasks to workers with appropriate domain knowledge and expertise.

*Application platforms* extend existing applications with custom human computation capabilities, thus enabling

crowd services in applications. These approaches do not depend on external platforms for human services as compared to previous categories. Freebase supported by a human computation platform called RABj [23], which allows users to distribute specific tasks to communities of paid or volunteering worker. Similarly, MOBS [24] provides a tool extension approach for enabling crowd sourcing of schema matching applications. Both RABj and MOBS are crowd sourcing platforms tailored for specific data management applications. We propose CAMEE; a human computation based approach for guided data cleaning. The objective of CAMEE is to facilitate task routing for effective utilization of human attention in collaborative data cleaning processes.

*Expert finding* has been the subject of a considerable amount of research in the Information Retrieval community [25]. The expert finding problem involves ranking the list of experts according to their knowledge about a given topic or query. Generally, some web-based or enterprise text corpus is utilized to uncover associations between experts and topics [26]. On the other hand, *expert profiling* is defined as the opposite process of determining the list of topics that an expert is knowledge about [27]. In both cases, current approaches mine existing text corpus to determine worker and topics associations. By contrast, in this paper we are interested in profiling expertise of workers for finding task and worker associations. We cast this problem in a data cleaning scenario where we building profiles by only using source data. We assume that the source data does not provide any evidence of worker expertise in form of person and topic associations. Instead we demonstrate the effective use of SKOS for the purpose of expertise profiling and task routing with in data cleaning scenario.

## SUMMARY AND FUTURE WORK

This paper presents an concepts based approach for routing data quality tasks to appropriate workers based on an their knowledge and expertise. An expertise model for representing worker profiles against a set of concepts from the dataset is described. The approach is validated with a simple routing algorithm for exploiting expertise model based on either concept selection or task performance. The approach is evaluated on real world datasets using human workers. The results demonstrate the effectiveness of using concept based profiles for soliciting higher number of responses from workers. In this paper we described the architecture of CAMEE and its use of SKOS concepts for modelling expertise for tasks and knowledge worker. As the part of future work we plan to expand our analysis of the system to effect of various expertise assessment methods and task routing methods on quality of task routing. Further research is also required into the effective balancing of the community workload under constraints such as cost, latency, and motivation. We plan to investigate the utility of CAMEE in real world information management scenario that deals with multiple data sources and heterogeneity problems, such as enterprise energy management [28].

## ACKNOWLEDGEMENTS

We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Some of the references cited in this paper are included for illustrative purposes only. The work presented in this paper is funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion- 2).

## REFERENCES

- [1] S. Lavalle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big Data, Analytics and the Path from Insights to Value,” *MIT Sloan Management Review*, vol. 52, no. 2, pp. 21–32, 2011.
- [2] A. Haug and J. S. Arlbjørn, “Barriers to master data quality,” *Journal of Enterprise Information Management*, vol. 24, no. 3, pp. 288–303, 2011.
- [3] R. Silvola, O. Jaaskelainen, H. Kropsu-Vehkaperä, and H. Haapasalo, “Managing one master data – challenges and preconditions,” *Industrial Management & Data Systems*, vol. 111, no. 1, pp. 146–162, 2011.
- [4] E. Curry, A. Freitas, and S. O. Ri, “The Role of Community-Driven Data Curation for Enterprises,” in *Linking Enterprise Data*, D. Wood, Ed. Boston, MA: Springer US, 2010, pp. 25–47.
- [5] A. Doan, R. Ramakrishnan, and A. Y. Halevy, “Crowdsourcing systems on the World-Wide Web,” *Communications of the ACM*, vol. 54, no. 4, p. 86, Apr. 2011.
- [6] E. Law and L. von Ahn, “Human Computation,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121, Jun. 2011.
- [7] D. Loshin, *Master Data Management*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

- [8] B. Otto and A. Reichert, "Organizing Master Data Management: Findings from an Expert Survey," in *Proceedings of the 2010 ACM Symposium on Applied Computing - SAC '10*, 2010, pp. 106–110.
- [9] K. Weber, B. Otto, and H. Österle, "One Size Does Not Fit All---A Contingency Approach to Data Governance," *Journal of Data and Information Quality*, vol. 1, no. 1, pp. 1–27, Jun. 2009.
- [10] S. O'Riain, E. Curry, and A. Harth, "XBRL and open data for global financial ecosystems: A linked data approach," *International Journal of Accounting Information Systems*, Mar. 2012.
- [11] U. Ul Hassan, S. O'Riain, and E. Curry, "Leveraging Matching Dependencies for Guided User Feedback in Linked Data Applications," in *9th International Workshop on Information Integration on the Web IIWeb2012*, 2012.
- [12] A. Miles and J. R. Pérez-Agüera, "SKOS: Simple Knowledge Organisation for the Web," *Cataloging & Classification Quarterly*, vol. 43, no. 3–4, pp. 69–83, Apr. 2007.
- [13] A. Isaac, J. Phipps, and D. Rubin, "SKOS Use Cases and Requirements." [Online]. Available: <http://www.w3.org/TR/skos-ucr/>. [Accessed: 28-Sep-2012].
- [14] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A crystallization point for the Web of Data," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 7, no. 3, pp. 154–165, Sep. 2009.
- [15] K. Belhajjame, N. W. Paton, S. M. Embury, A. A. A. Fernandes, and C. Hedeler, "Feedback-based annotation, selection and refinement of schema mappings for dataspaces," in *Proceedings of the 13th International Conference on Extending Database Technology - EDBT '10*, 2010, p. 573.
- [16] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD '08*, 2008, pp. 847–860.
- [17] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided Data Repair," *Proceedings of the VLDB Endowment*, vol. 4, no. 5, pp. 279–289, 2011.
- [18] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin, "CrowdDB : Answering Queries with Crowdsourcing," in *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, 2011, p. 61.
- [19] P. Wichmann, A. Borek, R. Kern, P. Woodall, A. K. Parlikad, and G. Satzger, "Exploring the 'Crowd' as Enabler of Better Information Quality," in *Proceedings of the 16th International Conference on Information Quality*, 2011, pp. 302–312.
- [20] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller, "Human-powered Sorts and Joins," *Proceedings of VLDB Endowment*, vol. 5, no. 1, 2012.
- [21] C. Grady and M. Lease, "Crowdsourcing document relevance assessment with Mechanical Turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 172–179.
- [22] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, 2012, p. 469.
- [23] S. Kochhar, S. Mazzocchi, and P. Paritosh, "The anatomy of a large-scale human computation engine," in *Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP '10*, 2010, pp. 10–17.
- [24] R. McCann, W. Shen, and A. Doan, "Matching Schemas in Online Communities: A Web 2.0 Approach," in *2008 IEEE 24th International Conference on Data Engineering*, 2008, vol. 00, pp. 110–119.
- [25] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, 2006, p. 43.
- [26] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch, "Broad expertise retrieval in sparse data environments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, 2007, p. 551.
- [27] K. Balog and M. De Rijke, "Determining expert profiles (with an application to expert finding)," in *Proceedings of the 20th international joint conference on Artificial intelligence*, 2007, pp. 2657–2662.
- [28] E. Curry, S. Hasan, and S. O'Riain, "Enterprise Energy Management using a Linked Dataspace for Energy Intelligence," in *Second IFIP Conference on Sustainable Internet and ICT for Sustainability*, 2012.

# **DOMAIN KNOWLEDGE BASED QUALITY FOR BUSINESS PROCESS MODELS**

(Research-in-Progress)

**Sarah Ayad**

CEDRIC-CNAM, 292 Rue Saint Martin, F-75141 Paris Cedex 03, France  
ayad\_sa1@auditeur.cnam.fr

**Samira Si-Said Cherfi**

CEDRIC-CNAM, 292 Rue Saint Martin, F-75141 Paris Cedex 03, France  
Samira.cherfi@cnam.fr

**Abstract** In recent years the problems related to modeling and improving business processes have been of growing interest. Indeed, companies are realizing the undeniable impact of a better understanding and management of business processes (BP) on the effectiveness, consistency, and transparency of their business operations. BP modeling aims at a better understanding of processes, allowing deciders to achieve strategic goals of the company. However, inexperienced systems analysts often lack domain knowledge leading and this affects the quality of models they produce. In this paper we propose to support this modeling effort with an approach that uses domain knowledge to improve the semantic quality of BP models. This approach relies on domain ontologies as a mean to capture domain knowledge and on metamodeling techniques. The main contribution of this paper is threefold: 1) the metamodels describing both a domain ontology and a BP model are described, 2) the alignment between the concepts of both metamodels is defined and illustrated, 3) a prototype implementing the approach is presented.

**Key Words:** Domain knowledge, Domain ontology, Semantic quality, Business process modeling, Quality improvement

## **1. INTRODUCTION**

Modeling is the intellectual activity of creating abstract and comprehensive representation of a system necessary to understand its existing or planned behavior. In practice, conceptual models have been recognized as playing an important role in communication and understanding among various stakeholders within a project. Business Process models are conceptual models supposed to give a complete description of the underlying business processes. Consequently, companies are today aware of the undeniable impact of a better tuning of business processes (BP) on the effectiveness, consistency and transparency of their business operations. This tuning requires a better understanding and an effective management of BP. However, to achieve the expected benefits it is necessary to rethink the approach of designing these processes. BP modeling is a prerequisite. It is now considered as an engineering activity aiming at providing the actors with a better understanding of the processes in which they are involved. But BP modeling is difficult. It is an expert task that needs to be performed by trained experts. And, what about quality? Quality can be defined as the total of properties and characteristics of a product or service that are relevant for satisfying specific and obvious requirements [1]. The business process modeling approaches share many similarities with conceptual modeling activities, but are much more complex [2] Indeed, a business process model captures a dynamic vision of the system through activities descriptions, generally done at a low level of abstraction; with a difficult issue of ending with a high level description for which a good acquaintance and understanding of domain knowledge is necessary. This is why the activity of modeling BP requires a high degree of pragmatic expertise generally referred to as empirical rules and

heuristics difficult to formalize and to share. Commercial tools for business process modeling activities mainly focus on the accuracy of models based on a set of syntactic criteria imposed by the notation and provide little or no guide to guarantee the quality of produced models.

We propose to assist the modeling activity with a quality centered approach that aims to exploit the domain knowledge. The domain knowledge in Information Systems discipline refers to knowledge provided by both methods and application domain [3]. In our approach we propose to exploit domain ontologies knowledge with alignment rules to identify similarities between BP models and domain ontologies elements. The aim is to improve the semantic completeness and expressiveness of BP models according to domain knowledge contained in the ontologies.

This paper is organized as follows. State of the art is described briefly in Section 2. The overall approach of our semantic is broadly described in the third section. The metamodels structuring both BP models and domain ontologies are described in detail in Section 4. Section 5 is dedicated to alignment rules. Finally Section 6 concludes and describes future research.

## **2. STATE OF THE ART**

A Business Process (BP) is a set of related activities that transform an input to create an output with added values [4]. Experts in information systems and professionals agree that the success of a company depends particularly of a good understanding of business processes [5]. To make a business process model understandable, reliable, and reusable it is important to ensure its quality. Several approaches that work in this direction exist in the literature. We have classified them into three categories: 1) Approaches focused on improving BP methods of analysis and design, 2) Process quality measurement, and 3) Process model quality measurement.

In the first category the approaches are intended to provide advice and best practices to ensure the best quality of models. The hypothesis is that improving the process development improves the quality of available products. As an illustration we can mention [6] where the authors propose a set of guides to improve various characteristics of a process model such as clarity, comprehensibility, or accuracy ("correctness"). Other authors focus on improving the comprehensibility of models by providing naming rules, documentation, and use of icons or symbols graphs [7, 8]. Other approaches, propose a set of best practices encapsulated in reusable and applicable patterns depending on the defined contexts.

The second category considers the quality level of business processes and their execution. In this family, we categorize the research on simulation and control of process as in [9] where the authors present a set of simulation tools for business process evaluation. Others focus on the verification of certain characteristics, when executing the process. In [10] for example, the authors present and discuss several techniques for the analysis of processes during execution such as verification, or for the discovery of a process ("process mining"), etc.

Our focus is in the third category that addresses the quality from the point of view of its evaluation and improvement. Process quality has been investigated in different disciplines. Consequently, a variety of standards have been introduced to define, manage, monitor, and improve that quality. In [11], the authors present a typology and an overall view of the business process model metrics. They mention the most important five measures: coupling, cohesion, complexity, modularity, and finally the size. The authors in [12] propose an approach based on GQM method (Goal-Question-Metric [13]) to help finding, among the set of quality characteristics, those that are relevant in a given framework and to deduce how to measure them. One of the characteristics that has been the subject of several proposals is the complexity [14, 15]. However, these studies are based primarily on structural characteristics of processes and their models.

Model complexity is directly related to their comprehensibility and their maintainability. The authors in [16] adapted the complexity metric of the software engineering process models and use it to study the complexity of general BP models. They then studied through an experiment the impact of process model complexity on their maintainability. In [17] the authors conducted an interesting experiment to try to understand the factors that impact the understandability of process models.

In conclusion, our analysis of the state of the art leads us to argue that the quality of BP model is mainly addressed in terms of structural and syntactic and rarely in terms of semantics. In the remainder of this paper, we present our approach which aims to go a step forward into a semantic quality based approach of BP model.

### **3. USING DOMAIN KNOWLEDGE TO IMPROVE QUALITY OF BUSINESS PROCESS MODELS**

Modeling activity in general and BP modelling in particular are creative activities conducted by modelers using a given notation or modelling language. The result is of course highly dependent on the modeler experience in the notation practice. It relies also on his/her interpretation of the reality, and on the decision he/she makes regarding the choice of concepts and details to be modeled. This explains the fact that several correct but different models could usually be generated from the same reality. However, these models are supposed to be faithful representations of the reality. Thus the definition of quality requirements for these models is, in fact, a mean to evaluate this modeling activity and ensure a better result. Many factors may be defined to characterize this quality. The semantic quality measures the degree of correspondence between the model and the domain. The semantic quality is related to both completeness and validity of the models; here the BP models [18].

To improve the quality of models produced, several approaches are possible:

- assistance in the development process phase by generic methodological guides from experience,
- measurement of the specifications quality,
- reusing approved specifications fragments etc.

Several authors pointed out the impact of lack of domain knowledge on the quality of produced models [19,20, 3]. In this paper, we propose to exploit knowledge of field, which are supposed to reflect the knowledge shared by a community of actors, in order to improve the quality of process models.

Many business domains has common domain knowledge more or less structured. For example, in the medical area, there exists a huge amount of knowledge on healthcare practices known as clinical pathways. In tourism business area, there exist classifications and even ontologies on tourism accommodations, tourism services etc..

The research question addressed by our approach is, given existing domain knowledge, how could we assist business process modelers using this knowledge to improve their way of modeling.

As our approach have to be generic and independent from the notations used for both domain knowledge expression and process modeling we use metamodels to express various BP models and domain ontologies. The metamodels are presented in sections 3.1.

Our approach proceeds as follows:

- first a mapping between BP model elements and domain ontology concepts is performed. This is necessary as the ontology and the BP models do not necessarily use the same vocabulary. The mapping rules are defined at the metamodels level to ensure their genericity. This part is presented in section 3.2.
- once the mappings validated by the analyst, a quality analysis based on the domain knowledge is performed on the BP model. The aim of this step is to detect some semantic quality defects. This step is detailed in section 3.3.
- finally, our approach suggests a set of improvement for each kind of detected quality defect. This part is presented in section 3.4.

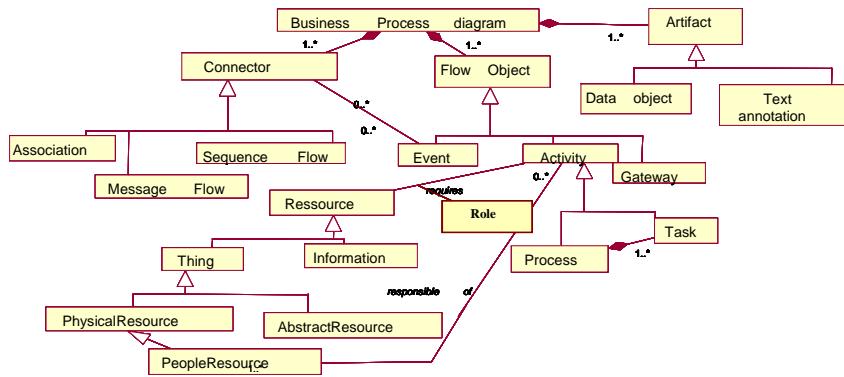
#### **3.1 Ontology and process model metamodels.**

In order to identify similarities between knowledge contained in the ontology and the one represented by the BP model, our approach relies on alignment. To ensure the generality of these rules, we have chosen to define them at a metamodeling level. Hence, the first contribution is the construction of metamodels representing ontologies and BP models.

### 3.1.1 Business Process Metamodel

There are several advantages of defining such a metamodel. First, the metamodel provides a synthetic vision of concepts used independently of specific notations helping in the understandability of models. Second, instead of defining mapping rules for each couple of BP modeling notation and ontology language we define the rules only at the metamodel level. Finally, since we consider that domain knowledge contains also knowledge embedded in methods and consequently in notations, we will use metamodels to integrate completeness, validation and correctness rules defined by BP notations to enrich our actual vision of domain knowledge.

The metamodel defined in this section and shown at Figure 1 was constructed as a synthesis of a selection of concepts proposed by several authors and according to several notations and more specifically the work presented in [21,22].



**Fig 1. Business process metamodel**

A business process model is composed of flows of objects and connectors. A flow object can be an event, an activity or a gateway [22]. An event that occurs is a fact and impacts the progress of a process. Our events can be of three types: initial, intermediate and final. An activity can be an atomic task if it is not decomposable or a process if it is complex and has a visible structure. A gateway is a mechanism that can manage the convergence or divergence of activities flow. A connecting element can be an association, a sequence or a message flow. An association is used as a simple link between two concepts. The sequence flow defines an execution order of activities. A message flow is used to represent exchange of information between two participants in the process.

Activities refer to resources. A resource is a concept which includes abstract concepts such as the human agent responsible for execution of the activity and information produced or consumed by it. The exact role of the resource in the process is explained by the concept of role.

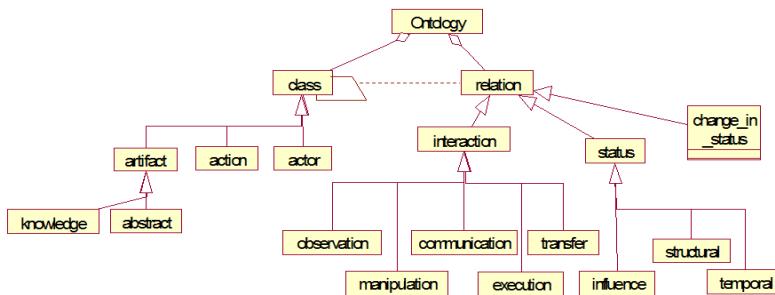
### 3.1.2. Ontology Metamodel

The ontology metamodel allows representing domain ontologies using the same concepts independently of the language for their implementation. There are several contributions in literature concerning ontology metamodeling. The authors in [23] introduced simple concepts and constructors (negation, conjunction, disjunction) to define complex concepts. They also defined several relationships including inheritance links, instantiation and constraints. In [24.] five types of concepts have been proposed to represent the functional requirements (function, object, and environment) and non-functional requirements (constraints, quality). In our approach, we consider an ontology as a set of classes and relationships. This vision is largely adopted. We distinguish between three types of concepts of type class: actor, action and artifact.

- An actor is an independent entity, able to perform actions.

- An action represents the execution of an action.
- An artifact is an inanimate object incapable of performing an action. An artifact may represent an information or an abstract concept.

However, most of metamodels take into account two kinds of relationships, namely inheritance and structural relationships. For the needs of our approach we adapted the classification of relationships proposed by [25], which has been initially defined to analyze semantics of relationships within a relational database. This classification offers several types of relationships allowing us to characterize precisely the nature of links between concepts.



**Fig 2. Ontology Metamodel (an extract)**

Relations are first decomposed into three categories:

- Status: represents relationships that may be structural (inheritance, composition, instantiation, etc.), influence (own, control, creation, destroy, etc.), or temporal (follow, require, etc.).
- Change of status: reveals the occurrence of remarkable events. This type of relationship is primarily used to express the interdependence of status in the life cycle of an entity.
- Interaction: represents short-term relationships between entities. Several semantic relations are defined for interactions such as communication, observation, execution, etc.

Figure 2 illustrates some concepts of the ontology metamodel.

### 3.2. Identifying Model-Ontology similarities

In the first step, the approach consists in discovering the mappings between business process model elements and the domain ontology elements. To make these alignment rules generic and independent of both the BP modelling notation and the ontology implementation language, we have defined two metamodels namely a BP metamodel and an ontology metamodel. The alignment rules aim to identify similarities between the process model elements and the domain ontology concepts. Once these similarities identified they serve as input for both semantic quality evaluation and improvements activities.

We have defined two kinds of mapping, namely type-based mapping and semantics-based mapping.

#### 3.2.1. Type-based mapping

This mapping involves the types of concepts in order to establish correspondences between the concepts at the meta-level. These correspondences allow reconciliation based on the types of concepts independently of their meaning. These rules are still essential to avoid typing errors. An extract of predefined metamodel concepts mappings is given in Table 1.

BP model metamodel concept	Domain Ontology metamodel concept
People resource	Actor
Abstract resource	Abstract
Information resource	Knowledge
Process / activity	Action

Table 8. Concept alignment

Similarly, we have established mappings between metamodel relations of BPM and those of the ontology metamodel. The result is given in Table 2.

BP model metamodel connectors	Domain Ontology metamodel relations
Sequence Flow Message Flow	Temporal Communication Transfer Execution Manipulation Observation Influence
Role	

Table 2. Relation alignment

### 3.2.2. Semantics-based mapping

The second type of mapping is richer, being based on the semantics of concepts. Let  $O$  a domain ontology and  $o \in O$  a concept from this ontology. Each concept has a set of synonyms (set of words), hyponyms, hypernyms and keywords related to it :

Note that we say the two names are partially equivalent if they have common names.

There are four classes of matching rules. The rules are all defined as functions having as input a BP model concept  $bpi$  and returning one or several concepts from the domain ontology.

The similarity computation uses the names of concepts, the synonyms, and keywords associated to ontology concepts. It is based on wordnet and distance algorithms from literature such as Resnik information content [26], Wu & Palmer path length [27] , Purandare & Pederson context vectors [28]. We use these algorithms through equivalence (applies when the names are composed of one word) and partial equivalence (applies when names are composed of several words) functions:

Our approach uses five types of semantic similarity functions:

- **Name based similarity:** returns a set of ontology concepts having the same name that the BP model concept.
- **Synonyms based similarity:** returns a set of ontology concepts having at least one synonym syntactically equivalent to the BP model element. We compute such similarity when no ontology concept is returned by the name based similarity.
- **Hypernyms Similarity:** based on the results obtained by name based and synonyms based similarity, this function returns for each ontology concepts of the result the set of its hypernyms. This allows findings from the domain knowledge concepts that are more general than those used in the BP models and that could help in completing the model by exploiting the related concepts and relationships.
- **Hyponyms Similarity:** based on the results obtained by name based and synonyms based similarity, this function returns for each ontology concepts of the result the set of its hyponyms. This allows précising BP models modeling elements by using more specific concepts provid-

ed by the ontology.

- **Keywords:** returns a set of ontology concepts having at least one keyword syntactically equivalent to the BP model element. We compute such similarity when no ontology concept is returned by the name based or the synonyms based similarity functions.

### 3.3. Quality defects Detection

First, similarities have to be identified between a BP model element, let it be  $bpm_i$ , and an element from the domain ontology  $o_i$ . Our approach exploits the knowledge from the domain ontology related to  $o_i$  to detect and measure semantic quality deficiencies. In order to exploit the knowledge related to  $o_i$  we use the mappings identified in section 3.2.

Second, we have identified a set of what we call quality deficiencies. These deficiencies result from modeling choices producing models that do not cover the intended requirements or with low expressiveness.

- **Ambiguity:** Ambiguity results from using different names and constructs to express the same reality. This makes models unclear and creates confusion when trying to understand them.
- **Completeness:** Completeness is related to an incomplete representation of the real world. This incompleteness can result from the complexity of concepts for which only a sub-set of the description is captured within the process model. One of the metrics that show the incompleteness is the Number of Human Resources which exploits the structure of the concepts, as an activity should have a Human resource responsible of its execution.
- **Abstraction level:** Abstraction level is related to the use of the suitable level of generality. Indeed, in some cases, using general concepts instead of specific and precise ones can decrease the efficiency of the processes. On the contrary, using very specialized terms may decrease the understandability of the models. The relevant choice of an abstraction level depends on several factors among which we can mention the nature of audience (developers or users), the objective of the model (explanation or implementation), etc.
- **Meaningless states:** meaningless states correspond to states and constructs from the models for which no correspondence is found in the corresponding ontology. This decreases the relevance of models and has an impact on its intelligibility.

### 3.4. Quality Defects correction

Third, the quality improvement activity consists in suggesting to the analyst or the quality expert a set of improvement guidelines to improve the quality of theirs models.

- **Correcting ambiguity defects:** consists in replacing the chosen concepts by an other more adequate from the list of synonyms proposed by the tool. Once again, the ontology helps by providing the list of synonyms from the ontology and the analyst has to choose among them the most suitable term.
- **Correcting incompleteness defects:** In case of incompleteness, the analyst can rely on the knowledge provided by the ontology to complete the missing parts of the model. On the other hand semantic constraint are defined on BP metamodel concepts level that my show the user what is it missing in his model. As each activity should have a human resource responsible of its execution so if the number of human resource is equal to zero means there is a human resource should be mapped to the activity. Additionally, keywords provided by the ontology can help the user to complete its model by requirements missing from his model.
- **Correcting the abstraction level:** Likewise, the user can rely on the knowledge provided by the ontology (Hypernyms/hyponyms) to choose the adequate abstraction level of the concept that will make the model more comprehensible.
- **Correcting the meaningless states:** When a BP model's element does not match any concept from the

ontology this could mean that this element is out of the domain.

## 4. Implementation of the approach

To illustrate our approach, we consider the example of "mission order" process. Our approach takes as input a business process model under construction and a domain ontology.

### 4.1. An illustrating example

Our BP model case study represents the business process followed by a researcher/ employee in our university who plans to attend a conference or to participate to an exchange program. In this case he/she should fill an official "Mission Order". This means in particular that he/she is covered by insurance, that the university pays the plane/train ticket, that the employee get some money in advance for his/her fees and that he/she is reimbursed after mission. The employee has first to get an authorization from his/her service/laboratory director/administrator. The administrator analyzes the request. Based on his decision, the employee fills a form called mission order (MO), sends it to the financial service and at the same time carries out mission formalities. The financial service calculates the reimbursement costs only if the employee sends mission costs proofs (tickets, bills etc.). Finally the BP is closed by the transfer activity. The BP model expressed in BPMN is presented in Figure 3.

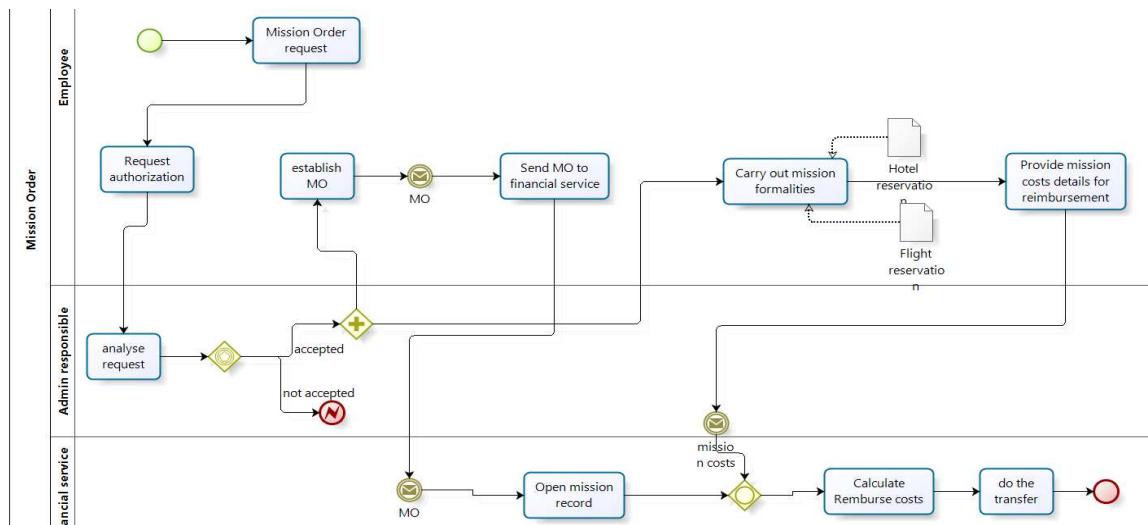


Fig 3. Instantiation of the BPM metamodel

In addition, an extract of the domain ontology "mission plan" is represented in Figure 4. The actor "employee" which is linked to actor "PhD student" by an is-a relation of type structural (status), is related to the action "Request authorization" by an influence relation. Also the actor "Financial service" is related by a synonymy relation to actor "Accounting service". A "requires" relationship relates "Request authorization" with the action "Ask for Delay". Moreover, an is-a relationship relates different actions to show different abstraction levels. For example, "Estimate costs" action is more general than "calculate compensation" action. "Analyze request" is more general than "Examine the applicant" and "Correct the application".

Finally temporal relations between "Send MO to financial service" action and "calculate reimbursement costs" action show the order of their execution.

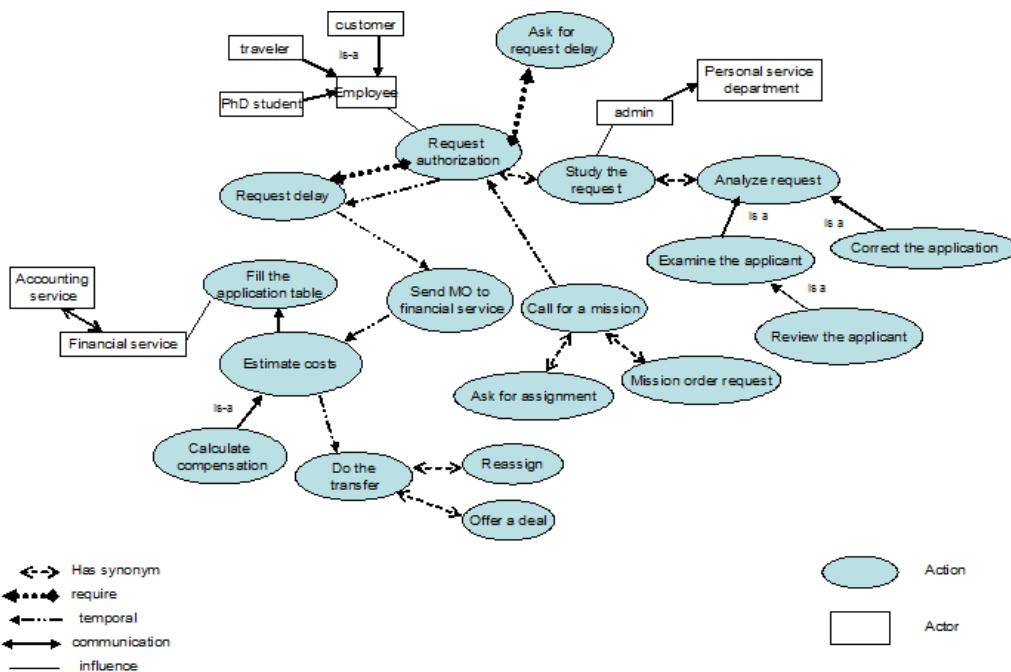


Fig 4. An ontology extract

We started implementing the approach in a prototype. The architecture is detailed in section 4.2.

## 4.2 Prototype Architecture

In order to support our approach for we are currently implementing it within a prototype. The general architecture of the prototype is presented in Figure 5.

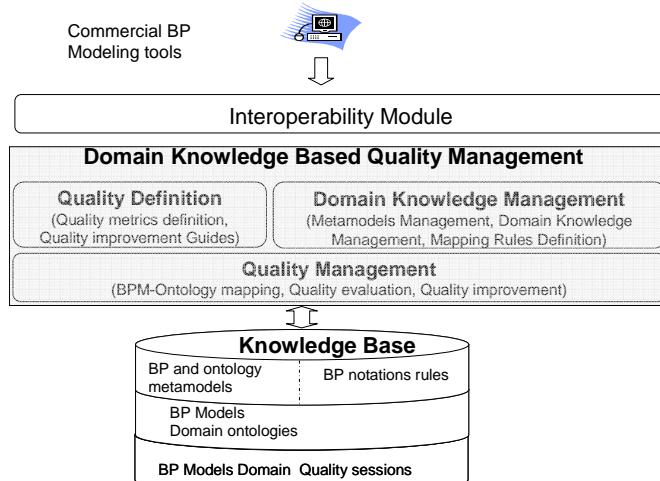


Fig. 5. Prototype Architecture

The overall architecture is structured around three main modules: the Interoperability module, the Domain Knowledge Based Quality Management (DKBQM) module and the Knowledge Base.

**Interoperability Module** allows supporting quality evaluation of BP models produced by commercial or free BP modeling tools. Indeed, the current version of the prototype is interoperable with several BP modeling tools such as Bizagi, Enterprise Architect or Star UML. These tools use different modeling notations: BPMN for Bizagi and enterprise Architect and Eriksson and Penker's notation for StarUML for example. These tools provide export utilities based on XML or XPDL. Moreover, the several BP modeling tools use different notations. To solve this problem, we have developed an interoperability module able to deal with several export languages. This module also annotates the exported models to make the BP models compatible with the metamodel presented in section 3.1.1.

**The DKBQM module** offers utilities to:

- define quality metrics. In the actual version quality metrics are written in java. A further version will include a metrics specification language based on the metamodels.
- define quality improvement guides. The improvement guides are written in OCL [29].
- define and improve the metamodels. The ontology metamodel is implemented as OWL [30] classes within protégé [31]. The domain ontologies are defined as instances of these classes.
- define notation constraints. Modeling notations have some rules helping analysts verifying correctness and syntactic completeness of models. We have included this kind of knowledge as domain knowledge. We have actually integrated some correctness rules written in OCL.
- define mapping rules. These rules allow finding similarities between domain knowledge ontology concepts and BP model elements. These rules use word similarity distances from the literature. Actually the prototype implements five algorithms for words similarity detection. More information on these algorithms is given in section 3.3.2.
- manage quality sessions. These sessions consist in selecting a PB model, a domain ontology and the application of the approach as illustrated in section 4.3.

**The Knowledge Base** stores the several artifacts (metamodels, domain ontologies, BP models, traces of evaluation sessions, several versions of BP models etc.).

## 4.3 Illustrating the approach

This section illustrates the approach on the BP models illustrated in figure 3.

### 4.3.1 Similarities and quality defect detection

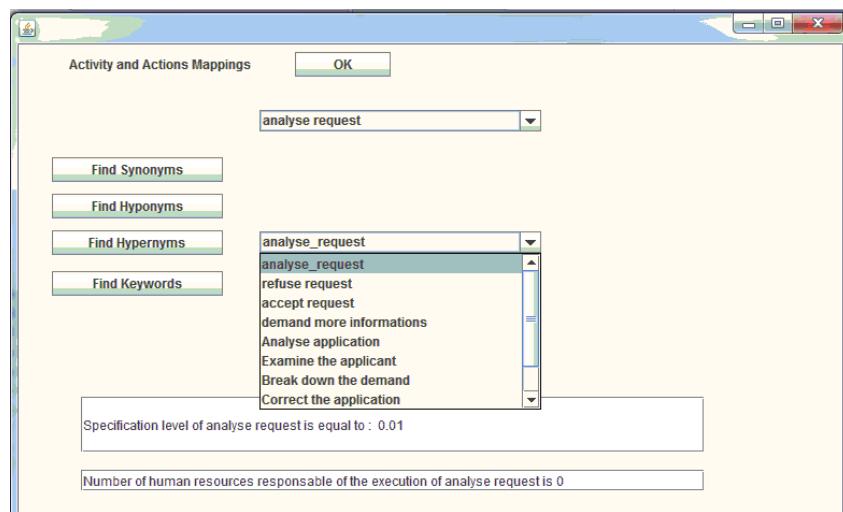
The first step applies mapping rules between the BP model and the domain ontology from figure 4. Based on the type based mappings, our prototype computes a list of action/actors for each activity/Human resource present in the BP model. These mappings are refined based on the semantic mapping rules. The result is a list of synonyms/hypernyms/ hyponyms given in table 3. Notice that the mapping is actually a word based mapping. We are actually improving the similarity distance definition to take into account sentences. This will improve the relevancy of results obtained in table 3.

BP activity	Action synonyms
Mission order request	<ul style="list-style-type: none"> <li>• Ask for assignment</li> <li>• Call for a mission</li> <li>• An operation command</li> <li>• Postulate to a work mission</li> <li>• Fill for a foreign mission</li> </ul>
Provide mission costs details for reimbursement	<ul style="list-style-type: none"> <li>• Supply for reimbursement</li> <li>• Provide the total spent</li> <li>• Provide the compensation paid</li> </ul>
Do the transfer	<ul style="list-style-type: none"> <li>• Reassign</li> <li>• Offer a deal</li> </ul>

Table 3. Extract of Synonymy results

Concerning the quality evaluation, the more an element from BP model have synonyms the more it is ambiguous. So based on results in table3, we conclude that "Do the transfer" activity is less ambiguous than "Mission order request" activity that corresponds to different concepts from the ontology.

Furthermore, based on the knowledge provided by the ontology some elements from the BP model are more general or less general than the concepts defined in the ontology. For example, the activity "analyse request" have many hyponyms as shown in figure 5. The user has to decide to maintain the generality level defined in the BP model or to change it for a more precise description by choosing one of the hyponyms. This could generate other changes. Changing an activity may imply changing the actor responsible of it and/or change the information resources required or produced by the activity etc. These changes are deduced from the ontology.



**Fig 5. hyponym mappings**

An extract of computed hyponyms are shown in table4.

BP activity	Hyponyms
Analyse request	<ul style="list-style-type: none"> <li>• Analyse application</li> <li>• Examine the applicant</li> <li>• Break down the demand</li> <li>• Correct the application</li> <li>• Review the application</li> </ul>
Carry out mission formalities	<ul style="list-style-type: none"> <li>• Fetch the mission order</li> <li>• Open mission recorder</li> <li>• approve order</li> <li>• validate the mission</li> </ul>
Provide mission costs details for reimbursement	<ul style="list-style-type: none"> <li>• Provide mission in France costs details</li> <li>• Provide mission in Europe costs details</li> <li>• Support with business expense</li> <li>• Provide prices</li> <li>• Provide borrowing costs/ production costs</li> </ul>
Estimate costs	<ul style="list-style-type: none"> <li>• Calculate compensation</li> <li>• Estimate costs</li> </ul>
Do the transfer	<ul style="list-style-type: none"> <li>• Delivery</li> <li>• Lend</li> <li>• Conveyance</li> </ul>

**Table 4. Extract of Hyponyms Results**

Additionally, more general concepts are suggested to each activity/human resource, for example "Administrator" can be replaced by more specific concepts such as "Decision maker" or "Executive". The approach proposes the several alternatives and the analyst has to decide about the more appropriate choice.

BP human resource	Hypernym
Employee	<ul style="list-style-type: none"> <li>Worker</li> </ul>
Administrator	<ul style="list-style-type: none"> <li>Decision maker</li> <li>Executive</li> </ul>
Financial Service	<ul style="list-style-type: none"> <li>Financial department</li> <li>Supplier</li> <li>Contractor</li> </ul>

Table 5. Extract of Hypernyms results

Finally, the context provides knowledge allowing the BP model enrichment and/or completion. This context is defined through the keys words and relationships related to domain ontology concepts. The keywords returned by keywords function propose to the user domain concepts that may help to enrich his model. These concepts are ontology concepts similarly related to the BPM concept chosen.

BPM activity	Ontology action	Related concept	
		concept	relationship
Establish MO	Establish Mission Order	Documents required (abstract) Delay (knowledge)	Assigned to
Reserve flight and hotel	Carry out mission formalities	Flight reservation (abstract) Hotel reservation (abstract)	requires
Calculate reimburse costs	Estimate costs	Commission (Knowledge) invoice	requires
Analyse request	Analyse the request	Mission record (abstract)	manipulate
Do the transfer	transfer	RIB (knowledge)	requires

Table 6. Extract of keywords result

Based on the results shown in table 6, time delay concept can be related to "establish MO" activity. And time delay is an important condition in any administrative file request. As a result the user can add timer and information about mission order time delay in his/her process. Also the activity "Analyse request" is done by manipulating the mission record so that it has to be an output sent to the administrator.

#### 4.4 Improving detected quality defects

The quality improvement activity consists in suggesting to the analyst a set of improvement guidelines to improve the quality of their models.

Correcting ambiguity defects: The ambiguity hampers the possibility to decide whether the statement from the model is meaningful according to the domain. In fact "Mission order request" activity does not present the domain specific contract. It's an application to fill so the verb "fill for" is more adequate and mission order is a foreign mission. As a result " Fill for a foreign mission " activity is more adequate.

Correction abstraction level defects: aims to use the suitable level of generality.

- In some cases, using general concepts instead of specific and precise ones can decrease the efficiency of the processes. For example, "analyse request" activity is a general activity that may skip a lot of detailed activities. So based on the hyponyms provided by the ontology the analyst can replace it by "Examine the applicant", "Review the application" and "Correct the application" activities given by the ontology.
- In other cases, using very specialized terms may decrease the understandability of the models. For example, the actor "administrator" can be replaced by "decision maker".

Correcting incompleteness defects: Enrich and complete our BPM can be done in two ways separately.

- Using keywords enriches the model with new activities and new resources. As we can see in table 6, the activity "establish MO" can be related to mission order time delay so a timer is added as output for this activity. And "carry out mission formalities" is enriched by adding abstract resources "mission program" and "mission date" as inputs. Additionally, "Do transfer" is related semantically to the concept "RIB" because Resnik information content metric [26] is higher than five which means that the distance from the root to these two concept is small. We can suggest to the user to follow "do transfer" activity by "ask for the RIB".
- Also one of the semantic constraints defined on the BP metamodel concepts concern HR semantic constraint which consist that each activity should have a human resource responsible of its execution. Our tool shows the number of Human resources responsible of the execution for each activity. As a result the user has to complete his model by matching the activity to a human resource.

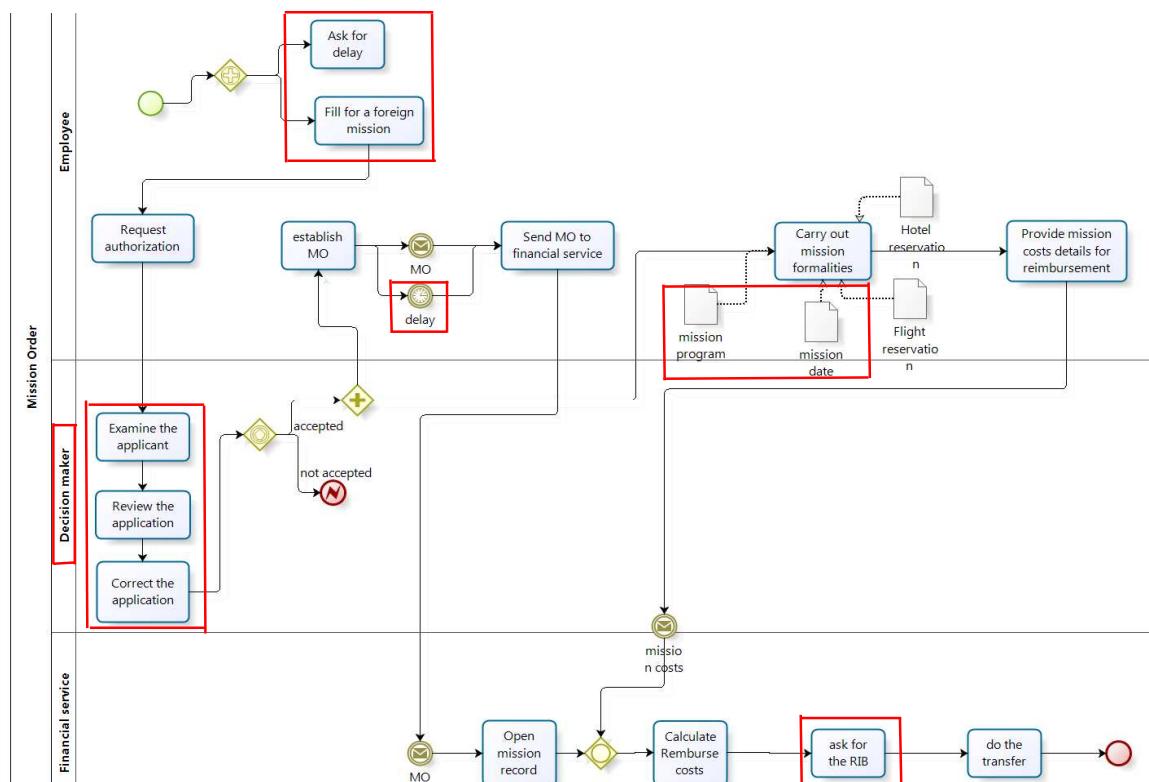


Figure 6. The BP model resulting from quality improvement

## **CONCLUSION**

The paper presented a work in progress on how to improve quality of business process models by exploiting domain knowledge. We presented an approach based on semantic quality analysis and quality improvement using of domain knowledge ontologies. The approach takes into account the variety of business process model notations by using a metamodel. The domain ontology is represented by the means of ontologies where semantics is enriched by using several kinds of relationships among the concepts. This knowledge is structured again through ontology metamodel.

An implementation and an example aiming to illustrate the approach are described.

The article presents an ongoing work that requires further research to improve it. We are actually working on the improvement of distances computing the mappings between the BP models and the ontologies. We are also working on quality metrics definition. This part of the work is not presented here. Finally, we are collecting real world domain ontologies and process models to conduct an experiment to validate the approach on a real case study.

To conclude, the main contribution is the enrichment of semantic model quality evaluation through the use of domain ontologies. The paper objective is to demonstrate the feasibility of the approach that we believe could be generalized to other conceptual models such as data models, development process models, requirement models etc.

## **REFERENCES**

- [1] The International Standards Organisation ISO.
- [2] Vanderfeesten I., Cardoso J., Mendling J., Reijers, Alast: Quality Metrics for Business Process Models. In: Fischer, L. (ed.) BPM and Workflow Handbook 2007 (May 2007), pp. 179-190. Key: citeulike:5757678
- [3] Khatri V. and Vessey I., Information use in solving a well-structured IS problem: the roles of IS and application domain knowledge. In Proceedings of the 29th international conference on Conceptual modeling (ER'10), Springer-Verlag, Berlin, Heidelberg, 46-58.
- [4] Johansson H.J. et al. (1993), Business Process Reengineering: BreakPoint Strategies for Market Dominance, John Wiley & Sons
- [5] Aguilar-Savén R. S., Business process modelling: Review and framework, International Journal of Production Economics, Volume 90, Issue 2, 28 July 2004, Pages 129-149, ISSN 0925-5273.
- [6] Becker J, Rosemann M., Uthmann C. V.: Guidelines of Business Process Modeling. Business Process Management 2000: 30-49
- [7] Mendling J., Reijers, Cardoso: What Makes Process Models Understandable? In: Lecture Notes in Computer Science, 2007, Volume 4714/2007, 48-63, DOI: 10.1007/978-3-540-75183-0\_4
- [8] Cardoso J., Jan Mendling, Gustaf Neumann, Hajo A. Reijers: A Discourse on Complexity of Process Models. Business Process Management Workshops 2006: 117-128
- [9] Van der Aalst W.M.P, ter Hofstede A.H.M., Kiepuszewski B, and Barros and A.P Workflow Patterns. Distributed and Parallel Databases, 14(3), pages 5-51, July 2003.
- [10] Jansen-Vullers M.and Netjes M.: Business Process Simulation: A Tool Survey. In Workshop and Tutorial on Practical Use of Coloured Petri Nets and the CPN Tools, Aarhus, Denmark, October 2006.
- [11] Van der Aalst W. M. P." Challenges in Business Process Analysis" in ICEIS (Selected Papers) 2007: 27-42
- [12] Vanderfeesten I., Reijers, Mendling J., aalst, Cardos: On a quest for good Process Models: The Cross-Connectivity Metric. In: Advanced Information Systems Engineering (20th International Conference, CAiSE'08, Montpellier, France, June 18-20, 2008, Proceedings) / Ed. Z. Bellahsène,

- M. Léonard. - Berlin : Springer, 2008. - ISBN 978-3-540-69533-2. - (Lecture Notes in Computer Science ; 5074). - p. 480-494
- [13] Basili V.R., Caldiera G., Rombach H.D., The Goal Question Metric Approach. Encyclopedia of Software Engineering, vol. 2, September 1994, p. 528-538.
- [14] Ghani A., Wei G. M., Muketha G. M., Wen W. P., Complexity Metrics for Measuring the Understandability and Maintainability of business process Models using goal-question-Metric (GQM). In: International journal of computer science and network security, Vol. 8 N° 5, p. 219-225, May 2008
- [15] Rolon E., Ruiz, Garcia , Piattini M. : Applying Software metrics to evaluate Business Process Models. In: CLEIEI Ecronic Journal, volume 9, number1, paper 5, june 2006
- [16] Gruhn V., Laue R.; Complexity metrics for business process models ; 9th international conference on business information systems (BIS 2006), volume 85 of Lecture Notes in Informatics
- [17] Mendling J., Recker J. and Reijers H.A. " On the Usage of Labels and Icons in Business Process Modeling" in IJISMD 1(2): 40-58 (2010)
- [18] Krogstie J., Lindland O. I., Sindre G., Defining quality aspects for conceptual models. In Proceedings of the IFIP international working conference on Information system concepts. 1995, pp 216-231
- [19] Davies I., Green P., Rosemann M., Indulska M., and Gallo S. How do practitioners use conceptual modeling in practice?. Data Knowl. Eng. 58, 3 (September 2006), 358-380
- [20] Shanks G., Conceptual Data Modelling: an empirical study of expert and novice data modellers, Australasian Journal of Information Systems, Vol 4, No 2 (1997)
- [21] Eriksson H.E., Magnus Penker, Business Modeling With UML: Business Patterns at Work, John Wiley & Sons, Inc. New York, NY, USA ©2000 ISBN:0471295515
- [22] Loja L., Neto V., Costa S., Oliveira J. : A business process metamodel for enterprise information systems automatic generation. In Brazilian Workshop on Model-Driven development. Brazil 2010.
- [23] Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. Knowl. Acquis. 5, 2 (June 1993), 199-220.
- [24] Kaiya H. and Saeki M. Using Domain Ontology as Domain Knowledge for Requirements Elicitation. In Proceedings of the 14th IEEE International Requirements Engineering Conference (RE '06). IEEE Computer Society, Washington, DC, USA, 186-195.
- [25] Purao S. and Storey Veda C. 2005. A multi-layered ontology for comparing relationship semantics in conceptual models of databases. Appl. Ontol. 1, 1 (January 2005), 117-139.
- [26] Resnik P, Using Information Content to Evaluate Semantic Similarity in a Taxanomy, International Joint Conference for Artificial 1995
- [27] Wu. and Palmer M. 1994. Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, pages 133–138.
- [28] Purandare A.and Pedersen T., Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces, HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning 2004
- [29] Object Management Group: UML 2.2 OCL specification (2010). Available as OMG document formal/2010-02-01
- [30] W3, C., OWL Web Ontology Language Overview, W3C Recommendation, February 2004
- [31] Protégé: <http://protege.stanford.edu>

# **APC-SIMULATOR: DEMONSTRATING THE EFFECTS OF TECHNICAL AND SEMANTIC ERRORS IN THE ACCURACY OF HOSPITAL REPORTING**

(Completed Academic Paper)

**Sami Laine**

Aalto University, School of Science, Finland  
<mailto:sami.k.laine@aalto.fi>

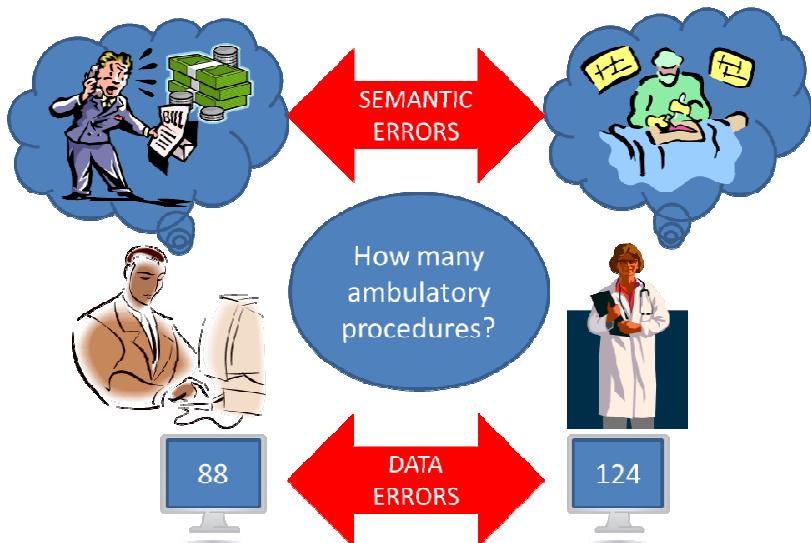
**Abstract:** In this paper we present the development of the APC-simulator (Ambulatory Procedure Calculator-simulator). It is an experimental tool for illustrating the effects of technical and semantic errors in hospital reporting. Many disciplines are concerned that currently unrecognized inaccuracies in raw data and derived information products endanger the validity of management decisions, policy recommendations and statistical research results. Healthcare reporting development experiences presented in this experiment support these concerns. During these development projects many inaccuracies were found to be significant. They often result from intertwining contextual reasons rather than from random failures to provide correct data. In this experiment, constructive research methods are applied to demonstrate interdependencies between technical and semantic factors with an experimental dashboard and empirical information. First, the dashboard calculates the errors that result from the technical data flow. Then descriptive empirical information explains what errors mean in reality. The experiment suggests that many inaccuracies could be fixed by redefining general concepts semantically to match their local meaning. In addition, the entire information production process should be monitored transparently from technical and human perspectives to avoid making invalid decisions.

**Key Words:** Administrative Data, Constructive Research, Data Accuracy, Data Quality, Healthcare Information Systems, Information Accuracy, Information Quality, Information Production Process, Information Product, Medical Informatics

## **INTRODUCTION**

In hospitals data are stored into many IT systems for various purposes: patient care, hospital administration, policy making and medical research. Unfortunately, data often contain errors or they can be interpreted in many different ways by various users. These problems have significant consequences since the validity of any decision depends on the quality and the meaning of the data it is based on.

Data quality challenges, such as errors or semantic ambiguity, become significant problems when data are used for different purposes in various secondary contexts (figure 1). The reason for this is that local data quality errors and semantic mismatches are often not recognized or understood well in the secondary usage. The practical significance of data errors and semantic heterogeneity becomes visible after noting that in our example the counts of ambulatory procedures in administrative reports derived from different systems varied by almost 50%.



**Figure 1: Data might contain wrong values or semantic mismatches between contexts.**

For example, the count of procedures might be based on bills or clinical events resulting in semantic mismatches. Also, the counts of data instances derived from data sources might differ from reality.

To find out the reasons for inconsistencies in administrative reports and to guarantee the validity of decisions one must trace the entire data flow from the first data entry situation to the final utilization case of the information products. This principle might seem self-evident, but in practice it is far from easy to accomplish in collaborative work situations of numerous actors. In practice it is very hard to know answers to questions such as:

- What actual reasons were there to enter the specific data instance?
- What exactly has been done to each data element across technical data flows?
- Where have the derived information products been delivered?
- What are decision makers actually doing with the information products?

Answers to previous questions have been investigated time and again while developing information systems for healthcare service providers. The motivation of the research was to find out systematically what kind of errors there exists in the information production processes. These findings could be then used to analyze what kind of consequences the identified errors have to the information products and decision making. This problem was studied by selecting a very limited information product, identifying errors occurring in the production process and analyzing their effects to the information product and decision making.

The next section introduces briefly theoretical issues of information accuracy and traceability. The third section describes the methods and the process of this study. The fourth section describes the characteristics of the constructed APC-simulator. The fifth section describes the simulation input and results while the sixth section continues to analyze the details of error calculations and contextual issues. The seventh section discusses the limitations of the study and its relationship to the reality of hospital reporting. The last section summarizes the findings and implications for further research.

## PERSPECTIVES ON INFORMATION ACCURACY

Inaccurate information means that data values differ from the true state of the phenomena they try to represent. Accuracy is a widely studied information quality topic. There exist useful practical [1] and academic [2] texts focusing solely on the accuracy dimension. Lately different disciplines, such as medical sciences [3] and statistical sciences [4], have expressed their concern about the accuracy of

currently available data from their own perspectives. They are all concerned that currently unrecognized inaccuracies in raw data and derived information products endanger the validity of current management decisions, policy recommendations and statistical research results. Decisions based on erroneous and biased data could lead to unnecessary costs and consequences would not be the anticipated ones.

### ***Classification of Semantic Information Accuracy***

In the APC-simulator, adjustable accuracy error rates are inserted into the calculations of a simple information product. Therefore different types of accuracy errors contributing to the error rates should be defined in more detail. Healthcare data suffers widely from inaccuracies [5, 6]. Medical science researchers are concerned that data can often be systematically biased rather than randomly erroneous [7, 8]. Data quality research has pointed out that inaccurate data are often not simply false but rather semantically heterogeneous [9]. Combining these perspectives information accuracy can be categorized in the following way (table 1).

Semantically “wrong” data.		Semantically “correct” data.	
Random errors	Systematic errors	Representational heterogeneity	Ontological heterogeneity

**Table 1: Semantic accuracy can be affected by actual errors but also by semantic mismatches between contexts.**

There are two types of semantically “wrong” and therefore inaccurate data: random errors and systematic errors. Random errors are individual failures to represent the true state of the phenomena. A doctor simply did not recognize the disease. A nurse forgot to press the confirmation icon in the user interface while sending digital referral. Systematic errors are patterns of failures and mistakes occurring repeatedly for similar reasons and in a similar way. A complicated user interface causes local data update problems in the same way all the time. Alternative user-friendly user interface provides better quality data leading to systematic bias between otherwise identical data sources. Some errors are related to the phenomena itself but many are results of the particular information production process.

There are also two types of semantically “correct” but still potentially inaccurate data: representational and ontological heterogeneity [9]. Representational heterogeneity means different formats of data requiring straightforward conversions. The currency can be represented in euros or pounds. The counts of sold services could be based on fiscal year or calendar year. Ontological heterogeneity refers to a problem of slightly different concepts. It often results from actual usage of generalized and ambiguous concepts. The count of ambulatory procedures performed in the hospital might be based on clinical events such as patients entering the operation room or financial issues like bills sent to the patients. Semantic heterogeneity leads easily to huge systematic errors unless data are used carefully in all contexts.

### ***Traceability in Information Production Processes***

Semantic inaccuracies can occur in any phase across the entire information production process. Tracing data flows and their error rates, as will be demonstrated later, is not a new idea. Ballou and Pazer presented decades ago a way to model data and process quality in multi-input and multi-output information systems [10]. They used data flow diagrams and algorithms to analyze impact of errors and quality controls to the selected outputs. Their work is continued to a more general modeling of information manufacturing systems [11] and even to a complete data quality governance methodology based on Total Quality Management [12, 13]. For example, Data Quality Flow Models (DQFM) [14] and IP Maps [15] are examples of methods used to track down data flows.

Technical aspects of tracing data flows in software systems are studied under the label of data lineage and provenance. According to Cui the goal of data lineage research is to provide systems and algorithms capable of tracing information products back to the sources which were involved in their production [16].

She notes coarse-grained data lineage systems provide metadata and process descriptions of data flows and storages back to the source in schema level. Her own research focused on fine-grained data lineage can trace data back to the original data elements in instance level.

### ***Theoretical Considerations of Tracing Accuracy Errors***

Information flows and errors are often traced by methods from two complementary perspectives: organizational information management and technical software systems. Information management methods (e.g., DQFM, IP Map) are used to model information production processes in higher abstraction level. They describe networks of processing and controlling units which are either human or technical actors. Software engineering methods produce features to technical systems (e.g., data lineage capability). They are used to trace backwards or forwards implemented software elements such as tables, processes, objects and instances. Together these complementary methods provide necessary formal information about information production processes.

Unfortunately, in practice automatic data impact analysis and data lineage functions are not yet widely used in heterogeneous software-system environments. For example, Gartner [17] has stated that the lack of a single unified metadata layer or capability that spans Enterprise Business Intelligence platforms components is a problem for many software providers and their technology platforms. In practice Gartner means technical capabilities related to the traceability of information flows: metadata modeling, data impact analysis and data lineage. IP Maps and similar modeling methods can be used to document information production processes when automatic technological support is unavailable or technically too complex for multidisciplinary communication. The problem is that manual modeling and documentation is laborious and error prone while tracing calculation rules and semantic mismatches across information production processes. In the future, these methods should be integrated to combine their strengths.

Most importantly, from the perspective of information accuracy, all previously mentioned methods suffer from the same problem. Their current ability to capture, store and provide contextual information about relevant human factors is limited. Therefore, current state-of-art research is extending IP Maps to CEIP Maps: Context-embedded Information Product Maps [18]. The same should be done also to technological platforms by adding more contextual metadata to current technical features. In this study, we provide an empirically inspired example why these methods should be supplemented with additional contextual human information.

## **CONSTRUCTIVE RESEARCH METHODS**

Constructive research refers to a scientific process of producing solutions to explicit problems [19]. Solutions can be theoretical and practical constructions, such as models, methods, organizations or prototypes. Constructive approach is often used in applied sciences, such as computer science, engineering or clinical medicine.

Kasanen et al. point out that anything that solves a problem cannot be called constructive research. The research process and the construction itself must be linked to theoretical background and practical relevance. In addition the research must contribute to the scientific discussion and provide evidence for being feasible and functioning in practice [19]. Design science is one of the many methodologies used widely in data and information quality research [20]. As design science can be seen as a part of the constructive approach, many of the guidelines suggested by Hevner et al. [21] were followed in this experiment to guarantee the quality of the research process, scientific constructions and derived results. In this case, the research methods combine observational, experimental and descriptive methods [21]. Informal interviews were used to observe the reality of the selected information production process. The experimental APC-simulator was used to visualize the phenomena and its characteristics. Finally, informed argumentation was used to describe the inconsistencies between reports by combining local empirical information and selected scientific findings from other places.

## **Constructive Research Process**

This study follows the constructive research process described by Kasanen et al. [19].

### **Identifying Research Questions**

The motivation of this research was to demonstrate technical mechanisms and human reasons contributing to the information accuracy problems in administrative reports. The best way to do this was to choose a familiar everyday healthcare concept and the simplest possible report calculation process which could be found to suffer from large-scale semantic heterogeneity.

The semantic concept was chosen to be ‘ambulatory procedure’. It is defined as ‘a surgery performed on a person who enters and leaves from the hospital on the same day’. This concept is very important in practice since healthcare procedures are deliberately being transformed towards ambulatory mode. It is more cost-effective and safer than alternative more intensive forms of procedures resulting to similar health outcomes. As medical sciences advance, the more operations can and will be performed in ambulatory mode.

The following research questions were derived after the concept was selected:

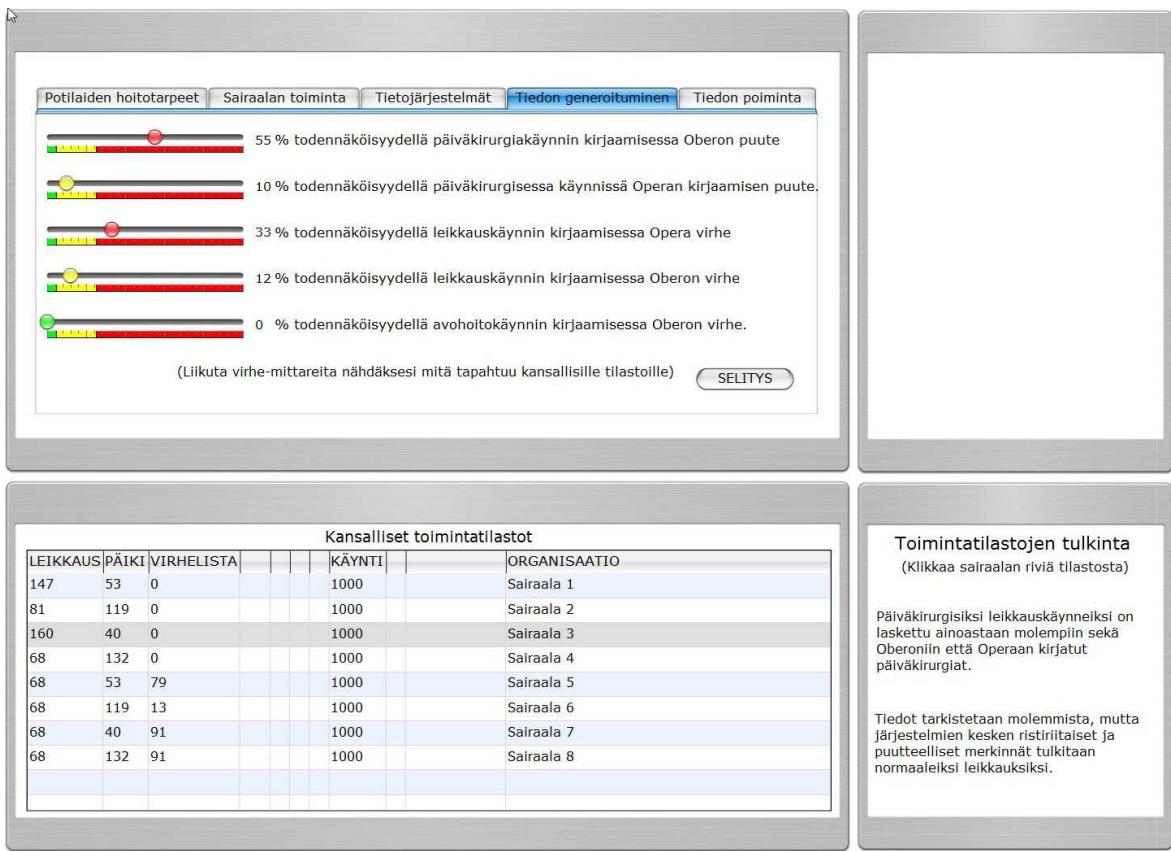
1. How to calculate the count of ‘ambulatory procedures’ based on a single Boolean attribute in a real hospital environment?
2. What kind of quantitative and qualitative inaccuracies there exists in the calculation process of ‘ambulatory procedures’?
3. How do these contextual inaccuracies affect the administrative reports including ‘ambulatory procedures’?
4. Why do these inaccuracies exist in the production process in the first place?

### **Gathering Understanding**

The research site was a university hospital providing tertiary care for almost half million people. The presented calculation rules, original report data and contextual details were encountered during a regional data warehousing project. The project’s goal was to build a centralized data warehouse to unify the data across separate information systems. It was found out that there were systematic and contextual reasons for inconsistencies between reports derived from different systems. Many error rates just could not be classified simply as random failures to enter absolutely correct data. Quite often there were intertwining technical and organizational issues leading to conflicts between contexts. The exact research case was chosen after semantic heterogeneity, existing error rates and contextual reasons were recognized to be a critical management problem. Later, additional empirical information was gathered by informally interviewing relevant practitioners with healthcare, statistical and technical education. The topics of the interviews included, for example, clinical work practices, user interfaces, software database structures, data extraction scripts, report calculation rules and report usage depending on the expertise of each practitioner. The purpose was to verify what these errors actually mean in practice and why they exist in the organizational reality in the first place. Finally, the most important details were confirmed by emails. In this way, the author’s previous work notes and the APC-simulator were iteratively and more objectively verified by others for scientific purposes.

### **Constructing Artefact**

Originally, the APC-simulator (figure 2) was implemented as an interactive dashboard on top of a spreadsheet. Report calculation rules and artificial data fields were implemented in the spreadsheet functions. Functionally, the spreadsheet simply calculated the impacts of accuracy error rates to various alternative reports. Then, the user interface layer was constructed to display alternative more user-friendly views including adjustable input fields, dynamic reports and information windows. The dashboard allowed users to adjust rates of each error type and see what happens to artificial reports in real time.



**Figure 2: The APC-simulator is an interactive dashboard. User can insert varying accuracy error rates to data. Each hospital (sairaala in Finnish) has a different calculation rule to report the count of procedures.**

### Demonstrating Results

The APC-simulator and empirical findings, described in following sections, were presented to many internal stakeholders and external contacts. The presentations were mostly informal meetings including, for example, chief medical officers, management personnel and researchers. The APC-simulator was used to illustrate visually the consequences of simple data accuracy errors to different reports currently suffering from inconsistent values. Error types and reports were then explained by using the information gained from previous informal interviews.

In this way, the reasons and mechanisms causing the huge inconsistency in different reports became more understandable. The reports were not anymore simply wrong or unreliable. They were noticed to be different for many good reasons that weren't visible for all stakeholders. The illustration also provided a way to visualize and explain the need for a better documentation about information products and information production processes. As a result, the APC-simulator provided a practical way to start discussions about data quality problems. Additionally, it was used to argue for a need to start a data quality research project to identify, fix and prevent similar currently hidden information accuracy errors in information production processes.

## DESCRIPTION OF THE APC-SIMULATOR

Following calculation rules and identified contextual explanations are all based on the reality although there does not exist exactly these hospitals or exactly these error rates. The amount of hospitals and mag-

nitudes of error rates have been chosen to illustrate the characteristics and the consequences of the studied phenomena. Also, to simplify the presented experiment some of the APC-simulator's properties will be ignored in this article. The APC-simulator contains three layers of adjustable variables: medical services, data error rates and report calculation rules.

### The amount of patients and types of services

The APC-simulation assumes four identical hospitals each providing perfect quality medical services. Also the patient population is absolutely identical. Therefore the same amount of identical patients enters each hospital and they all receive similar perfect care. The APC-simulator allows adjusting the counts of patients entering the hospitals: a) patients receiving ambulatory procedure, b) patients receiving normal procedure and c) patients not receiving any procedure (table 2). This variable selection covers all possible patient cases also in reality. To simplify the article the variables related to the type C are not used although they do exist in the implemented APC-simulator.

PERFORMED MEDICAL SERVICES		
Service Type	Correct Data Value (EPR:OR)	Patient Count (Adjustable)
Ambulatory procedures	1:1	100
Normal procedures	0:0	100
Visits without procedures	Ignored	Ignored

**Table 2: The amount of patients is set to be identical for all hospitals.  
The counts can be adjusted interactively.**

### The percentage and type of input errors

Hospitals have identical information systems: electronic patient records (EPR-system) and operation room management (OR-system). During patient visits healthcare practitioners are expected to insert a mark in each system for each patient receiving an ambulatory procedure. In theory, each patient should have value 'true = 1' or value 'false = 0' in both of the systems in all moments of time. However, in technical data there are exactly four possible combinations: '1:1', '1:0', '0:1' and '0:0'.

The APC-simulator allows altering the error percentages. There exist exactly four types of errors: missing true in EPR, missing true in OR, invalid true in EPR and invalid true in OR (table 3). To simplify the APC-simulator every hospital has identical error profile. In practice, it would be trivial to add individual error profile for each hospital. However, that would only make the APC-simulator more complicated to follow. In this case, one would choose exactly the same error rates for each hospital anyway to communicate the conclusions. The purpose of the simulator is to illustrate error rates and their impacts rather than simulate the real reporting environment in its whole complexity.

DATA INPUT ERRORS			
Error Types	Correct Value (EPR:OR)	Entered Value (EPR:OR)	Error Rates (Adjustable)
Missing true in EPR	1:1	0:1	40 %
Missing true in OR	1:1	1:0	5 %
Invalid true in EPR	0:0	1:0	5 %
Invalid true in OR	0:0	0:1	10 %

**Table 3: Error rates are adjustable and user sees in real-time how altering percentages make reported counts diverge from actually produced services.**

### The data source and calculation rules

In practice, each hospital has a choice to report ambulatory procedures by calculating their count from

one of the systems or using both of them (table 4):

DATA SOURCES AND HOSPITALS		
Data Source	Calculation Rule	Hospital
EPR	EPR	Hospital A
OR	OR	Hospital B
EPR, OR	EPR and OR	Hospital C
EPR, OR	EPR or OR	Hospital D

**Table 4:** All hospitals share identical data and error profile to highlight the importance of calculation rules. This theoretical situation occurs also in practice when different departments build own departmental reports from the same system.

## RESULTS OF THE APC-SIMULATION

While dynamically setting up the patient counts for each patient type (table 2), one can see the reports in real-time. Each hospital has identical counts of ambulatory procedures, normal procedures and regular patient encounters in dashboard all changing in perfect harmony.

However, after altering the error percentages for various error types (table 3) one can see the reports diverge rapidly although in the APC-simulator each hospital has same data and identical error profile. The reason for the emerging inconsistency is the last variable: calculation rules (table 4). Erroneous data contain contradictory values in source systems and calculation rules treat sources differently.

One possible simulation input data (table 5) results in following reports (table 6). The magnitudes of presented error rates are informed guesses chosen to demonstrate the issue rather than scientifically validated facts. They were chosen to reflect the fact that in reality the total difference between systems is near 50%. Each error rate was selected to roughly match their estimated scopes in relation to each other.

USER INPUTS					
MEDICAL SERVICES - REALITY			INPUT ERRORS – RAW DATA		
Service Type	Correct Data	Patient Count	Error Types	Raw Data	Error Rates
Ambulatory Procedure	1:1	100	Missing true in EPR	0:1	40 %
Normal Procedure	0:0	100	Missing true in OR	1:0	5 %
			Invalid true EPR	1:0	5 %
			Invalid true OR	0:1	10 %

**Table 5:** User can adjust ‘patient count’ and ‘error rates’ to simulate their impacts to the reports for each hospital.

The high amount (40%) of missing ambulatory procedures in the EPR-system results from subjective decisions of individual healthcare practitioners. Public service practitioners often do not want to send a more expensive bill to students, elderly, single mothers or even to any of the patients. They simply do not update the variable to the ‘true’-state for financial and social reasons.

The smaller amount (10%) of invalid ‘true’-values in the OR-system results from data entry policies and technical limitations of existing software systems. The operation room staff plans procedures according to their local operational and patients’ medical requirements. These values can be considered quite reliable since they are used in their internal processes and are generated by automatic timestamps. Unfortunately sometimes the reality does not follow original plans. A patient receiving an ambulatory procedure must stay in hospital overnight, a planned ambulatory procedure must be changed to another

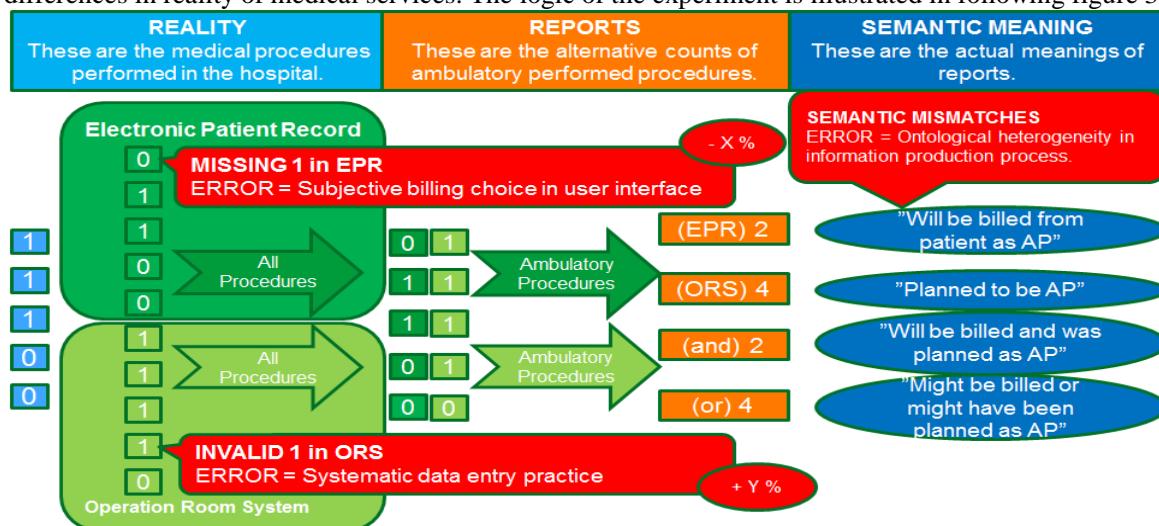
one or something else happens. After the patient leaves the immediate operation situation, these values will not be updated back to ‘false’ in OR-system. Not even if the situation changes later for a reason or another. Therefore, reports published from the OR-system contain always more ambulatory procedures than were actually performed from the perspective of the generalized definition.

The smaller inaccuracies (5%) were chosen mainly for simulation purposes. It might be that also these error types exist. For example, in the EPR-system one might make a mistake between several user interface windows and send an ambulatory bill to a wrong patient. It is more likely that mistakes like these will be corrected later after patients notice wrong bills. Therefore these two smaller error types are not so likely because of the technical characteristics of current software systems, local clinical processes and human motives. In addition, these two other error types would be more clearly random human mistakes rather than systematic semantic mismatches described earlier.

SIMULATION RESULTS				
Data Source	Calculation Rule	Hospital	Reported Normal Procedures	Reported Ambulatory Procedures
EPR	EPR	Hospital A	135	65
OR	OR	Hospital B	95	105
EPR, OR	EPR and OR	Hospital C	143	57
EPR, OR	EPR or OR	Hospital D	88	112

**Table 6: Reports are calculated from identical data including identical errors but using different calculation rules.**

The presented APC-simulation demonstrates how none of the reports match with the reality. Data in both systems includes errors and therefore it does not represent exactly what it is assumed or documented to represent. In addition, each hospital produces very different numbers from identical data by choosing different calculation logic to calculate their official count of ambulatory procedures. All differences in reported counts result from accuracy errors and internal manipulation logic rather than describe any differences in reality of medical services. The logic of the experiment is illustrated in following figure 3.



**Figure 3: The experiment simply demonstrates the effects of adjustable error rates and calculates a set of different reports. Additionally it can be used to clarify the semantic meanings of each error and report.**

The APC-simulation demonstrates how the final information product is highly dependent on usually

unrecognized accuracy error rates in raw data and internal characteristics of the calculation rules. If decision makers or researchers rely just on the reports (table 6) they might think that there are differences between hospitals. On the contrary, they should know the existing inaccuracies as well as all the calculations rules to be able to make valid conclusions from different reports.

The experiment is used to illustrate the phenomena called ontological heterogeneity [9]. The highest error rates are relative and result from semantic mismatches between slightly different concepts in information production processes. The EPR-system contains accurately information about ambulatory procedure patient bills and the OR-system stores accurately the planned work flows of operation room services. However, their semantic details in the local context do not fully match with each other nor the generalized definition used by administration and national statistics.

## INTERPRETING THE RESULTS

### ***Analyzing Impacts of Error Rates in Technical Data Flow***

Hospital D claims to have produced 112 ambulatory procedures while hospital C reports only 57 ambulatory procedures. Neither is even close to the real count (100) which they both have produced. The count of hospital D appears to be almost twice that of the hospital C although the largest error rate was set to be just 40% and other error rates much less than that. This demonstrates how accuracy errors can multiply their impacts because of internal calculation logic in information production processes.

On the other hand, inaccuracies might seem to vanish because of lucky choices in data sources. Hospital B reports 105 ambulatory procedures based solely on the OR-system data and getting quite close to the true value of 100. However, even they have a severe hidden data quality problem in the OR-system although the final overestimation is only 5%. As one can see, in reality 15 patients have wrong information in the OR-system. The true value is missing from the records of five patients while ten patients have an additional invalid true. Errors can lessen the impact of each other even in the same source system.

The actual data values make it also visible how one cannot always determine the semantic accuracy errors from input or output data. In this case it is impossible to know even whether data instance ‘0:1’ is an inaccurate value resulting from missing ‘1’ from ‘1:1’ or invalid ‘1’ in ‘0:0’. It might be that contextual information about work practices, user interface structures or application logics could give hints which one is the more common error in the local reality.

### ***Identifying Underlying Reasons for the Error Rates***

It would be easy to just argue that healthcare practitioners should enter the data accurately, completely and timely to avoid error rates in the first place. Unfortunately many inaccuracies happen for a good reason. Error rates become more understandable after contextual meanings of each data element, error rate and report are revealed. In this way, one can better understand why these hidden error rates exist in the reality and what could be done to avoid similar situations in the future.

### **Reasons for Errors in the Electronic Patient Record System**

The high amount (40%) of missing ambulatory procedures in the EPR-system was noted to result from the subjective decisions of individual doctors and nurses. It could also be seen as a semantic mismatch between different contexts. First, the hospitals official data entry guideline clearly states that the status of ambulatory procedure should be documented in the EPR-system. That is because it is technically impossible to document it in the OR-system. Secondly, the user interface in the EPR-system screen states explicitly handling only a billing issue and this text field is seen constantly by healthcare practitioners. There reads ‘bill from patient as ambulatory procedure’. Finally, IT-department uses database attribute descriptions which match the administrative documentation but not the actual usage and user interface.

All these semantic domains are used by different user groups. Administration uses guidelines and

organizational structures to control the work practices. Medical practitioners work with the actual software systems. They rarely have time or will to read and follow all the details of administrative guidelines. Finally, the IT people extract data from system tables and rely mostly on the technical documentation.

One underlying reason for the semantic mismatch is that the administration has tried to standardize data definitions and provide information for secondary uses in local administration and national statistics. Everything in the information production process was changed but the tiny text field in the user interface. The proprietary software system's user interface and patient billing logic were not altered to support this particular customer requirement.

Administrative efforts to teach users to enter data to support administrative requirements do not work in practice when they are in opposition to contextual requirements and motivations. Neither can technical database specifications nor report calculation rules change the true meaning of entered data in live patient care contexts. In practice, the simple user interface text just overrules all the other semantic meanings across the whole information production process.

### **Reasons for Errors in the Operation Room Management System**

The smaller error rate (10%) of additional ambulatory procedures in the OR-system is based on the lack of updating the variable even if the situation changes. Unfortunately, neither this error rate can be easily fixed just by changing the data entry guideline to update the variable for several reasons. The root cause is actually a complex combination of technical software restrictions, organizational practices and a problematic nature of the administrative concept itself. The OR-system does not have any good technical option for users to insert such a value just for administrative purposes. Software technology providers are not easily willing to change their user interface properties or internal application structures of their proprietary software to suit better the needs of a single customer in one country. There are also other uses for non-updated variables in internal management and development. It is useful to have a series of different counts of 'ambulatory procedures'. Each has own contextual meaning relevant for internal operation room processes. The count of planned ambulatory procedures decreases step by step for various reasons. There is a need for different reports each derived from different timestamp or other variables. Forcing a single version of the truth or updating variables could result in the loss of necessary version history or destruction of critical semantic differences leading to other reporting problems.

The concept of ambulatory procedure itself is problematic in actual medical work contexts. In reality the patient might stay overnight in another ward or clinic rather than in the immediate one performing the procedure. Practitioners in the operation room might not know what happens later in other places. They might not know when to update it back to 'false'-status even if they were able to do so.

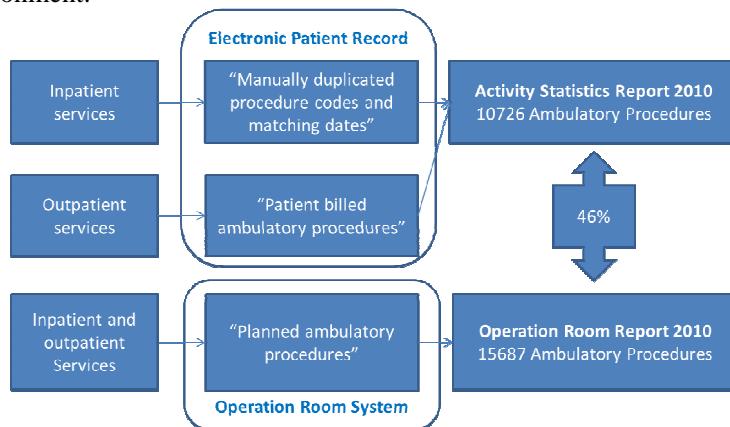
### **Potential Solutions to Fix the Errors**

The definition seems simple - a surgery performed on a person who enters and leaves from the hospital on the same day. In reality, the definition requires facts that are currently scattered and volatile for complex contextual reasons. These facts are currently handled simultaneously by two separate information systems, several organizational units and different practitioners. The current system architecture and the selected variables make it impossible to derive directly and accurately the count of performed ambulatory procedures based on current national definitions. It would be possible to get a better estimation of the correct count by selecting alternative variables while integrating data to a data warehouse. This was being done by the time of the data warehousing project and it has been done in some other similar hospitals. Extracted patient visit timestamps from the EPR-system and procedure codes from the OR-system could be matched in an enterprise data warehouse to get a closer estimation of the correct count. However, even then there would exist many alternative counts of ambulatory procedures each having a slightly different semantic meaning. One would still have to choose between alternative calculation rules to interpret the situations when patients end up receiving several procedures,

inpatient episodes and outpatient encounters during the same day. Each new variable and calculation rule would also introduce a new source of potential inaccuracies. Most importantly, even identical calculation rules and variables can produce heterogeneous information products if there exists differences in other parts of the information production processes. For example, contextual details in work practices, division of labour, user interfaces, human motivations, domain terminology etc. can introduce systematic accuracy error rates such as described in this case. One should know these contextual details to be able to detect potential inaccuracies and semantic mismatches that are embedded in the information products constructed for varying purposes.

## LIMITATIONS OF THE STUDY

The previous technical calculation process and empirical explanations are only a partial description of a more complex reality (figure 4). The case was chosen to illustrate the most important general characteristics of the studied phenomena rather than providing a full picture of all details found in the real reporting environment.



**Figure 4: The experiment is a simplified snapshot of a real reporting case. It describes the process of counting ambulatory procedures during outpatient services.**

The experiment demonstrates the information production process from technical and contextual perspective in the outpatient services. In addition, patients can be admitted into hospital wards. The inpatient service case was ignored to limit the scope of the experiment. In reality, it is very similar except ‘subjective billing choice’ would be changed with ‘manual duplication of procedure code from the OR-system to the EPR-system’. Missing codes remain a similar problem since practitioners do not always duplicate the codes accurately to additional free-text fields in multiple systems just for administrative purposes.

In addition, variables describing the status of the ambulatory procedure could be looked up from many different system tables in each system. It is not always possible to know how proprietary software links its internal variables and system tables. Some of the variables might be triggered or virtual duplicates between system tables while others might be individual fields in each system table. In any case, there are actually much more potential reports than those four presented in the simulation. Each would describe a slightly different semantic variation of ‘ambulatory procedures’.

Currently, the exact error rates and all their error types in their full detail are not known either. It is well known that the counts in reports differ significantly according to the report sources and this has caused distrust to the reports. The official activity statistics (generated from the EPR-system) in the year 2010 depicts 10726 ambulatory procedures. The operation room management report (generated from departmental software) identifies 15687 ambulatory surgeries in the same year.

The purpose of the APC-simulator and this article is to illustrate the interdependencies between usually

hidden accuracy errors and their origin such as semantic heterogeneity. In reality, even technically sound data often contain similar huge but hidden accuracy error rates. Their scope and origin can be found out only by using multiple complementary methods across information production processes. In the future, more comprehensive case studies should be used to trace semantic origins and error rates across information production processes. These studies should systematically combine empirical field studies, data analytics and modeling methods. By combining methods from multiple scientific disciplines, one could trace more complex semantic concepts in a wider network of information production.

## CONCLUSIONS

The empirical findings presented here suggest that there exist significant inaccuracies in healthcare data and information products. There is also a lot of scientific evidence that healthcare data and information products include regularly inaccuracies [3, 5, 6, 8]. Inaccuracies in data endanger the validity of management decisions, policy recommendations and statistical research results unless they are recognized, fixed and prevented.

The experiment demonstrated how inaccuracies in data can diminish or multiply because of the internal characteristics of information production processes. To determine the accuracy of information products, one must know these internal characteristics. Therefore, the entire information production process should be made completely transparent to avoid unpredictable behavior of hidden error rates.

The empirical explanations highlighted how significant error rates can be actually semantic mismatches between contexts. These mismatches can occur in any part of the information production process because of subtle contextual details. A type of user interface input field or a systematic data entry practice. The analyses of technical data flow made it also visible how the semantic accuracy of a semantic concept cannot be determined by studying only data. Technically identical data can still be semantically heterogeneous. To recognize semantic accuracy errors, data flows should be supplemented with contextual information about human factors and semantic details in the actual information production process.

Ontological heterogeneity like presented in this experiment could be fixed by redefining the concepts to their local meanings [9]. In practice, the semantic contexts (e.g., data entry guidelines, user interface, application logic) should all match. Then data could be redefined to match accurately this local definition rather than trying to change it to match ambiguous definitions of secondary uses. Unfortunately this is not always possible since secondary uses might expect information that cannot be produced directly in organizational reality. In these cases, the subtle semantic differences between primary meanings and secondary expectations should be made visible. Also, the magnitude of accuracy error rate between contextual meanings should be monitored to guarantee the validity of decisions.

On the whole, the experiment highlights a need for further research. Information accuracy errors and semantic lineage should be made visible for all stakeholders across information production processes. Additional contextual information about human factors and technical details in information production processes and information products could be used to determine their fitness for different purposes. Only in this way, one can guarantee the validity of decisions in management, development, research and regulation of healthcare services.

## REFERENCES

- [1] Olson, J.E. *Data Quality: The Accuracy Dimension*. Morgan Kaufman, 2003.
- [2] Redman, T.C. "Measuring Data Accuracy: A Framework and Review" in *Information Quality*, eds. R.Y. Wang, E.M. Pierce & S.E. Madnick, M.E. Sharpe, Inc., Armonk, NY, USA, 2005. pp. 21-36.
- [3] van Walraven, C., Bennett, C. and Forster, A. J. "Administrative database research infrequently used validated diagnostic or procedural codes." *Journal of clinical epidemiology*, 64 (10), 2011. pp. 1054-1059.

- [4] Veaux, R. D. D. and Hand, D. J. "How to Lie with Bad Data." *Statistical Science*, 20 (3). 2005. pp. 231-238.
- [5] Lofthus, C. M., Cappelen, I., Osnes, E. K., Falch, J. A., Kristiansen, I. S., Medhus, A. W., Nordsletten, L. and Meyer, H. E. "Local and national electronic databases in Norway demonstrate a varying degree of validity." *Journal of clinical epidemiology*, 58 (3). 2005. pp. 280-285.
- [6] Malin, J. L., Kahn, K. L., Adams, J., Kwan, L., Laouri, M. and Ganz, P. A. "Validity of Cancer Registry Data for Measuring the Quality of Breast Cancer Care". *Journal of the National Cancer Institute*, 94 (11). 2002. pp. 835-844.
- [7] Powell, A. E., Davies, H. T. O. and Thomson, R. G. "Using routine comparative data to assess the quality of health care: understanding and avoiding common pitfalls." *Quality and Safety in Health Care*, 12 (2). 2003. pp. 122-128.
- [8] van Walraven, C. and Austin, P. "Administrative database research has unique characteristics that can risk biased results." *Journal of clinical epidemiology*, 65 (2). 2011. pp. 126-131.
- [9] Madnick, S. and Zhu, H. "Improving data quality through effective use of data semantics." *Data&Knowledge Engineering*, 59 (2). 2006. pp. 460-475.
- [10] Ballou, D. P. and Pazer, H. L. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems." *Management Science*, 31 (2), 1985. pp. 150-162.
- [11] Ballou, D., Wang, R., Pazer, H. and Kumar, Tayi, G. "Modeling information manufacturing systems to determine information product quality." *Management Science*, 44 (4), 1998, pp. 462-484.
- [12] Wang, R. Y., Lee, Y. W., Pipino, L. L. and Strong, D. M. "Manage Your Information as a Product." *Sloan Management Review*, 39 (4). 1998. pp. 95-105.
- [13] Huang, K., Lee, Y. and Wang, R. *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River, New Jersey, 1999
- [14] Lee, Y. W., Leo L. Pipino, Funk, J. E. and Wang, R. Y. *Journey to data quality*. MIT Press, Cambridge, Mass., 2006.
- [15] Shankaranarayanan, G., Wang, R.Y., Ziad, M. "IP-MAP: Representing the Manufacture of an Information Product." In *Proceedings of the 2000 Conference on Information Quality*, 2000. pp. 1-16.
- [16] Cui, Y. *Lineage Tracing in Data Warehouses*. Ph.D. thesis, Stanford University, 2001.
- [17] Hagerty, J., Sallam, R. L. and Richardson, J. "Magic quadrant for business intelligence platforms," Gartner, Inc. 2012.
- [18] Lee, Y., Chase, S., Fisher, J., Leinung, A., McDowell, D., Paradiso, M., Simons, J. and Yarsawich, C. "CEIP Maps: Context-embedded Information Product Maps" In *Proceedings of the AMCIS 2007*, pp. 12-31.
- [19] Kasanen, E., Lukka, K. and Siionen, A. "The Constructive Approach in Management Accounting Research." *Journal of Management Accounting Research*, 5. 1993, pp. 243-264.
- [20] Madnick, S. E., Wang, R. Y., Lee, Y. W. and Zhu, H. "Overview and Framework for Data and Information Quality Research." *Journal of Data and Information Quality*, 1(1). 2009. pp. 1-22.
- [21] Hevner, A. R., March, S. T., Park, J. and Ram, S. "Design Science in Information Systems Research." *MIS Quarterly*, 28 (1). 2004, p p. 75-105.

# ASSESSING ACCURACY DEGRADATION OVER TIME WITH A MARKOV-CHAIN MODEL

(Completed Academic Paper)

**Alisa Wechsler**

Ben-Gurion University of the Negev, Israel

[alisav@bgu.ac.il](mailto:alisav@bgu.ac.il)

**Adir Even**

Ben-Gurion University of the Negev, Israel

[adireven@bgu.ac.il](mailto:adireven@bgu.ac.il)

**Abstract:** Accuracy, among the most discussed data quality dimensions in literature, reflects the extent to which data values match a baseline perceived to be correct – e.g., the true real-world attribute values, or another validated dataset. Even when data values are accurate when acquired, their accuracy may degrade over time - certain properties of real-world entities may change, while the data values that reflect them are not being updated. Drawing on that assumption, this study suggests a Markov-Chain model that describes accuracy degradation over time – this by assessing the likelihood of a data attribute to transition from one state to another within a given time period. Evaluation of the model with real-world data shows its potential contribution for a few key data-quality management tasks, such as the prediction of accuracy degradation, and the development of data auditing and maintenance policies.

**Key Words:** Data Quality, Accuracy, Currency, Markov-Chain Model

## INTRODUCTION AND BACKGROUND

Accuracy, the extent of data correctness, is among the most discussed data quality (DQ) dimensions. A data item is considered to be inaccurate if its value doesn't match the correct real-world value, or another baseline value that was validated to be correct (Even and Shankaranarayanan, 2007). Errors in data acquisition (e.g., flawed data-entry) and processing (e.g., calculation errors) (Ballou et al., 1998) are common causes for inaccuracies. However, this study argues that even if data values are being recorded and processed correctly - inaccuracies might still occur. Real-world entity may change over time (e.g., a person may change address, marital status, occupation, and other attributes), and if the data is not kept up-to-date – it becomes inaccurate, as it no longer reflects the correct real-world value. This cause for inaccuracies can be linked to another commonly-discussed DQ dimension – currency (or recency), which reflects failures to keep data items up-to-date (Even and Shankaranarayanan, 2007; Heinrich et al., 2009; Heinrich and Klier, 2011).

We suggests that in certain DQ management scenarios the dimensions of accuracy and currency are closely interlinked – as data becomes less current, it is also likely to become less accurate. The model developed in this study links the likelihood of a certain data item to become inaccurate to the time passed from the last update of that data item. Compared to other commonly-discussed DQ dimensions such as completeness, validity, and currency – accuracy is much more challenging to manage and improve. This argument is backed by the comparison in Table 1, which summarizes typical solutions for handling the defects associated with the four DQ dimensions (obviously, these are examples only – DQ literature has discussed many others possible solutions). The comparison highlights a few key differences between accuracy vs. the other dimensions:

- a) Completeness, validity and currency can be detected and corrected independently, based on the data itself, while detecting accuracy requires a certain external baseline for comparison.
- b) With completeness, validity and currency, the rules for detection are clearly defined and easier to

implement (e.g., by running a well-defined SQL query), while the comparison to a baseline might be more challenging (e.g., when no key attributes are available for comparison).

c) Completeness, validity, and currency degradation can be prevented (or, at least aided to an extent) by technical solutions, which are available today in many software packages - e.g., tools for programming data-entry screens, database management systems (DBMS), and data processing utilities in business-intelligence (BI) systems (a.k.a. ETL – Extraction, Transformation, Loading). On the other hand – preventing, or even alerting, on data accuracy defects requires some human “wisdom”, beyond a purely technical solution – e.g., defining a set of business rules that would “flag out” data values that appear to be erroneous.

<b>Dimension</b>	<b>Completeness</b>	<b>Validity</b>	<b>Currency</b>	<b>Accuracy</b>
<b>Defects reflected</b>	Missing data values	Mismatch between data values and the attribute's domain	Outdated data values	Incorrect data values
<b>Detection (In Existing Datasets)</b>	Querying the dataset for records with undesired NULL values	Querying the dataset for records with attribute values that contradict the desired domain	Querying the dataset's for timestamps that indicate too-large time margin since last update	Comparing records and data values against a baseline perceived to be correct (e.g., the real-world entity, a validated dataset)
<b>Correction: (Automatically, Using Software Tools)</b>	Imputation – filling missing values, based on similarity to other records	Setting a default value, that matches the domain	Triggering data collection or update requests	Updating values, based on the validated baseline
<b>Prevention: (Front-End)</b>	Defining attributes as mandatory in data-entry tools	Using visual aids (e.g., “radio buttons”, “drop-down lists”) that enforce value domains	Alerts on a record, or values within a record, that are not up-to-date	Alerts on value that appear to be incorrect, based on some business rules
<b>Prevention: (Back-End)</b>	Defining an attribute as “NOT NULL”	Adding “CHECK” constraints, that prevent storing values that conflict with the attribute's value domain	Expedite data processes	Setting business rules (e.g., in a form of database “triggers”) that will alert on “suspicious” values

**Table 1. Typical Solutions for Handling Defects – a Comparison of DQ Dimensions**

Acknowledging the relative difficulty – the DQ literature has addresses the issue of accuracy management and improvement in a plethora of studies. As discussed above, the task of identifying and correcting accuracy defects is inherently challenging and expensive (Olson, 2003; Even and Shankaranarayanan, 2007). Accuracy improvement often requires a baseline, against which data items can be compared - either the real-world entity itself (e.g., surveying a person and validating his/her personal details), or another data source perceived to be accurate (e.g., validating customers' addresses against a list provided by the post services).

However, in many real-world DQ management scenarios, such solutions cannot be applied - a reliable and validated data sources is not always available, or very costly if available, and auditing a large dataset against real-world entities might turn out to be too expensive. Other solutions – e.g., improving the design of data-entry screens, training end-users, and redesigning data-acquisition processes (Olsen, 2003) – can possibly reduce the chances of error, but not eliminate inaccuracies entirely. A few studies (e.g., Ballou and Pazer, 1995; Even et al., 2010; Askira-Gelman, 2010) have pointed out the tradeoffs between

the desired goal of raising accuracy to the highest possible level versus the high costs involved, and proposed approaches for assessing these tradeoffs and setting the optimal accuracy-level target. However, even if some inaccuracies are acceptable – it is still critical to assess, monitor, and improve the level of accuracy, otherwise quality might degrade to a point that data resource might become unfit to use.

DQ literature has suggested a few different methods for assessing and estimating accuracy levels. A commonly-used metric defines accuracy level as the ratio between the number of incorrect data items and the total number of data items (Pipino et al., 2002). This ratio definition can be extended by adding utility weights that reflect relative importance from the end-users' standpoint (Even and Shankaranarayanan; 2007). Sessions and Valtorta (2009) use a Bayesian Network algorithm for assessing data accuracy based on the links between data values. Fisher et al. (2009) develop an accuracy metric that considers error distributions toward detecting systematic errors. Askira-Gelman (2011) analyzes the link between accuracy assessments at the input level (raw data) versus the output level (the decision made), showing that the association is not necessarily positive.

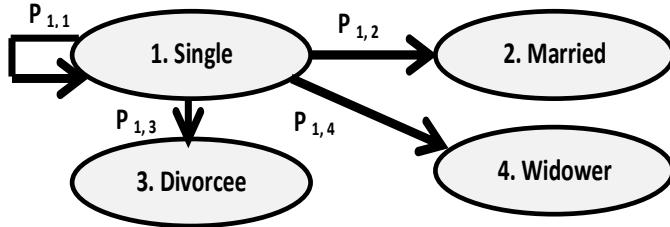
While these assessment methods reflect different views of accuracy and the mechanisms behind it – they have a few issues in common. First, with all these methods, accuracy cannot be assessed independently, based on the data itself. The assessment requires a certain baseline for comparison and/or some manual intervention and, as discussed earlier – in many real-world scenarios such solutions are not viable, or might turn out to be too expensive. Second, the assessment reflects the current state of accuracy, without showing changes in behavior over time. Data resources are dynamic in their nature – data can be added and/or updated and, as a result, the accuracy state may change. We suggest that quantifying and analyzing accuracy progressions and trends over time are critical for cost-effective accuracy management. This leads to the third issue – accuracy-assessment models that do not take into account the behavior over time, cannot provide predictive capabilities.

These issues motivated our development of a Markov-Chain (MC) model that describes accuracy degradation over time. MC models are a common approach for describing stochastic processes (Ross, 1996). They have been used in a plethora of scientific and applicative contexts – e.g., Customer Relations Management (Pfeifer, 2000), Queuing theory (Heifergott et al., 2010), and Computerized Simulations (El-Haddad, 2010); however, they have rarely been applied in the context of data quality management. We suggest that the developed model can contribute significantly to some important data quality management tasks – estimating accuracy when a baseline for comparison is unavailable or limited, predicting accuracy degradation of newly acquired data, and prioritizing accuracy auditing and improvement efforts. We next describe the model development and discuss its potential contribution. We follow that with evaluation of the model with real-world data. Finally, we offer some concluding remarks and discuss limitations and directions for future research.

#### **ACCURACY DEGRADATION OVER TIME: A MARKOV-CHAIN MODEL**

In our study, we adapt the Markov-Chain (MC) model of stochastic processes to reflect changes in the attribute values of a real-world entity. Our model is applied for a tabular dataset with N records (indexed by [n]), each reflecting an instance of a certain real-world entity, and M columns that reflect entity attributes (indexed by [m]). A data item in attribute [m] of record [n] is said to be accurate if its value  $X^{n,m}$  reflects correctly the real-world value, and inaccurate otherwise. Our model addresses scenarios in which data values are accurate (i.e., reflects correctly the real-world attribute) at the time of acquisition ( $t=0$ ), and remains unchanged else if updated later on purpose. If certain properties of real-world instances change over time, and the associated data items are not being updated accordingly – with some likelihood, those data items will no longer reflect correctly the current real-world values, and results in some accuracy degradation.

The Markov-Chain model is based on the likelihood of a certain object to transition from one state to another within a given time period. We first address a single data item, which describes a certain property of a real-world entity. A data item may transition, within a given time period, from one state (a certain data value) to another. For example (Figure 1), the “Marital Status” of a certain person may transition between four states – “Single”, “Married”, “Divorcee”, or “Widower”. If a certain person is single (state 1) at the beginning of a certain time period ( $t=0$ ), with some probability ( $P_{1,1}$ ), s/he will stay single by the end of that period ( $t=1$ ). However, with some probability s/he may become married, divorcee, or widower ( $P_{1,2}$ ,  $P_{1,3}$ , and  $P_{1,4}$ , respectively). Similarly, we can define transition probabilities between all other states.



**Figure 1.** Transition Probabilities for the “Single” state

The model is developed for each data attribute independently, and targets attributes with discrete value domains – i.e., a finite set of  $J$  possible values, indexed by  $[j]$  (e.g., Marital Status, Occupation, or Region of residence). Time is modeled as a discrete variable ( $t = 0, 1, 2, \dots$ ), where the values reflects equal time intervals (e.g., day, month, year). Notably, extensions of the MC model have also addressed multi-dimensional attribute vectors, as well as continuous value domains and time variables (e.g., El-Haddad et al., 2010; Heifergott et al., 2010) – and our model can be extended accordingly in the future. The fixed-size time intervals reflect, in our model, periodical data auditing. By the end of each time interval, we decide whether or not to audit and correct certain data values. Within a given time interval, the data value in attribute  $[m]$  may transition from state  $[i]$  to state  $[j]$  (or remain at state  $[i]$ ) with a probability of  $P^m_{i,j}$ , such that  $\sum_{j=1..J} P^m_{i,j} = 1$  for each  $[i]$ . The transition probabilities for attribute  $[m]$  can be represented in a form of a matrix  $P^m$ , and the model assumes that this matrix is identical for all the records in the dataset, and doesn't change over time.

$$P^m = \begin{pmatrix} P^m_{11} & \cdots & \cdots & P^m_{1J} \\ \vdots & P^m_{jj} & \ddots & \vdots \\ (1); & \ddots & \ddots & \vdots \\ P^m_{J1} & \cdots & \cdots & P^m_{JJ} \end{pmatrix}$$

The matrices  $\{P^m\}_{m=1..M}$  may help assessing attribute volatility. An attribute  $[m]$  is said to be stable if all diagonal-cell values  $\{P^m_{jj}\}_{j=1..M}$  are nearly 1, while others are nearly 0. At the extreme case, the attribute is said to be stagnant – once its value is set, it stays permanent and its accuracy will not degrade over time. An attribute  $[m]$  is considered volatile when some diagonal-cell values  $\{P^m_{jj}\}_{j=1..M}$  are much smaller than 1, while non-diagonal cells are substantially greater than 0.

The MC model assumes that the transition matrix  $P^m$  is known, or can be reasonably estimated from data samples. As  $P^m$  is assumed not to change over time,  $P^m(t)$ , the  $t$ -steps transition matrix of attribute  $[m]$  (i.e., the set of probabilities that a certain value in attribute  $[m]$  will change from state  $[i]$  to state  $[j]$  after  $t$  periods) is the  $t$ -power of the transition matrix:  $P^m(t)=(P^m)^t$ . Further, the MC model assumes “memory-

less” transitions – meaning that the probability of having a certain value  $X^{n,m}_{t+1}$  in attribute [m] of record [n] at the end of period t+1, depends only on the transition matrix  $P^m$ , and on the value ( $X^{n,m}_t$ ) at the end of period t, and not on earlier values:

$$P^n\{X^{n,m}_{t+1} = j | X^{n,m}_t = i_t, \dots, X^{n,m}_0 = i_0\} = P^m_{ij}(X^{n,m}_{t+1} = j | X^{n,m}_t = i_t) \quad (2)$$

With no updates,  $X^{n,m}_t$  is accurate at time t if the real-world value has not changed, or if changed and transitioned back to the original value. We define  $A^{n,m}_j(t)$  as the expected accuracy of data item [n,m] at time t, given a current value of j. It equals to the likelihood that a real-world value of j at the time of acquisition (t=0), is still j at time t:

$$A^{n,m}_j(t) = P^m_{jj}(t) \quad (3)$$

Be averaging, we can assess the expected accuracy level of a record, an attribute, or the entire dataset ( $A^{R(n)}(t)$ ,  $A^{C(m)}(t)$  and  $A(t)$ , respectively), given the set of known data values at time t:

$$A^{R(n)}(t) = \frac{1}{M} \sum_{m=1}^M A^{n,m}_j(t), \quad A^{C(m)}(t) = \frac{1}{N} \sum_{n=1}^N A^{n,m}_j(t), \quad A(t) = \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N A^{n,m}_j(t) \quad (4)$$

It can be shown that the expected accuracy is a number between 0 and 1, where the averaging method adheres to the DQ metrics guidelines in (Even and Shankaranarayanan, 2007). With different relative importance of records and attributes, these definitions can be extended to use a weighted-average formulation proposed in that work. With a MC model, it can be shown that for a data-value of j in attribute [m] at a certain time, the expected time for transitioning out to a different value, is:

$$T^m_j = \frac{1}{1 - P^m_{jj}} \quad (5)$$

Given that average-time estimation, one could set a policy for next-time audit, given a certain current value. It can be shown that the time for transitioning out of state j, can be estimated with an exponentially distribution, where  $\lambda^{n,m}_j$  is the rate of data item [n, m], currently at state j, to leave that state and  $\alpha$  is an auxiliary parameter.

$$F_j(t) = 1 - \exp\left\{-\alpha\lambda^{n,m}_j t\right\} \quad (6)$$

Until time t, if the real-world value had not transitioned, the associated data item is accurate. Therefore, using the approximation, the expected dataset accuracy can be expressed as:

$$A(t) = \frac{1}{NM} \sum_{m=1}^M \sum_{n=1}^N \exp\left\{-\alpha\lambda^{n,m}_j t\right\} \quad (7)$$

Using the exponential approximation for evaluation is substantially less time consuming than using the entire model, as it doesn't require matrix multiplication; hence such an approximation has importance in cases where the dataset has a large number of records, or where the prediction involves a large number of time periods.

The proposed model can support a few important DQ management tasks:

- a) **Estimating accuracy level:** measuring the accuracy of a large dataset is challenging and expensive, as it requires a baseline for comparison (e.g., the real-world values, or another dataset that was validated for correctness). The suggested model permits accuracy estimation of a dataset, or subsets within, given the current values and the time since their last update - without requiring assessment against a baseline.
- b) **Predicting future accuracy degradation:** when new items are recorded, the model can help predicting their accuracy behavior over time. Knowing the recorded value  $j$ , the model can help predict the accuracy level at time  $t$ . It can also predict the time until the accuracy will decline below a certain desired threshold value – and recommend auditing at that time.
- c) **Prioritizing data maintenance efforts:** the model may help assessing the accuracy behavior of data subsets (records and/or attributes), and setting auditing and maintenance priorities accordingly. As discussed earlier, the set of transition matrices can help differentiating between stable versus volatile attributes. Further, given the current data values, records with a higher likelihood of inaccuracy can be detected and audited.

A key challenge with the proposed framework is estimating the transition matrices  $\{P^m\}$ . Such estimation requires a large-enough sample of data records, which includes the history of data-value transitions. However, given such a sample – the matrix component  $P_{i,j}^m$  will be estimated by the number of times that attribute [m] transitioned from value [i] to value [j], divided by the total number of attribute [m] transitions from of value [i] (including “transitions” from [i] back to [i]). The next section demonstrates a case in which the availability of such a data sample permitted estimation of transition matrices and reasonable prediction of accuracy behavior over time.

## MODEL EVALUATION

The evaluation described in this section used a dataset published by the Central Bureau of Statistics. The dataset contains 25 economic-performance indicators on 124 industrial sectors, which were collected annually over a 14-year period (a total of 3224 records). Data as such may have major importance, in a variety of decision-making scenarios. For example, government agencies may use it to guide allocation of financial aid to certain industries, promoting certain industries oversees, or setting differentiating taxation policies. Such data can also be used by the private sector - e.g., for guiding investment decisions, or setting loan and interest-rate policies. In accordance with our modeling assumptions – the data items were updated in fixed time intervals and, as the data source is considered highly reliable, it was reasonable to assume that the numbers provided are accurate. However, as the status and the financial performance of certain industries may change over time – if decision makers do not have the most up-to-date data available, and use older data instead, their decisions are likely to be biased.

Year	Industry 1			Industry 2		
	Sales Rank	Revenue Rank	Export Rank	Sales Rank	Revenue Rank	Export Rank
<b>0</b>	1	1	2	8	5	3
<b>1</b>	1	1	2	9	5	3
<b>2</b>	1	1	2	9	5	3
<b>3</b>	1	1	2	8	5	2
<b>4</b>	2	1	2	8	4	2
<b>5</b>	3	2	2	8	4	2
<b>6</b>	3	2	2	6	4	1
<b>7</b>	3	2	2	7	4	2
<b>8</b>	3	2	2	5	3	2
<b>9</b>	3	2	2	6	3	1
<b>10</b>	3	3	2	5	2	1
<b>11</b>	4	3	2	4	2	1
<b>12</b>	4	3	2	5	2	1
<b>13</b>	4	3	2	5	2	2

**Table 2. Annual Rankings of Two Sample Industries**

For the purpose of this study, we evaluated three key financial attributes, among the 25 available – Sales ( $m=1$ ), Revenue ( $m=2$ ), and Export ( $m=3$ ). These characteristics have continuous value domain, and had to be “discretized”. This was done by classifying each industry/year record 10 equally-size deciles (1 being the highest), based on the value range of each indicator (e.g., the “Pharmaceutical” industry is ranked in the 2<sup>nd</sup> decile in terms of Sales, 4<sup>th</sup> in Revenue, and 1<sup>st</sup> in Export), this division relies on the assumption of uniform value distribution. Over time, some industries improved their ranking, while others declined; hence, data records that were published a few years back do not reflect the accurate ranking. The data behavior over time is demonstrated by the two industry examples in Table 2. The ranks of industry 1 appear to be more stable than the ranks of industry 2 - but while industry 1 demonstrates degradation, in terms of relative positioning, industry 2 demonstrates some increase. The examples also highlights a difference in the behavior of the different indicators – The Sales rank appears to be more volatile than the two others, the Revenue rank is a bit less volatile, and the Export rank seems to be relatively stable (for Industry 1, it is identical for the entire 14-year period).

While the examples above show a difference in the stability of the three attributes – their correlations (Table 3), based on the raw numbers, are high, positive, and significant. The correlations between the ranks (after “binning” the raw numbers) are lower, but still relatively high, positive and significant (the numbers shown are for the last year, however, all the other years show similar highly-positive correlations, with similar levels significance).

Metric	Raw Numbers		Ranks	
	Revenues	Export	Revenues	Export
<b>Sales</b>	0.992**	0.956**	0.893**	0.398**
<b>Revenues</b>	-	0.985**	-	0.695**

**Table 3. Pearson Correlations**  
(\*\* - significance level of 0.01 or less)

The high correlation between the attributes may explain the similarity in the transition matrices estimated for the three. This similarity is further highlighted in Table 4, which assesses the expected time for a

certain value to transition to another (in this case – the expected time it would take for a certain industry to change its relative ranking in terms of Sales, Revenues, or Export).

Rank (j)	Sales			Revenues			Export		
	P <sup>1</sup> <sub>ii(1)</sub>	Trans. Time (T <sup>1</sup> <sub>i</sub> )	Rec. Audit Periods	P <sup>2</sup> <sub>ii(1)</sub>	Trans. Time (T <sup>2</sup> <sub>i</sub> )	Rec. Audit Periods	P <sup>3</sup> <sub>ii(1)</sub>	Trans. Time (T <sup>3</sup> <sub>i</sub> )	Rec. Audit Periods
<b>1</b>	0.624	2.657	3	0.632	2.717	3	0.603	2.522	3
<b>2</b>	0.350	1.538	2	0.356	1.553	2	0.351	1.541	2
<b>3</b>	0.338	1.510	2	0.335	1.503	2	0.342	1.519	2
<b>4</b>	0.240	1.315	1	0.239	1.315	1	0.241	1.318	1
<b>5</b>	0.307	1.443	1	0.306	1.442	1	0.301	1.431	1
<b>6</b>	0.230	1.299	1	0.233	1.303	1	0.227	1.294	1
<b>7</b>	0.185	1.228	1	0.188	1.232	1	0.184	1.226	1
<b>8</b>	0.231	1.301	1	0.237	1.311	1	0.226	1.291	1
<b>9</b>	0.261	1.352	1	0.270	1.370	1	0.258	1.348	1
<b>10</b>	0.488	1.952	2	0.484	1.938	2	0.476	1.910	2

**Table 4. Assessment of Time to Transition, and Recommended Audit Periods**

The expected time ( $T^m_j$  – in Equation 5) for each possible state (10 ranks, in our case), is calculated based on the probability to stay at the same state (the main “diagonal” in the transition matrix  $P^m$ ). The results show that for all three financial indicators - the ranking stability is higher at the “edges”. When a certain industry is ranked “high” (decile 1) or “low” (decile 10) – it is likely to stay at that decile for a longer time, compared to industries that are ranked in the mid-range ranks. Based on the assessments of expected transition time – audit recommendations can made. After how many periods would it be recommended to audit and evaluate the data? If the transition time is relatively short – the data item should be audited within a relative short period after the last update. If the transition time is expected to be higher – it would be reasonable to postpone the auditing for that data item.

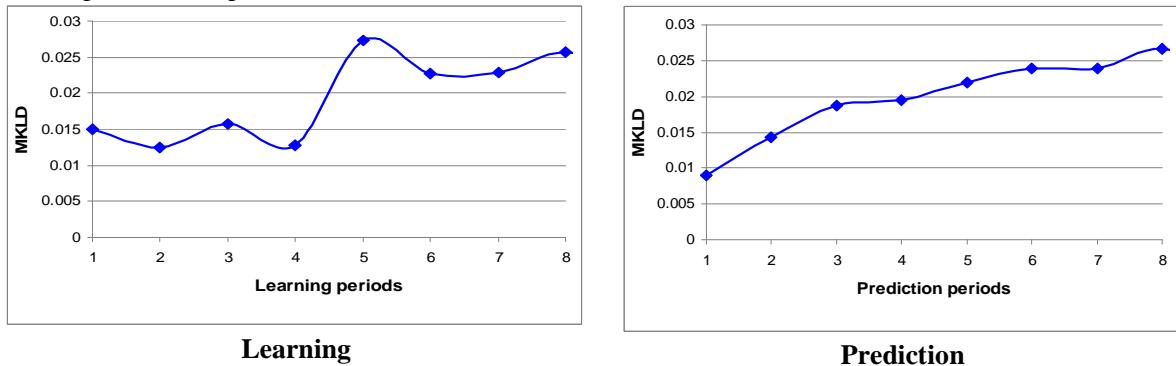
The extent of accuracy degradation over a time period of Z years can be therefore assessed by comparing the rankings in year Y records versus the rankings of year Y+Z records. We consider a record to be inaccurate if at least one of the three rankings has changes over that period. Our analysis reflects two manifestations of the time variable: **a) Learning:** the number of periods between the first update and the last update, and **b) Prediction:** the number of periods between the year of last update and the year of accuracy assessment. To estimate the transition matrices, we randomly chose 80% of the records as a training set, and used the rest 20% as a test set. We repeated this process 10 times with different random permutations and averaged the results. To assess performance, we used the Kullback-Leibler Distance (KLD) metric (Do, 2003). Here, we use it to measure the distance between the predicted accuracy level  $A^{PR}(t)$  versus the actual  $A^{AC}(t)$ , where the lower is the KLD, the better is the prediction:

$$KLD(t) = A^{AC}(t) \cdot \log \frac{A^{AC}(t)}{A^{PR}(t)} + (1 - A^{AC}(t)) \cdot \log \frac{(1 - A^{AC}(t))}{A^{PR}(t)} \quad (7)$$

Where,

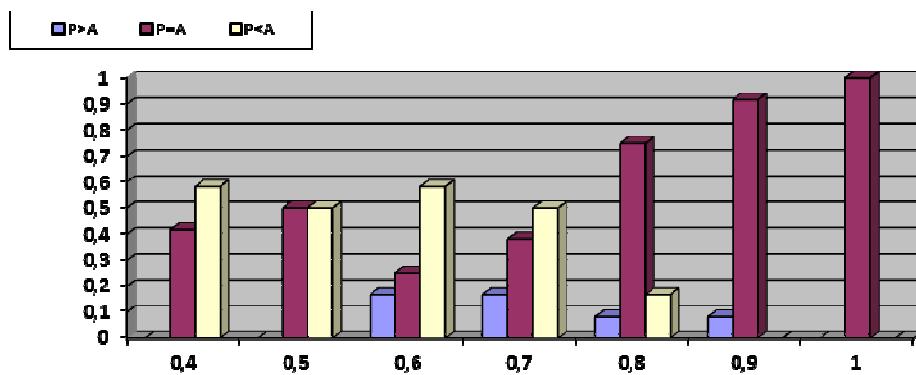
- KLD(t): The Kullback-Leibler Distance at time t
- $A^{AC}(t)$ : The true dataset accuracy at time t, where last update is at t=0

-  $A^{PR}(t)$ : The predicted dataset accuracy (the model's output) at time  $t$   
 Following these analysis principles, we used the dataset to assess the potential contribution of the model for the data quality management tasks discussed earlier. To estimate current accuracy levels, the Mean KLD (MKLD) of the ten predictions was calculated versus learning (Figure 2a) and prediction (Figure 2b) times. The MKLD values are relatively small (less than 0.027), reflecting strong performance. As expected, both learning and prediction performance degraded with a larger number of periods, and the prediction performed better with short-term time-periods. Further, in this particular case - transition matrices  $\{P^m\}$  were assumed to permanent over time; however – it is reasonable to assume that over a long period of time, the transition behavior will change; hence, learning over a too-long period might hinder prediction capabilities.



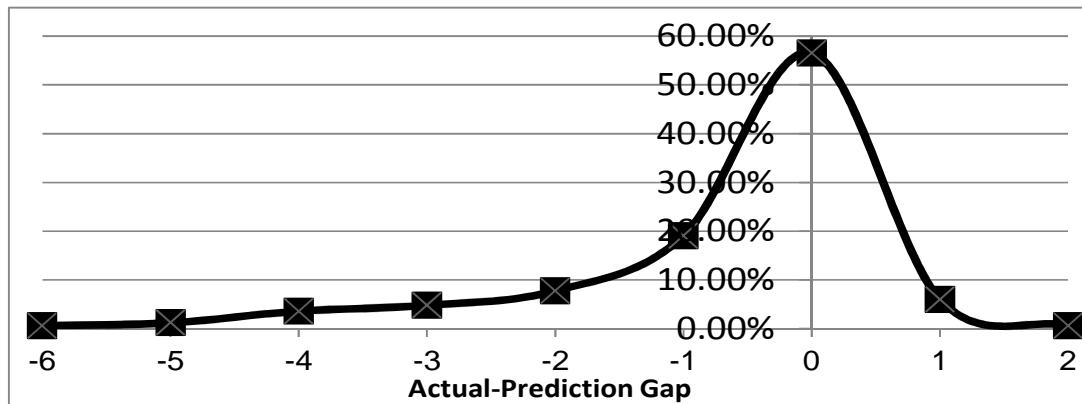
**Figure 2. Mean KLD (MKLD) over Time – (a) Learning vs. (b) Prediction**

To assess the prediction of future accuracy degradation, we asked: given a certain threshold, can we predict the number of periods that it will take the accuracy of a perfectly-correct dataset to decline below that threshold? We assessed the predicted number versus the actual for thresholds ranging between 0.4 and 1. The results (Figure 3) show that in the majority of cases the prediction was either identical to the actual or lower.



**Figure 3. Prediction (P) versus Actual (A) – Behavior for Different Thresholds**

The gap distribution (Figure 4) shows that in 56% of the cases, overall, the prediction was precise, in 37% it was lower than the actual, and only in 7% the prediction was higher than that actual. These results have important data quality management implications, as the model is shown to be stringent – in the majority of cases (93%) we would have audited data items on time or earlier.



**Figure 4. Prediction vs. Actual – Gap Distribution**

Finally, we examined the reliability of the exponential approximation (Equation 7), for prediction of accuracy degradation. For each record, we calculated the average parameters that minimize the error between the model prediction and the exponential approximation for different prediction periods between 1 and 12. As can be seen (Figure 4), the exponential approximation of accuracy level is only slightly below the actual model's prediction.

To summarize, in the evaluation above we showed the potential contribution of the MC model that we introduced in this study to a few important DQ management tasks:

**Estimating accuracy level:** the model permitted estimation of accuracy levels with a relatively high precision. As expected, the precision is higher with a larger number of learning periods (a larger “training set”), and with a shorter time-gap from the time of data acquisition.

**Predicting future accuracy degradation:** the model predicted with a relatively high precision the time it will take for the accuracy level to go below a certain threshold level. The prediction was shown to be stringent – meaning that the predicted time was either identical or lower than the actual. Only in ~7% of the cases, the prediction would have recommended too-late data auditing.

**Prioritizing data maintenance efforts:** the metrics that can be developed from the MC model can help setting and prioritizing data management effort. For example, using the set of “Time to transition” metrics ( $T^m_j$  – in Equation 5), we can set in advance a recommended time for auditing a record and/or, based on the data-value that was recorded – instead of setting a “one size fits all” policy to all records and attributes.

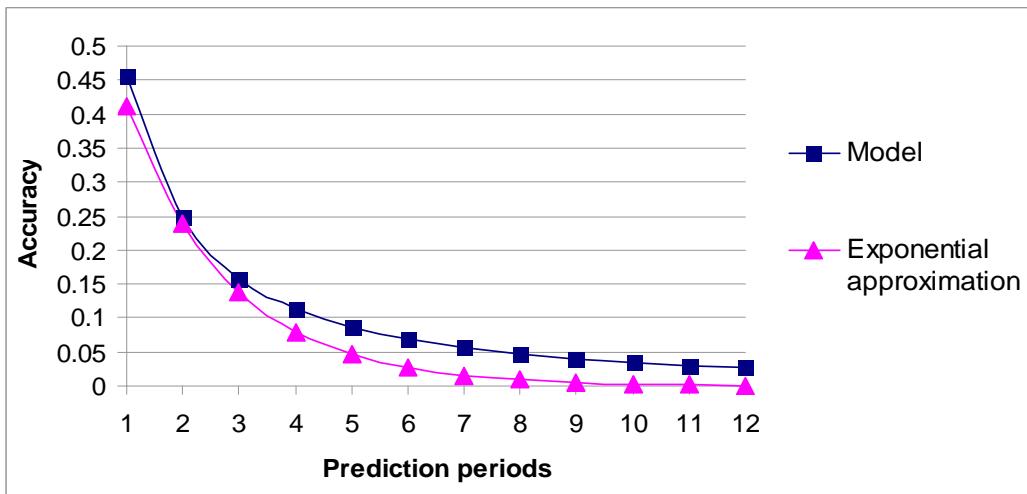


Figure 4. Exponential model and the Markov model predictions

## CONCLUSIONS

Accuracy defects are difficult and expensive to handle – hence, the importance of developing analytical tools and methodologies for understanding the mechanisms behind accuracy degradation and predicting future accuracy defects accordingly. This study contributes to that end by developing of a Markov-Chain model for simulating and predicting accuracy-degradation behavior over time. As demonstrated in our evaluation with real-world data, the model has the potential to aid a few important data quality management tasks – estimating future accuracy levels, predicting accuracy degradation, and prioritizing accuracy maintenance efforts accordingly. Obviously, the results presented here are only preliminary, and some additional evaluation is required in other data management contexts and with other real-world datasets. Notably, Markov-Chain models have rarely been applied in the context of data quality management, and we see it as a contribution of its own. This direction can be explored further, and we see it as promising avenue for developing methodologies and tools for aiding data quality management efforts.

Some of the modeling assumptions made may hold only to a limited extent in real-world cases and required further enhancement and evaluation. Data audits and corrections are not always done at fixed-length time intervals; what required a different modeling of the time variable. Data attributes (e.g., salary, length, and duration) are often continuous, and in some real-world scenarios, “discretization” into a set of bins, as done in this study, is not a valid solution. Further, the transition matrix  $P^m$  for a certain attribute may not be identical for all records (e.g., when it depends on the value of other attributes), may change over time, and may not adhere to the “memory-less transitions” assumption. A plethora of studies have extended the Markov-Chain model to address such limitations, and the solutions offered can be adopted for extending the framework developed here.

## REFERENCES

- [1] Askira-Gelman, I. Setting Priorities for Data Accuracy Improvements in Satisficing Decision-Making Scenarios: A Guiding Theory, *Decision Support Systems*, 48(4), 2010, pp. 507-520
- [2] Askira-Gelman, I. GIGO or not GIGO: The Accuracy of Multi-Criteria Satisficing Decisions, *The ACM Journal of Data and Information Quality*, 3(2), Article 9, 2011, pp. 1-27
- [3] Baldi, P., S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview, *Bioinformatics*, 16(5), 2000, pp. 412-424

- [4] Ballou, D. P., and Pazer, H. L. Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff. *Information Systems Research*, 6(1), 1995, pp. 51-72
- [5] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.
- [6] Do, M. N. Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models, *IEEE Signal Processing Letters*, 10(4), 2003, pp. 115-118
- [7] El-Haddad, R., Lecot, C., L'Ecuyer, P., and Nassif, N. Quasi-Monte Carlo Methods for Markov-Chains with Continuous Multi-Dimensional State Space. *Mathematics and Computers in Simulation*, 81(3), 2010., pp. 560-587
- [8] Even, A., and Shankaranarayanan, G. Utility-Driven Assessment of Data Quality, *The DATA BASE for Advances in Information Systems*, 38(2), 2007, pp. 76-93
- [9] Even, A., Shankaranarayanan, G., and Berger, P.D. Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting, *Decision Support Systems*, 50(1), 2010, pp. 152-163
- [10] Fisher, C.W., and Matheus, C. C. An Accuracy Metric: Percentages, Randomness and Probabilities, *The ACM Journal of Data and Information Quality*, 1(3), Article 16, 2009, pp. 1-28
- [11] Heifergott, B., Hordijk, A., and Leder, N., Series Expansions for Continuous-Time Markov Processes, *Operations Research*, 58(3), 2010, pp., 756-767
- [12] Heinrich, B., Kaiser, M. and Klier, M. A Procedure To Develop Metrics For Currency And Its Application In CRM, *The ACM Journal of Data and Information Quality*, 1(1), 2009 pp. 5-28
- [13] Heinrich, B., and Klier, M. Assessing Data Currency – A Probabilistic Approach, *Journal of Information Sciences*, 37(1), 2011 pp. 86-100
- [14] Jiang, Z., Sarkar, S., De, P., and Dey, D. A Framework for Reconciling Attribute Values from Multiple Data Sources, *Management Science* 53(10), 2007, pp., 1946–1963.
- [15] Olson, J. E. *Data Quality: The Accuracy Dimension*, Morgan Kaufmann Pub., 2003
- [16] Pfeifer, P. E., and Carraway, R.L., Modeling Customer Relationships as Markov Chains, *Journal of Interactive Marketing*, 14(2), 2000, pp. 43-55
- [17] Pipino, L.L., Lee, Y.W., and Wang, R.Y. Data Quality Assessment, *Comm. of the ACM*, 45(4), 2002, pp. 211-218.
- [18] Ross, S. M. *Stochastic Processes*, Wiley, 1996
- [19] Sessions, V., and Valtorta, M. Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm, *Journal of Data and Information Quality*, 1(3), Article 14, 2009, pp. 1-34

## **DETERMINANTS OF ACCURACY IN THE CONTEXT OF CLINICAL STUDY DATA**

(Research-in-Progress, IQ Concepts, Metrics, Measures, and Models)

### **Meredith Nahm**

Duke University, Center for Health Informatics, Durham, NC  
[meredith.nahm@duke.edu](mailto:meredith.nahm@duke.edu)

### **Joseph Bonner**

Michigan State University, Biomedical Research Informatics Core, East Lansing, MI  
[Joseph.Bonner@hc.msu.edu](mailto:Joseph.Bonner@hc.msu.edu)

### **Philip L. Reed**

Michigan State University, Clinical & Translational Sciences Institute, East Lansing, MI  
[phil.reed@hc.msu.edu](mailto:phil.reed@hc.msu.edu)

### **Kit Howard**

Kestrel Consultants, Ann Arbor, MI  
[kit@kestrelconsultants.com](mailto:kit@kestrelconsultants.com)

**Abstract:** The link between data quality and research is inextricable; after all, scientific conclusions are based on data. However, key determinants of information quality in research have not been articulated. Likewise, there are no formal constructs relating aspects of research design to data quality. In the absence of such theories, investigators and research teams formulate independent mental models and rely on personal experience to design data collection and processing operations for their studies.

We applied an iterative consensus process among four experts each with experience over the spectrum of prospective and retrospective research in both industry and government funded settings to identify key determinants of the accuracy of research data. From this work, we posit that the relative timing of three key data-related milestones 1) occurrence of the event of interest, 2) data collection about the event, and 3) data cleaning, impact information quality and research results, and therefore should be included in a broad spectrum of research design decisions that impact results. We offer a link between aspects of data collection and processing and data quality and apply the resulting framework to a case study to illustrate its use.

**Key words:** Information quality, information quality assurance, data accuracy, clinical research, theory

## **INTRODUCTION**

Methodology for collecting and managing clinical study data evolved through separate communities of practice such as industry clinical trials intended for marketing authorization, observational clinical registries of many varieties, academically oriented, government-funded clinical trials, and secondary analysis studies.[1, 2] Each developed community-specific methods that do not directly translate to other areas of practice.[2] Historically, when there was less variety in the data collected, each community optimized practices for the type of data collected and managed.

Ten years ago at the first author's institution, any given multicenter clinical trial had two or three data sources, usually including data collected on data collection forms, data from central laboratories, and data from external reading centers. Today, however, the number of data sources for any one study[3] as well as the overall complexity of clinical trials[4] is increasing. Today, the data sources on a trial are likely to include data captured directly from patients, (e.g., patient reported outcomes) as well as data collected directly from devices, retrospective data from healthcare settings, and a host of other sources. Further, today many trials are using warehoused clinical care data and real-time lab and other messages in the healthcare setting to screen and identify patients eligible for clinical trials, and many registries receive part or all of their data directly from patients or from healthcare information systems. Studies also use data collected via different methods (e.g., web-based data entry with on-screen checks, single entered data without on-screen checks, paper forms, etc.)

Cost pressure on development of new therapeutics,[5, 6] government requirements for results reporting[7] and data sharing[8] and federal incentives for meaningful use of healthcare data[9] are all increasing reuse of both healthcare and research data for secondary analyses. Evidence is pointing toward the need to integrate clinical research into the health care process.[10] Similar to increased interest in information quality following large scale data warehousing in other industries,[11] attention in the healthcare sector is turning towards secondary use of healthcare data. Two recent Institute of Medicine (IOM) meetings, Digital Data Priorities for Continuous Learning in Health and Health Care[12] and Sharing Clinical Research Data: A Workshop,[13] shared a focus on information quality.

The increase in secondary use of health care data coupled with the increasing number of data sources for any given study necessitates development and testing of relevant theory to guide selection of capable data sources, collection and management processes and selection of data sets adequate to support intended analyses.[14] Such theory should be based on underlying characteristics of data and information rather than particulars of any one use context, and should explain interactions of events and other factors impacting data quality in a manner that guides practice. Such theories, once tested and evaluated should be universally applicable to extant and planned data and should inform selection of data sources or selection of methods for data collection and cleaning such that the resulting quality will support the intended use of the data.

## **BACKGROUND**

Research studies are often categorized as prospective or retrospective. Prospective is essentially looking forward from an event of interest. Retrospective is looking backwards in time from some event of interest. For example, a prospective study is one where the unit of study, e.g., a patient, has a condition or receives a treatment and is followed forward over time from cause to effect and then compared with another group of people who are not affected by the condition or did not receive the treatment.[15] The parallel and less common terms prolective and retrolective refer to the timing of data recording. Prolective refers to data recorded after initiation of the study. Retrolective refers to data recorded before initiation of the study. Here study initiation is usually taken to mean sample selection; the impact of timing of data recording relative to sample selection can be a source of bias,[16] e.g., selecting a sample after data are collected leaves the potential for knowledge of the data to bias the sample selection. Because of their emphasis on data recording it is unfortunate that the terms prolective and retrolective are not in broader use.

Categorization of research as retrospective, prospective, prolective or retrolective reflect the importance of the relative timing of study events, including experimental control, to the strength of conclusions that can be drawn. However, these concepts focus on how the timing impacts control and representativeness

of the sample rather than data accuracy. Our work here significantly expands the concept of timing of data recording and goes further to apply it directly to impact on data accuracy.

In practice, categorization of the research design has come to be associated with corresponding intuited impression of data accuracy. For example, data analyzed for retrospective, observational clinical studies are in general presumed “dirtier” than data analyzed for prospective controlled studies. Finer distinctions have not been formalized and the determinants of data accuracy or even “data cleanliness” in a research context are not clearly articulated. Such determinants could be of utility in prospectively assessing data sources, and can also serve as a framework to guide practitioners in the common task of designing or matching appropriate data collection and management methods to a given research scenario.

Data quality has associated costs including both 1) cost to achieve a desired level of quality, and 2) cost incurred for failing to achieve a necessary level of quality. Cost of Quality[17] ideas originated and flourished in manufacturing through the work of thought leaders such as W. Edwards Deming and Joseph M. Juran, and have since been applied to other areas, e.g., accounting[18] and software development,[19] where it has been shown that correction costs increase exponentially the further downstream an error is detected and corrected.[20, 21] Walker provides an example of these costs with address data in the context of a company that ships products using consumer supplied mailing addresses.[21] Briefly, the 1-10-100 rule conveys that there is an order of magnitude increase in cost as one goes from the cost to prevent an error, to the cost of finding & fixing that same data error after occurrence, to the cost of a failure due to the error. Upstream prevention and early detection is a cornerstone of the International Organization for Standardization (ISO) 9000 series of standards. ISO standards establish international standards for quality management and quality assurance applicable to most industries. The proverb: “An ounce of prevention is worth a pound of cure” is not foreign to medical research. In fact, from organizational data, we know that an “on-screen” error check costs a few dollars to implement and address at the research site during data entry, and costs an estimated \$35 if the data discrepancy is identified after data have been submitted to the data center,[22] and further, costs much more if caught during analysis or after submission for regulatory review. The 1999 Institute of Medicine Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making workshop report articulates industry fears that one questionable data value can cast doubt on an entire regulatory submission.[23] However, in the context of significant cost pressure on therapeutic development and in light of more recent risk-based approaches for regulated clinical trials,[24] the likelihood of a serious data error and the potential impact should be weighed against the cost of preventing or fixing such data errors.

It is for the purpose of identifying and quantifying candidate data error prevention and mitigation processes that we pursue the framework presented here. The framework uses *relative timing* of key data milestones, e.g., occurrence of an event of interest, collection of data about the event, and data cleaning activities as major determinants of accuracy of the resulting data.

## METHODS

The authors all have extensive experience in research data management or informatics, and the experience covers the full spectrum intended in the NIH definition of Clinical Research.[25] As do many people, each of the authors uses a working mental model to recommend data collection and management strategies to research projects, and in the case of multiple available sources of data, to make judgments about which data sources should be used. The purpose and approach of this work is to probe the working mental models of experts and to make explicit the criteria that these experts use in decision making about design and operationalization of research data collection and management. To do this, the first author drafted a skeleton diagram and explanatory text and iteratively circulated it among the remainder of the authors for input and further discussion. The resulting consensus framework is presented here to prompt

broader discussion and evaluation.

We refer the reader to the diaphoric definition of data (DDD) and general definition of information (GDI)[26] to explore the philosophical underpinnings of definitions of data and information. Unfortunately, in practice the terms data and information are often used interchangeably, with the term “data” used when the speaker refers to a value(s) that have associated meaning. Data and information quality are multidimensional concepts.[27] Our work presented here focuses keenly on the accuracy dimension. Accuracy is in most cases an intrinsic dimension in that it is independent of any external context or use. Reflecting the DDD and GDI in this work would prompt consistent use of the word information, i.e., data plus meaning, throughout the work, and while we have not done this in deference to common use of the terms, we wish to make explicit our acceptance of DDD and GDI.

Due to the theoretical nature of this work, we are compelled to make explicit our conceptual and operational definitions of data accuracy. We conceptually define data accuracy as the property exhibited by a datum (a value) when it reflects the true state of the world at the stated, or implied, point of assessment. It follows that an inaccurate, or errant datum, therefore does not reflect the true state of the world at the stated or implied point of assessment. Data errors are instances of inaccuracy. Data errors are detected as discrepancies upon some comparison. The comparison might be between the data value and a “source of truth”, a known standard, a set of valid values, a redundant measurement, independently collected data for same concept, an upstream data source, some validated indicator of possible errors, or aggregate statistics. We use the term ‘error’ explicitly in the context of any deviation from accuracy no matter what the cause. For example, a problem in programming that renders an originally accurate value incorrect, e.g., a programming problem in a data transformation, is considered to have caused a data error. Because data are subject to multiple processing steps, some count the number of errors (consider a data value that has sustained two problems that each would have individually caused an error). From an outcomes perspective, it is the number of fields in error that matters rather than the number of errors, thus, we count the number of data values in error.

Operationally, an instance of inaccuracy or data error, is any discrepancy identified through such a comparison that cannot be explained by documentation.[28] The caveat, “not explained by documentation” is operationally necessary in practice because efforts to identify data discrepancies, i.e., potential errors, are undertaken on data at different stages of processing. Such processing sometimes includes transformations on the data that may purposefully change the value. In these cases, a data consumer should expect the changes to be documented and traceable through the data processing steps, i.e., supported by some documentation. The process of identifying and resolving data discrepancies is often referred to as data cleaning.

There are of course many factors that may impact data accuracy. The goal of our work is to develop a framework that can be applied at the research design stage to assess the suitability of existing data or to select capable processes for prospective data collection. Thus, we choose following as the backbone and foundation of our framework:

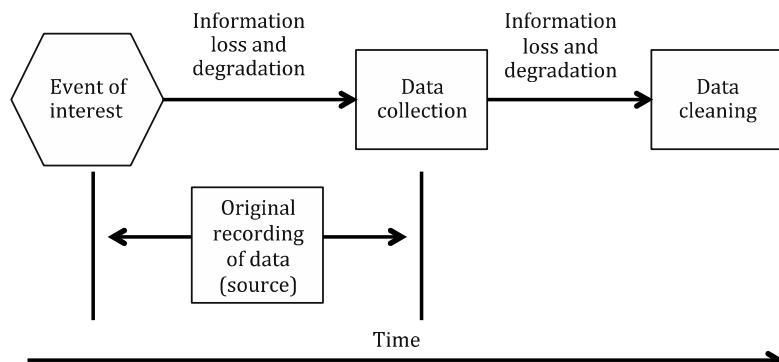
1. factors that we judge most impactful on data accuracy
2. factors that occur in the collection of most research data
3. factors that can be controlled during study design, and
4. factors that are easily discerned from extant data.

## **FRAMEWORK**

We present three crucial data-related milestones: 1) the event under study, 2) data collection about the event, and 3) data cleaning. We posit that the relative timing of these milestones is a key determinant of

achievable data accuracy. We further posit that the impact of the relative timing of these key data-related milestones can be affected by existence, completeness, and use of a recording of the event. (Figure 1)

The relative timing of the data collection about the event and data cleaning with respect to the occurrence of the event itself impacts the achievable data accuracy. Where data are recorded closer in time to the event, they are likely more accurate. If we consider delay in data collection to mean that data are jotted down on some temporary medium such as a scrap of paper or worse held in memory until they can be recorded, then it is intuitive that delay in data collection (recording for the purposes of the research) increases likelihood of inaccuracy. For example, end of assessment and end of shift charting are historically commonplace in healthcare. Similarly, where data are cleaned closer in time to the occurrence of the event and their collection, the data are likely to be more accurate. Consider for example, on-screen checks that notify a registration clerk of a missing or inconsistent zip code during patient registration as opposed to identifying and attempting to resolve the discrepancy after a patient is admitted or worse, after the patient has left the facility. This impact of the temporal distance on accuracy can be mitigated where an event-contemporaneous recording exists and is used to clean data discrepancies.



**Figure 1: Key Data-related Milestones Impacting Data Accuracy.**

### ***The source***

Figure 1 displays the original recording of the data on a separate line specifically to denote that the existence of a source may be, and often is, separate from the collection of data for a particular research need. The International Conference on Harmonization (ICH) E6 guideline definition 1.52 defines a source as the original recording of the data or certified copy thereof.[29] In the case of electronic recording of data, e.g., a blood pressure monitor, the original recording or source may itself be used as the data collection, i.e., an electronic certified by virtue of validation, copy of the source. This is not always the case. The mitigating impact of a contemporaneous and complete source is important in our framework.

Data collection about an event can occur in different ways. For example, a patient encounter can be audio or video recorded, in which case the audio or video recording would be considered the source. Consider a surgical procedure. We may be interested in the procedure type, the date and time, and the presence of any complications. If a camera in the operating room produces a recording of the event, then a contemporaneous record of the event exists. Completeness of such a recording with respect to certain parameters of interest may be achieved with an accurate system date and time stamp. However, such a recording would not be complete with respect to complications not discernable on the recording, or those

complications occurring after the patient leaves the recording area. Likewise, in the case of pregnancy and birth related events, a birth record (certificate) may exist however, this record most likely will not reflect with fidelity the circumstances or exposures.

A complete and contemporaneously recorded source doesn't exist in many cases, e.g., an event may be held in memory – to be documented later at the end of the shift, and the source may or may not be a complete representation of the event or may be otherwise be fallible. Consider for example diet recall used for epidemiology studies. It is simply impossible to remember accurately the frequency of items consumed. Additionally, because details of the event itself are not preserved, in many clinical studies, the source (the original recording), contemporaneous or not, is used as the place from which data about the event are collected, e.g., abstracted from a medical record.

Availability, fidelity, contemporaneity and completeness of an original recording of an event impacts the ability to detect and ultimately resolve discrepancies. A contemporaneously recorded source, in the absence of fidelity problems and subject to the completeness of the representation, captures as closely as possible the details of an event for later reference. With such a source, data discrepancies detected after the fact can be resolved. Figure 1 depicts that a source may be contemporaneous or not with respect to the event of interest. Closing the time gap between the occurrence of the event and the recording of such a source decreases information loss and degradation due to the passage of time in the source, e.g., memory loss.

There are two ways that information accuracy can be impacted by the source 1) the source can be used through comparison to identify discrepancies, and 2) where other standards for comparison are used to identify discrepancies, the source can be used in discrepancy resolution to confirm a discrepant value or to provide the accurate value. Often the source is not in a format amenable to comparison for discrepancy identification and is used in the manner of the latter.

Completeness of representation of the source impacts its use in discrepancy identification and resolution. For example, for a researcher studying association between physical proximity of providers and patients during encounters on patient satisfaction, an audio recording would be an inadequate representation, and a video may only give approximate proximity. However, in the case of symptoms stated by the patient, a good audio recording of the complete encounter would be a complete representation, and could be used to identify or resolve discrepancies in charted data. Incidentally, asking the patient to review the charted symptoms during the encounter would also be a possible way to contemporaneously identify and resolve data discrepancies. From the example, we see how existence, completeness of representation, and fidelity of the source determines the available methods for preventing, controlling, or correcting data errors. Thus, the existence, contemporaneity fidelity and completeness of a source are important determinants in data accuracy.

### ***Three data-related milestones***

In addition to an original recording of data about an event, the three data-related milestones: 1) occurrence of event of interest, 2) data collection about the event, and 3) data cleaning (Figure 1) are present in virtually all research and impact data accuracy. In any given research scenario, the arrangement in time of these events may take on one of several possible configurations. Most permutations of the three events are non-sensical, e.g., data collection prior to the occurrence of the event, or identification and resolution of discrepancies prior to data collection. The four arrangements occurring in reality (Figure 2) comprise distinct classes of approaches to data collection and cleaning. These classes are differentiated by the relative timing of data collection and data cleaning with respect to the event of interest and to each other. Usually more than one of the four classes is possible for any given

research scenario, and often, considering these different classes of approaches results in very different options for data collection and processing. We posit that these classes, differentiated based on the relative timing of key data-related milestones, determine the type of data discrepancies that can be detected and corrected. Therefore, research design decisions that layout key data-related milestones will determine in part the study's achievable data accuracy.

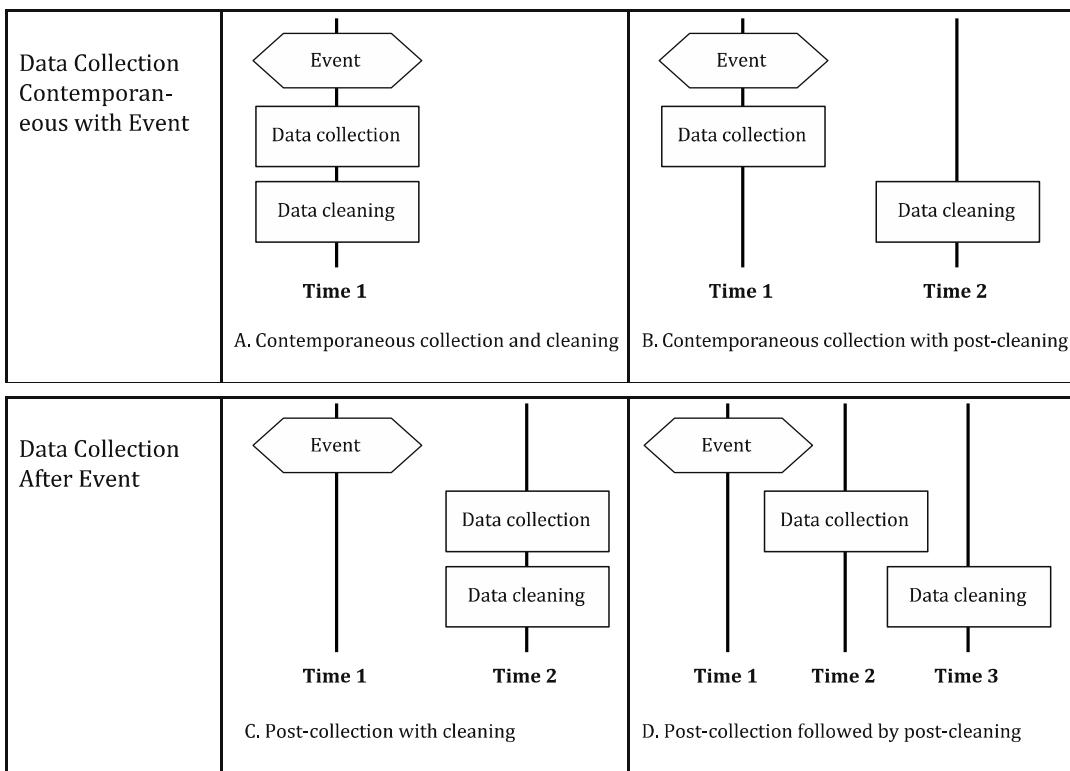
In the ideal case data collection and cleaning are contemporaneous with the event under study, i.e., the source is the data collection. In this ideal case (exhibit A in Figure 2), all three data milestones occur together in time, giving data discrepancies the maximum likelihood of being identified and resolved while "in" the event. Consider the example of a family photo; when pictures are viewed immediately on a digital camera, they can be retaken when someone's eyes are closed. Waiting some duration of time after the event to view the picture may remove the possibility of a retake. Also like the family photo example, it is sometimes possible to operationalize near contemporaneous observation, recording, and cleaning of research data. Here, we make explicit that even in the ideal case, some loss or errors occur and data cleaning may be needed. Consider the more clinically oriented example of a computer assisted telephone interview (CATI) to obtain a detailed medical history of both the interviewee, e.g., a mother, and her child. In this case the ability to have real time discrepancy checks would prompt the interviewer to ask for clarifications of inconsistencies and provide cues to insure that all needed responses are obtained from the mother. It may be impossible or prohibitively expensive to re-contact the interviewee. Therefore the capability to provide the interviewer a reminder to reconcile discrepancies during or at the end of the interview is a clear strength. The truly ideal case, of course, is lossless and completely simultaneous occurrence of the event with data collection and cleaning.

Other temporal arrangements of data collection and cleaning are also common and include: B) data cleaning after contemporaneous event and data collection C) data collection and cleaning occurring together but after the event of interest has happened, and, D) data collection occurring later in time than the event of interest, followed still later in time by data cleaning. These arrangements are depicted in Figure 2. Recall that the existence of a source other than the collected data may be present in any of these arrangements. Existence of such a source further impacts availability and applicability of options for data cleaning. For example, if interviews are recorded, fidelity of the data written on the data collection forms can be measured. If additional accuracy is needed, a redundant and independent data collection process may be pursued. Alternatively, where the interviews were not recorded, neither measurement nor review are possible.

From first principles, the relative timing of data-related milestones impact the data collection and cleaning methods that can be employed in the following ways:

1. In the absence of an independent recording of the event of interest, collected data cannot be compared with a "source of truth" as we might want to do to identify data discrepancies. In most cases, this means that the data cannot be verified, nor can they be changed with confidence.
2. Data cleaned after collection, and in the absence of an independent recording of the event of interest, cannot be verified, nor can they be changed with confidence.

The problem in both cases is that the source of observed truth (the event itself) has passed and the collected data are all that remain. Without some independent source of the observed truth, e.g., an audiovisual or other complete record of the event, the collected data must be accepted as they are because no basis upon which to make changes (corrections) exists.



**Figure 2: Relative Timing of Key Data-related Milestones**

Any of the arrangements B, C or D are rendered closer to the ideal where the event of interest is recorded in a way that preserves needed informational aspects of interest (a real source of truth), i.e., information preservation decreases the importance of data collection and cleaning contemporaneous with the occurrence of the event of interest. In this case, for example, video recorded clinical encounters, collecting data after the fact, e.g., a clinician charting at the end of a shift, can achieve the same accuracy as contemporaneously collected data. With the necessary resources, collected data or suspected inaccuracies can be verified against the recorded truth (the true source), similar to viewing the actors rather than the shadows in Plato's cave allegory. The cases above point to actions that can be taken at study planning to prevent or mitigate predictable sources of error. For example, in the cases of tests run on biological samples, independent samples can be taken and handled separately or even sent to completely separate labs, thus, the risk of losing data due to sample loss, delays in shipping, or other damage is mitigated. Mitigating or preventing impact number 1 above is accomplished by preserving a "source of truth".

Impact 2 above stems from the time lag between data collection and data cleaning. For example, prior to internet and mobile device based data collection, some types of data were usually collected by having the patients complete forms that were subsequently sent to a data center for entry. In this scenario, discrepancies could not usually be corrected due to time lag between form completion and processing at the data center creating a lengthy recall period. Such correction was of questionable validity and entailed an arduous process of contacting the study sites and the sites in turn contacting the patients to ask them how they were really feeling three weeks prior. Due to recall issues, such cleaning was misguided, thus, patient completed forms were not usually "cleaned". With internet and mobile device based data collection researchers have the ability to include on-screen discrepancy checks where appropriate, and

other data cleaning or integrity algorithms such as logging the time of entry. This relocation of the discrepancy detection and resolution process further upstream works not just because of the timing of data cleaning with respect to data collection, but because this gives the best chance of having the “source of truth”, in this case the patient, closer to the point of correction, i.e., higher likelihood that the discrepancy can be reliably corrected. Further, the effect varies because the intervening variable, time lag between occurrence of the event and the data collection. Impact 2 above is thus mitigated by a two-pronged approach, 1) decreasing the distance (geographical, temporal, or level of expertise) between the “source of truth” and the data collection, and 2) cleaning data as close as possible to the “source of truth”.

The further removed data collection is from the occurrence of the event of interest, and the further the data cleaning is from the occurrence of the event of interest and the data collection, 1) the fewer options available for preventing, or mitigating, identifying and resolving discrepancies, and 2) the more resources will be required to achieve levels of data accuracy obtained in the ideal contemporaneous case. Figure 3 explores the impact of relative timing on error prevention and data cleaning options. Also, for each arrangement, Figure 3 identifies main attractive features. In practice, these advantages and limitations of course need to be weighed and considered with respect to the particular research scenario.

With any of the above options, discrepancies may be identified by inquiring of, or comparison to the source or recorded source where available, comparison to a known standard, consistency checks with other data values, aggregate checks and distributional comparisons, or comparison with independently collected data. However, and described well in control theory, some errors are detectable, others are not, and some errors are correctable, while others are not. Thus, some errors are neither detectable nor correctable and some errors are detectable but not correctable. The arrangement chosen for a given research scenario alters which types off errors fall into the detectable and correctable categories.

The four arrangements above are one, albeit a key and a priori factor that may determine information accuracy. Admittedly, much of the argument for contemporaneous recording and cleaning rests on the concern for ephemeral states that would not allow going back at a later time for validation. There are of course many shades of grey between static and ephemeral states that should be mentioned. For example, although one may misplace a lab result, it may or may not be the case that a redraw later is equally valid for a given data use. Another example is the measurement of a patient’s weight. Weight may be fairly stable -- except in the case of surgical amputations, delivering a baby, or during aggressive treatment of congestive heart failure. Depending on the desired data use, the weight at the prior or next clinical encounter may or may not be appropriately imputed, e.g., last observation carried forward (LOCF). These phenomena are easily factored into the framework by considering the event of interest, how quickly it changes, and whether or not a replacement observation is a sufficient “source of truth”.

Importantly, the relative timing of key events is known in advance and can be used to plan capable data collection and processing processes, or for existing data, can be used to asses the likelihood that data are capable of supporting a secondary use. Further, some error causes may be suggested by the framework. For example, recall-based errors where information collection is removed in time from occurrence of the event would be suggested in arrangement C and D but not in arrangements A and B. There are however important characteristics of data errors that are not at all or not completely predicted by the framework, such as the location in the data set, the distribution and the extent of errors.

For example, random key errors in the last digit of the blood pressure are not likely to adversely impact conclusions based on aggregate data whereas severe outliers caused by an improperly calibrated measuring device may have an impact. Likewise, the distribution of the errors within a dataset may correlate with particular arrangements A-D but is not likely determined by the relative timing of key

events.

		Data cleaning	
		Contemporaneous with collection	After collection
Data collection	Contemporaneous with event	<b>Scenario A Strengths</b> Identifying discrepancies and resolving while still with the source allows maximum chance for resolution of all detectable discrepancies, and less costly resolution  <b>Scenario A Limitations</b> Can not control the timing of data collection and cleaning for existing data	<b>Scenario B Strengths</b> Prevent loss due to time lag between collection & event,  <b>Scenario B Limitations</b> Opportunity to resolve discrepancies while still with the source is lost. If the source is recorded, data can be compared with recorded source to identify discrepancies; this is usually more costly due to going back to a recorded source.. Further, discrepancies must be filtered to separate data processing errors from others.
	After event	<b>Scenario C Strengths</b> Data processing errors can be identified and corrected real-time.  <b>Scenario C Limitations</b> Allows loss due to time lag between occurrence of the event and data collection; recall may be degraded as may the ability to obtain the necessary data from the recording of the source.  Ability to resolve discrepancies is impacted by existence/fidelity of source recording and fidelity of data processing as well as by availability of the source to, "go back to".	<b>Scenario D Strengths</b> If lower accuracy can be tolerated, may be the lowest cost option.  <b>Scenario D Limitations</b> No opportunity to correct while subject is still there/while source is fresh.  Limited or cost-increasing options for correcting data processing errors.  Additionally allows loss due to time lag between collection & cleaning and the occurrence of the event.

**Figure 3: Strengths and Weaknesses in Four Arrangements of Key Data-related Milestones**

## DISCUSSION

When one thinks of research data within the framework, the distinction between common notions of prospective as collecting data in the future, and retrospective as use of previously collected data becomes more fine-grained and with distinctions that are actionable at study design or at decision to use existing data. For example, prospective studies often employ questionnaires asking about past events, collect medical history, and details about the onset of symptoms – all past information with data collection occurring some time after the event of interest occurs, and in some cases, in ways that is difficult to tell how far out in time from occurrence of the event of interest the original data were collected. Further, retrospective studies may rely on data that were collected contemporaneously with the occurrence of the event of interest, or even recorded directly, e.g., by a device. In the latter case, we might consider these data having more potential for accuracy than the previous “prospective” example. The method of data collection and cleaning should not be assumed by the label applied to the study, e.g., prospective versus

retrospective, but instead should be determined by the relative timing of data collection and cleaning with respect to the event of interest.

Existence of a data cleaning milestone in the model does not imply that all situations should include data cleaning. Instead, we include a data cleaning milestone for consideration of its utility in a given scenario. In fact, in many industries and particularly in observational clinical studies, discrepancies are often enhanced through independent data sources or identified and taken into account in the analysis, but not corrected. The cost of data cleaning is sometimes justified by the 1-10-100 rule.[20,21] The larger question, however is, for a given data use, is the benefit of data error prevention or cleaning worth the cost? Based on variability in practice across industries, we say probably not. In clinical research, where data from some studies are used once to answer a question, and then not used again, it is not worth it to clean data beyond what is required to support the planned analysis. In the case of clinical registries, where data are collected during routine care, it may not be feasible to clean data at input. In general, when likelihood of secondary use is increased, or when the risk of a serious problem is too high, then data are cleaned. Alternatively, when there is a low likelihood of re-use beyond the planned analysis – then a risk benefit analysis will likely favor the data accuracy to support planned analysis and nothing more.

Tcheng, et al., previously describe two categories of causes of data accuracy problems in the path from the true state of the patient to the documented state of the patient; 1) representational inadequacy, and 2) information loss and degradation.[30] Representational inadequacy is concerned with the choice of data elements used to document the state, including elements of context necessary for interpretation of the data values. Representational inadequacy is a cause and special case of information loss and degradation. While representational inadequacy is important and relevant to data accuracy, the impact on accuracy is known *a priori*, e.g., issues of level of abstraction and precision of chosen data element. These concerns are at the level of the data element or even operationalization rather than at the level of the data value and should be addressed when deciding operationalizations for important concepts and which data elements are collected or documented to support the operationalizations. Here, we are concerned with the accuracy of the collected values, i.e., information loss and degradation of data values and occurring after representational decisions, i.e., choice of operationalizations and data elements, have been made. Our framework furthers the work of Tcheng et al. by delineating some key determinants of information loss and degradation. Tcheng et al. present their model in the context of data retrieval for secondary use, while here we are concerned with both prospective and extant data.

Reliability and validity, like representational adequacy, are properties of the data element rather than the data value. To further contextualize the theory presented here, we add that reliability, i.e., degree to which a measure provides consistent results on repeated measurement attempts, of a measure sets limits on the achievable accuracy. Validity, i.e., the degree to which the measurement itself corresponds to the state one is trying to assess, is also a property of the measure itself and its application. Validity, possibly more foundational than accuracy, tells us whether or not the value should be used at all, e.g., the United States Food and Drug Administration (FDA) strongly considers validation of instruments used to collect data submitted for regulatory decision making.[31]

Helms describes error production in data processing operations performed on data.[32] We characterize this under the Tcheng et al. framework as information loss and degradation in data collection and cleaning. Our scope of consideration includes the original recording of the event of interest and data collection, and thus, is considerably broader than that described by Helms. Further, such a broad scope is necessary based on recent empirical work showing association of larger error rates in the data collection processes than those resulting from downstream data processing methods [33].

Reasons why investigators and research teams select suboptimal data collection and cleaning processes are varied.[34] Briefly, these reasons often include lack of knowledge, lack of resources, and competing design priorities. Rapidly evolving data collection technologies are removing barriers that previously prompted suboptimal data collection and cleaning processes. While the framework presented here cannot affect issues of resources or competing priorities it can mitigate lack of awareness and knowledge.

## APPLYING THE FRAMEWORK TO A CASE STUDY

We have elsewhere described a scenario where 12-lead Electrocardiograms (ECGs) were collected in a clinical trial.[2] In the scenario, when data tables were being reviewed prior to delivery to the study sponsor, a quality control auditor discovered that the average values for one clinical investigational site were very different from the others, so different that the data were characteristic of small mammals rather than of humans. Upon investigation it was found that the ECG data recorded on the study data collection forms was as shown on the tables and listings, i.e., the discrepant values were not the result of data entry error. It was ultimately found that the ECG machine at the site was faulty. Unfortunately, the discovery was made over a year after the date of data recording. Prior to the data table quality control check, there was no aggregate review of the ECG data, i.e., quality considerations of these important safety data were not part of the research design.

### *Completeness and existence of a source*

Analyzing the case according to the framework, the ECG print-out is a complete (with respect to the parameters of interest) recording of the event, i.e., the electrical activity of the myocardium during the period of interest. This complete and contemporaneous recording of the event was available for verification, and ultimately how the cause of the problem was identified.

### *Possible arrangements of key data milestones*

As stated, the scenario most closely resembles arrangement D – data collection after the event with subsequent data cleaning. Here, the data collection is the recording of the ECG parameters on the study data collection form. There was no cleaning contemporaneous with data collection. Arrangement D could have been improved by earlier collection and transmission of data to the data center or by applying the data cleaning as close to data arrival as possible. Essentially closing the time gap or achieving arrangement A, B, or C would have been an improvement. In particular, comparison of aggregate statistics across sites as data were collected and transmitted to the data center would have detected the problem during the active treatment phase of the trial and subsequent errant data could have been prevented.

The ideal case would be to take the information quality aspects into account during study planning and implement arrangement A, where two or three consecutive ECG readings were taken and transmitted or otherwise subject to range and consistency checks as soon as they were recorded, preferably while the patient was still in the examining room. At the time of the case, such real-time data transmission and checking was not possible. We do not argue that the ideal case is appropriate for every case, merely that the logistics and cost required to produce the necessary level of information quality be considered during the research design.

## LIMITATIONS AND FURTHER WORK

While the reasoning here is from first principles and seems intuitively solid, we prioritized the factors considered, and thus do not address the full compliment of factors impacting accuracy of research data. Further, only “controllable or discernable” factors were considered (controllable at least in prospective

data). Thus, as discussed earlier, the framework proposed here is not complete over all causes of information errors in clinical research. For example, as described elsewhere,[33] the distinct factors impacting the accuracy of the medical record abstraction process alone numbered close to three hundred. These factors and others come into play when considering as we do here, the broader scope of data recording, collection and cleaning processes across clinical research. Although we posit one framework here, additional inquiry is of course required to determine those factors that are of greatest impact to data accuracy. The framework proposed here represents hypothesis generation through qualitative consensus - what a few experienced practitioners and researchers deem to be key and *a priori* determinants of data accuracy. Thus, the determinants are likely necessary but are certainly not sufficient for data accuracy. As such, the model is incomplete; only testing and trial will affirm or not whether these determinants in fact act as we hypothesize that they do, under what conditions, and whether these factors outweigh others.

The theory posited here, although based on combined expert experience, requires testing. Such testing might involve observation of two studies using two different arrangements (Figure 2) and subsequent comparison of the data error rates. The ideal test would be a controlled trial within one study where comparable clinical investigational sites are randomly and blindly allocated to different arrangements. Support for theory would be offered if discrepancy rates scaled as indicated in Figure 2 while accounting for possible confounding factors.

## IMPLICATIONS AND CONCLUSIONS

Data processing has not often been considered as something that should impact research design. Instead, data processing has often been an afterthought, or considered mere operational detail rather than the stuff of academic discourse. We make the argument that data collection and processing are important considerations in research operational design; after all, the scientific method relies upon conclusions drawn from data and reproducibility of research results. Further, data collection, processing and cleaning can be costly. There is a growing trend in industry clinical trials toward risk-based approaches,[35] i.e., according to the potential for impact on study and regulatory outcomes. In non-trial clinical research settings data quality is often sacrificed for budgetary constraints. However with the increasing secondary use of clinical research data, there is a natural tension between risk-based approaches for the initial data use and the potential benefits of secondary uses. Application of key determinants of data accuracy supports risk-based approaches, and can ease secondary use tension by also providing a means to assess availability and suitability of data to support a particular use.

The experts involved in this work were easily able to reach consensus on a set of factors that we deemed most impactful on data accuracy, that occur in the collection of most research data, that can be controlled during study design and that are easily discerned for extant data. Through this work, we provide 1) a framework that can be tested, and 2) an ideal case against which data collection efforts can be assessed rather than the ambiguity that exists in clinical research today. However, the consensus-based method by which we derived the theory doesn't support further conclusions. Therefore, it is important to test both the importance of the identified factors to data accuracy, and the asserted relationships between the factors.

Although historically, clinical research data have been subject to extensive discrepancy identification, verification and resolution processes, the economic pressure on the therapeutic development can no longer withstand a "clean every data value approach". The framework proposed here can be used to inform risk-based analysis and approaches by providing candidate processes for which risk, cost and benefit can be considered. High-risk high-benefit data elements or data sets can be subjected to more rigorous processes. Further, the actual cost and benefit of the process variations can be considered to identify areas where an ounce of prevention will outweigh a pound of cure.

## ACKNOWLEDGMENTS

This work was supported by the National Center for Research Resources (NCRR) and the National Center for Advancing Translational Science (NCATS) a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research through grant number UL1RR024128, and the National Library of Medicine through grant number 1K99LM011128-01A1. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. Further, we gratefully acknowledge insightful comments on early drafts of the manuscript received from Michael G. Kahn, MD, PhD, Department of Pediatrics, University of Colorado, Denver.

## REFERENCES

- [1] Nahm, M. Clinical Data Management in Perspective, Data Basics, Society for Clinical Data Management Vol 13: No. 2, 2007.
- [2] Nahm, M. Data Gone Awry, Data Basics, Society for Clinical Data Management Vol 14: No.2, 2008.
- [3] Nahm, M., Current and Future Trends in Clinical Data Management, Data Basics, Society for Clinical Data Management. Vol. 11: No 2, Summer 2005.
- [4] Getz, Kenneth A., Wenger, Julia, Campo, Rafael A., Seguine, Edward S., Kaitin, Kenneth I., Assessing the Impact of Protocol Design Changes on Clinical Trial Performance. American Journal of Therapeutics 15, 450–457 (2008)
- [5] U.S. Department of Health and Human Services, Food and Drug Administration US Food and Drug Administration. Innovation Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products. March 2004. Available from <http://www.fda.gov>.
- [6] Malakoff D. Clinical trials and tribulations. Spiraling costs threaten gridlock. Science. Oct 10 2008;322(5899):210-213.
- [7] The Food and Drug Administration Amendments Act of 2007 (US Public Law 110-85). Passed on September 27, 2007. Available from United States Government Printing Office, <http://www.gpo.gov>
- [8] National Library of Medicine, National Institutes of Health, Data Sharing Plan Requirements page. Last updated January 2, 2011, accessed October 1, 2012, available at <http://www.nlm.nih.gov/ep/datashare.html>
- [9] Centers for Medicare and Medicaid , Meaningful Use EHR Incentive Programs. Last modified August 27, 2012, accessed on October 1, 2012, Available from <http://www.cms.gov>
- [10] Califff R.M., Filerman G.L., Murray R.K., Rosenblatt M., The Clinical Trials Enterprise in the United States: A Call for Disruptive Innovation, April 13, 2012. Available from <http://www.iom.edu>
- [11] Lee YW, Pipino LL, Funk JD, Yang. WR. Journey to data quality. Cambridge, MA: MIT Press; 2006.
- [12] Grossmann, C., Powers B., Sanders J., Rapporteurs, Institute of Medicine of the National Academies , Digital Data Improvement Priorities for Continuous Learning in Health and Health Care - Workshop Summary, prepublication summary, September 28, 2012, available from <http://www.iom.edu>
- [13] Institute of Medicine of the National Academies, Sharing Clinical Research Data: A Workshop, October 4-5, 2012, Washington DC, pre-meeting summary. Available from <http://www.iom.edu>
- [14] Nahm M., Chapter 10 Data Quality in Clinical Research in Richesson R., Andrews J. Eds., Clinical Research Informatics. Springer, 2012.
- [15] Agency for Healthcare Quality and Research (AHRQ), glossary entry for Prospective Observational Study, accessed May 18, 2012. Available from <http://www.effectivehealthcare.ahrq.gov/index.cfm/glossary-of-terms>
- [16] Krauth, Joachim, Experimental Design: A Handbook and Dictionary for Medical and Behavioral Sciences.

- [17] ASQ Quality Costs Committee, Principles of Quality Costs: Principles, Implementation, and Use, Third Edition, ed. Jack Campanella, ASQ Quality Press, 1999, pages 3–5.
- [18] Spiceland J. David, Sepe James F., Nelson Mark W., Intermediate Accounting eBook th Ed., Chapter20: Accounting Changes and Error Corrections. McGrawHill 2010. Available from <http://connect.mcgraw-hill.com>.
- [19] Boehm Barry W. , Papaccio Philip N., Understanding and Controlling Software Costs. IEEE Transactions on Software Engineering, v. 14, no. 10, October 1988, pp. 1462-1477.
- [20] Ross, J.E., What is the 1-10-100 Rule? Total Quality Management, February 2009. Accessed August 28, 2012, Available from <http://totalqualitymanagement.wordpress.com/2009/02/25/what-is-1-10-100-rule/>
- [21] Walker, B., The real cost of bad data, the 1-10-100 Rule. A Melissa Data White Paper. Accessed August 28, 2012, Available from [www.melissadata.com/dqt/1-10-100-rule.pdf](http://www.melissadata.com/dqt/1-10-100-rule.pdf)
- [22] Kush, Rebecca Daniels, Ph. D. et. al. eClinical Trials: Planning and Implementation. Thompson Centerwatch, Boston, MA 2003.
- [23] Davis, Jonathan R., Nolan, Vivian P., Woodcock, Janet, Estabrook, Ronald W., Eds., Assuring Data Quality and Validity in Clinical Trials for Regulatory Decision Making Workshop Report from the Roundtable on Research and Development of Drugs, Biologics, and Medical Devices, Division of Health Sciences Policy, Institute of Medicine. National Academy Press, Washington DC. Available from [http://www.nap.edu/openbook.php?record\\_id=9623](http://www.nap.edu/openbook.php?record_id=9623)
- [24] Department of Health and Human Services, U.S. Food and Drug Administration, Guidance for Industry Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring, Draft Guidance. August 24, 2011. Available from <http://www.fda.gov>.
- [25] Dictionary entry for “Clinical Research”, National Institutes of Health Glossary and Acronym List, accessed July 10, 2012. Available from <http://grants.nih.gov/grants/glossary.htm#C>
- [26] Luciano Floridi, Semantic Concepts of Information, latest revision Jan 28, 2011 <http://plato.stanford.edu/entries/information-semantic/>
- [27] Wang R, Strong D. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12(4):30.
- [28] Nahm ML, Pieper CF, Cunningham MM (2008) Quantifying Data Quality for Clinical Trials Using Electronic Data Capture. PLoS ONE 3(8): e3049. doi:10.1371/journal.pone.0003049
- [29] United States Food and Drug Administration, Guidance for Industry E6 Good Clinical Practice: Consolidated Guidance. 1996. Available from [www.fda.gov](http://www.fda.gov).
- [30] Tcheng, James E., Fendt, Kaye, Nahm, Meredith, Data quality issues and the electronic health record. DIA Global Forum, 2011.
- [31] U.S. Department of Health and Human Services Food and Drug Administration, Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, December 2009. Available from <http://www.fda.gov>
- [32] Helms, Ronald w., Data quality issues in electronic data capture. Drug Information Journal, Vol. 35, pp. 827–837, 2001.
- [33] Nahm M., Data accuracy in medical record abstraction. Doctoral dissertation. University of Texas at Houston School of Biomedical Informatics (formerly School of Health Information Sciences). May 2010.
- [34] Nahm M, Zhang J, Operationalization of the UFURT methodology for usability analysis in the clinical research data management domain., Journal of Biomedical Informatics 2009 42(2):327-33.
- [35] U.S. Department of Health and Human Services Food and Drug Administration, DRAFT Guidance for Industry Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring, August 24, 2011. Available from <http://www.fda.gov>

## **CUSTOMISED DATA QUALITY IMPROVEMENT**

**Philip Woodall**

University of Cambridge  
phil.woodall@eng.cam.ac.uk

**Alexander Borek**

University of Cambridge  
ab865@cam.ac.uk

**Ajith Kumar Parlikad**

University of Cambridge  
ajith.parlikad@eng.cam.ac.uk

**Abstract:** Several DQ improvement techniques (ITecs) have been proposed in order to guide the process of DQ improvement. However, while these techniques aim to be general, organisations wanting to improve DQ often have specific requirements that these general ITecs do not always meet. Therefore, this paper proposes a procedure for creating a new ITec that is driven by organisational requirements and is based on the existing approaches to DQ improvement. In particular, we model the activities suggested by existing ITecs and provide insights into how the existing ITecs overlap in their construction and how the internal activities are interconnected. Using a case example, we show how this model can be used to select the most suitable improvement path for an organisation based on their requirements.

**Keywords:** Data Quality, Information Quality, Improving Information Quality, Information Quality Improvement, Data Quality Improvement, Improvement model, Improvement Techniques.

### **1. INTRODUCTION**

The quality of an organisation's data is critical to its success, and poor data quality (DQ) can spell disaster for an organisation—see for example the citation of DQ as a causing factor in the explosion of the Challenger space shuttle and the mistaken shooting down of an Iranian civilian aircraft [4]. Two key steps that organisations can take in their quest for improved DQ are DQ assessment and DQ improvement [1]. DQ assessment is used to determine the current quality level of data. DQ improvement is the process of initiating a change to the data, either through changing the process which creates/changes the data or cleaning the data itself [1], [3] in order to make data “fit for use”. As is commonly accepted in the DQ literature, data that is fit for use is of high quality [10].

DQ improvement is the focus of this paper, rather than assessment, and a number of DQ improvement techniques (ITecs) have been proposed in order to guide this process (e.g., [2][3][5][6][7][8][9]). While these techniques aim to be general, organisations wanting to improve DQ often have specific requirements so that strict adherence to these general ITecs is either not necessary or do not always yield successful results. For example, no single ITec describes how to conduct an improvement exercise that involves determining root causes, selecting relevant DQ tools, trialling simple solutions to the DQ problems and validating the improvement results. Organisations may also want the flexibility to skip parts of the improvement process that are not needed to meet their requirements. In this case, existing ITecs individually do not provide comprehensive guidance on the options and alternative paths that can be taken at particular stages in the DQ improvement process. They also do not indicate what information would be missing should certain activities be avoided and how this would affect the rest of the process.

The research question we address in this paper is: “How can an effective and robust DQ improvement technique be built that meets an organisations’ DQ requirements?” A process that addresses this will need to meet the following criteria: (1) It should be able to accommodate all the requirements posed by the organisation; (2) It should avoid unnecessary activities; and (3) it should show the alternative ways in which an organisation can perform the improvement based on any changes to requirements (4) finally it should ultimately yield successful improvement results for the organisation.

In order to address this problem, we present a model of the DQ improvement process that shows all the activities from the existing ITecs and how they overlap and interconnect. This was done through an extensive literature review to identify the existing ITecs, and extracting the activities that comprise each ITec. These activities were then analysed to determine what inputs are needed for each activity and hence, what the ordering constraints are for each activity. Activities that aimed at achieving the same goal were merged as duplicate activities. An independent review process was used to ensure that the extracted activities were correctly interpreted.

This model can be used to select the most suitable improvement path for an organisation based on their requirements. For practical implementation in an industrial context, a four-step process that uses this model is also described. Academically, the model seeks to determine the fundamental pieces of DQ improvement that are common throughout all ITecs. It is hoped that it provides a basis from which new ITecs in particular applications getContexts can show what parts are solely application/context dependent compared to the parts that are applied in all contexts.

The rest of this paper is organized as follows: Section 2 describes DQ improvement and how it is viewed for this research. Section 3 describes the process of identifying the ITecs from the literature, extracting the activities from these ITecs, extracting the links between the activities, and finally the model of DQ improvement. Section 4 describes the four steps that can be used with the model to determine a suitable DQ improvement path, and section 5 gives an example of how to use the steps with the model based on an industrial scenario. Section 6 discusses the limitations of the work and section 7 presents the conclusions of the paper.

## **2. DATA QUALITY IMPROVEMENT**

The terms data and information are used synonymously in this paper. One way of viewing a DQ methodology is to split it into two main parts: assessment and improvement. State reconstruction is sometimes referred to as the initial step; however, most of the existing techniques start with the assessment phase and incorporate the workings of state reconstruction into the assessment part [1]. As noted before, this paper focuses on the improvement part, which uses the results of a DQ assessment to provide an understanding of the current level of DQ, to inform the improvement exercise of what data needs to be improved. In the existing literature, it is common to find proposals of a DQ methodology which comprises both DQ assessment and improvement (e.g., [3], [7], [9]).

At a further level of detail we view DQ improvement as a series of steps that are executed to deliver better data quality in the organisation. In this paper these steps are referred to as activities and indicate, at a relatively high level, what needs to be completed. The activities can therefore be conducted in different ways depending on how the data professional wants to implement them.

DQ improvement can be classified in two different ways: data-driven and process-driven. Data-driven approaches “take existing data that is defective and correct the deficiencies to bring it to an acceptable level of quality” [3]. In contrast, process-driven approaches improve DQ by redesigning the processes that create or modify the data [1]. The literature review conducted for this research found that not only do

all ITecs fit into these two categories, but most of the ITecs are in fact a combination of both and they demonstrate how to manage a process or data-driven approach within their constituent activities. The model therefore includes both perspectives of improvement.

### 3. REVIEW AND EXTRACTION OF DQ IMPROVEMENT ACTIVITIES

As mentioned before, a literature review was conducted to identify existing ITecs before extracting the constituent activities. The Scopus search engine, ACM and IEEE digital libraries, Google books and proceedings of the International Conference on Information Quality were used to search for studies containing ITecs. Conferences that have particular tracks that are related to DQ, such as the European Conference on Information Systems (ECIS) and the Americas Conference on Information Systems (AMCIS) were also searched. Furthermore, as a follow-up search, the references section of each paper was also checked for additional studies containing ITecs. This secondary search produced many papers from various areas that further searches using other search engines were deemed to be unnecessary. The following inclusion/exclusion criteria were applied to guide the selection of studies (papers, books etc.) that were included in the review:

Studies were selected if:

1. the study contains a generically applicable ITec and describes what activities are involved
2. the study describes a DQ methodology and part of the methodology is an ITec

Studies were rejected if:

3. the study contains an ITec that has not been subject to a rigorous review (as required by papers in high ranking journals or peer reviewed books)
4. the ITec in the study has not been subject to an actual implementation and successful trial of the approach that resulted in some benefit with regards to DQ
5. the study does not describe an ITec and its activities in sufficient detail to enable the reviewer to clearly and easily extract and document the activities
6. the study describes only DQ assessment and not an ITec

These criteria were chosen in order to ensure that the final model of DQ improvement, which uses the activities from each ITec, will be practically useful by ensuring that each selected ITec is implementable, produces demonstrable good results and can be understood sufficiently to allow the reliable extraction of activities. The final list of selected ITecs and the studies which propose them are shown in Table 9. ITec names are followed by ‘-i’ because they are part of a full DQ methodology (that is, having both a DQ assessment and improvement part) and this suffix is used to distinguish the ITec from the full methodology.

ITec name	Name	Reference
EDQP-i	Executing Data Quality Projects	[6]
CDQM-i	Comprehensive Data Quality Methodology	[2]
COLDQ-i	Cost-effect Of Low Data Quality	[5]
DQFG-i	Data Quality Field Guide	[8]
SODQA-i	Subjective Objective Data Quality Assessment	[7]
TDQM-i	Total Data Quality Management	[9]
TQdM-i	Total Quality data Management	[3]

Table 9: Final List of Selected ITecs

### *Extraction of Activities*

For each of the ITecs in Table 9, the activities were extracted from the sources describing these ITecs and the resulting activities are shown in Table 2. An ‘extraction table’ was used for each ITec to document details of all activities in an ITec including: the activity name, description of the activity, a cross-reference to the location of the description of the activity in the study, and any other comments. Using a structured review process, the extraction tables were checked by an independent reviewer to ensure that the extracted activities were correctly interpreted from the ITecs, are at a consistent level of granularity, and are not overlapping. Furthermore, some activities are common to more than one ITec and in this case, the relevant activities were merged into a single activity; the process of merging the relevant activities was also validated by two independent reviewers. Table 2 shows all the improvement related activities and includes an index number and abbreviation (for convenient reference to the activity in this paper), the activity name, description, and the ITec(s) containing the activity.

TQdM-i is significantly different from the other ITecs because it is actually two ITecs: one for a process-driven and the other for a data-driven approach. This distinction is made via the activities where the activities for each approach are different and the only overlapping activity is ‘Execute the improvement’. The activities from only the TQdM-i process-driven approach have been included in this research because they are sufficiently general with regards to DQ improvement. By contrast, the data-driven ITec in TQdM-i is very specific to data warehouses. And whilst this is very useful for guiding data warehouse related DQ projects, this research focuses on generic DQ improvement techniques (see literature review inclusion criteria 1). The other selected ITecs demonstrate how to manage a process or data-driven approach within their constituent activities.

<b>Index</b>	<b>Activity name</b>	<b>Description</b>	<b>Source ITec(s)</b>
1 Probs	Select processes or problems to focus on	Identify a process or a DQ problem that DQ improvements can focus on and are most likely to yield significant benefits if the data can be improved.	TQdM-i COLDQ-i DQFG-i
2 Team	Build a DQ team	Select people who will manage and implement the improvement activities.	COLDQ-i DQFG-i TQdM-i
3 Root	Identify root causes of DQ problems	Investigate and identify all causes of a problem to determine its actual (root) cause(s).	TQdM-i TDQM-i DQFG-i SODQA-i EDQP-i
4 Options	Develop and select alternative data quality improvement options	Develop alternative data quality improvement options/remedies and select the option(s) to implement. For example, an option might be to ensure that people update the company database more frequently or distribute the updates to remote sites more often. Another option could be to perform data cleansing on the database at selected time intervals.	SODQA-i TQDM-i DQFG-i CDQM-i EDQP-i TQdM-i
5 Tools	Select tools for improvement	Select suitable tools (e.g. software, or formal methods) for improvement. An example of software includes data cleansing tools.	CDQM-i COLDQ-i
6 Plan	Plan the DQ improvement process	Develop a plan, which outlines how the DQ improvement process will be conducted (taking into consideration all constraints such as time, cost, availability of resources etc.).	CDQM-i
7 Costs	Conduct a cost/benefit analysis of improvement options	Develop a cost/benefit analysis using a prioritised list of improvement options as a basis. This should take into account the cost of the DQ improvement exercise and the costs of having poor data quality.	CDQM-i
8 Model	Define a metadata model	Define a metadata model and extract all relevant meta data to improve the current understanding of the existing data.	COLDQ-i
9 Define	Define DQ rules	Define the rules to which data must adhere. These could be existing business rules.	COLDQ-i
10 Rules	Determine what DQ rules currently exist	Determine what DQ rules currently exist and to what extent these rules are currently being followed.	COLDQ-i
11 Ext	Manage your suppliers	Determine what rules will be imposed on external data providers (for example, a set of DQ expectations and penalties for non-conformance).	COLDQ-i
12 Trial	Trial simple solutions to the DQ problem	Identify simple solutions to DQ problems as a starting point and trial these with the aim of demonstrating that the trial solutions work.	DQFG-i
13 Exe	Execute the improvement	Implement DQ improvement actions in a controlled manner to improve DQ. This may include the actual execution of software, or the initiation of actions to change business processes.	all ITecs
14 Check	Verify the effectiveness of improvement actions	Verify that the selected DQ improvements do solve the problem.	TQdM-i COLDQ-i DQFG-i EDQP-i
15 Comm	Communicate the results of the DQ improvement	Communicate and share the results of the DQ improvement with relevant people	EDQP-i

**Table 10: All DQ Improvement Related Activities**

### **Links between Activities**

When constructing a new ITec consisting of a set of activities, it is necessary to identify what order the activities need to be carried out in so that the resulting ITec is usable and does not contain unimplementable links between activities. To mention the most straightforward example, it does not make sense to verify the effectiveness of improvement actions before any improvement actions have been executed. To give an indication of the order in which all 15 activities should be carried out, the inputs of each activity were identified to determine exactly what ordering was intended in the original studies describing the ITecs. For the activity inputs that were explicitly recorded in the original sources, Tables 3 to 11 list these inputs. In each table, the ‘input activity’ states the abbreviated activity that is listed as an input to the activity shown at the top of each table, a description is also given for why this is listed as an input as well as the source ITec that lists the input.

<b>Activity number</b>	2 (Build a DQ team)	
<b>Source ITec(s)</b>	<b>Input activity</b>	<b>Description</b>
COLDQ-i	Probs	Various teams are built to address each problem. (p 475 step 9)
DQFG-i	Probs	the team is built using people that relate to parts of the information chain for each of the DQ problems. (p132 1st bullet).

**Table 11: Extraction of Inputs for Activity 2**

<b>Activity number</b>	3 (Identify root causes of DQ problems)	
<b>Source ITec(s)</b>	<b>Input activity</b>	<b>Description</b>
TQdM-i	Probs	A candidate process for improvement is shown as an input. (p293 1st input).
DQFG-i	Team	People close to the problem know about why it occurs. (p133).

**Table 12: Extraction of Inputs for Activity 3**

<b>Activity number</b>	4 (Develop and select alternative data quality improvement options)	
<b>Source ITec(s)</b>	<b>Input activity</b>	<b>Description</b>
EDQP-i	Root	Root causes is mentioned explicitly as an input to activity 4 (p 209, table 3.35)
DQFG-i	Root	It is stated that root causes can be used to help formulate the solutions. (p 133 step 4 D).
SODQA-i	Root	Figure 2 shows root causes as an input (p 216)
TQDM-i	Root	In the study, activity 4 starts after root causes (p 65 first sentence of ‘improve IP’ section)
TQdM-i	Root	Root cause is described as an implicit part of doing activity 4. (p 293 step 2)

**Table 13: Extraction of Inputs for Activity 4**

<b>Activity number</b>	5 (Select tools for improvement)	
<b>Source ITec(s)</b>	<b>Input activity</b>	<b>Description</b>
CDQM-i	Options	The activity matrix that defines the options is used to help select the tools. (p187 first sentence section 7.4.8).

**Table 14: Extraction of Inputs for Activity 5**

<b>Activity number</b>	6 (Plan the DQ improvement process)	
Source ITec(s)	Input activity	Description
CDQM-i	Costs	The plan should take into consideration all constraints such as time, cost, availability of resources etc. (see definition in Table3).

**Table 15: Extraction of Inputs for Activity 6**

<b>Activity number</b>	7 (Conduct a cost/benefit analysis of improvement options)	
Source ITec(s)	Input activity	Description
CDQM-i	Options	The improvement options are needed as input so that the cost can be evaluated for each option.(p188 7.4.10 1st sentence).

**Table 16: Extraction of Inputs for Activity 7**

<b>Activity number</b>	12 (Trial simple solutions to the DQ problem)	
Source ITec(s)	Input activity	Description
DQFG-i	Root	Root causes can be used to help implement the trial solutions. (p 133 step 4 D)

**Table 17: Extraction of Inputs for Activity 12**

<b>Activity number</b>	13 (Execute the improvement)	
Source ITec(s)	Input activity	Description
EDQP-i	Plan	The improvement plans are noted as an input. (p214 see first input in table).
DQFG-i	Trial	The trial solutions are rolled out in full once they have proven their worth. (p133 step 5 A and B).
TQdM-i	Options	The recommended improvement options are noted as an input. (p298 first input).

**Table 18: Extraction of Inputs for Activity 13**

<b>Activity number</b>	14 (Verify the effectiveness of improvement actions)	
Source ITec(s)	Input activity	Description
EDQP-i	Exe	The results from executing the improvement are defined as an input. (p223 see 1st 3 inputs in table).
TQdM-i	Exe	The measured results are defined as an input to verifying the effectiveness of the results. (p299 1st input).

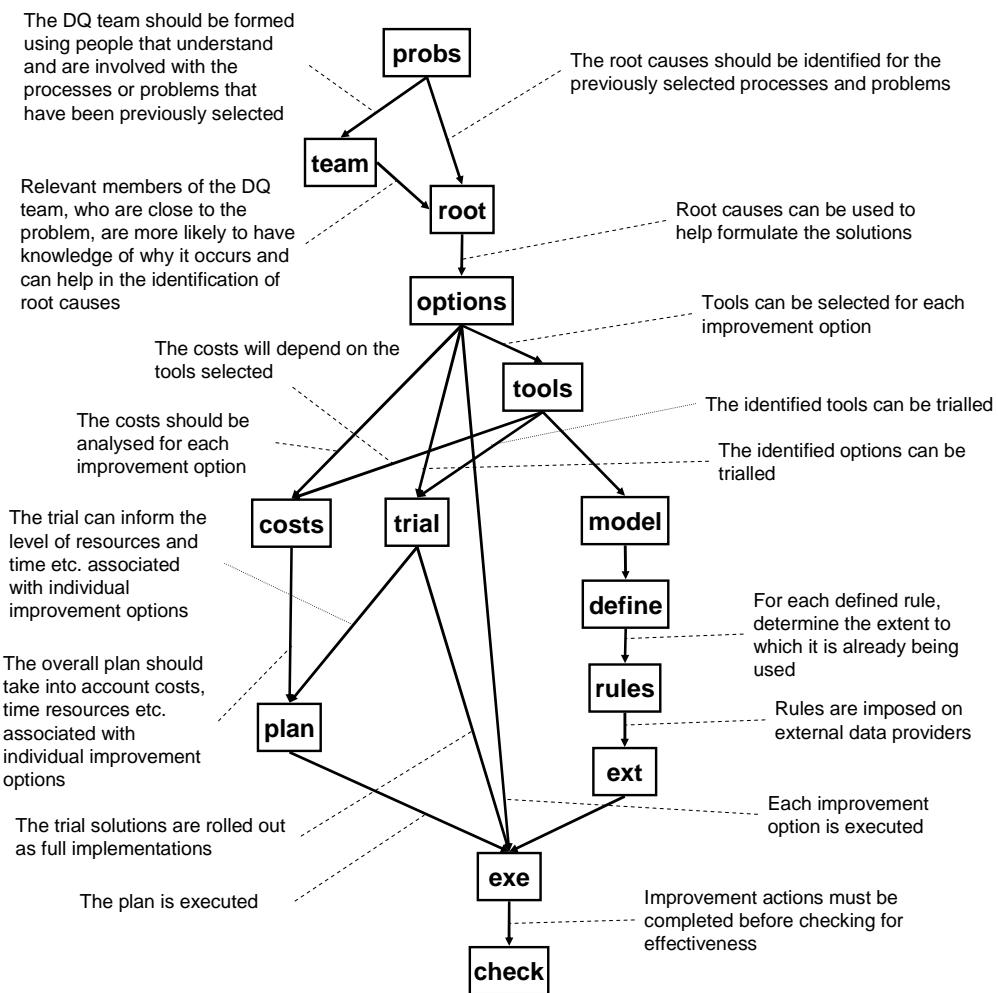
**Table 19: Extraction of Inputs for Activity 14**

#### 4. DQ IMPROVEMENT MODEL

The final model of DQ improvement, based on the activities and the ordering constraints, is shown in Figure 6. The boxes in Figure 6 represent the activities and the text in each box references the abbrevi-

ated name of the activity shown in Table 2. Activities are linked with arrows when one activity follows another, and the reason for this link is described in the text at the end of the dashed lines.

Activity 15 (Communicating the results of the improvement) is not shown in Figure 6 because the EDQP-i ITEc that proposes this activity specifies that it should be done over the course of the improvement project; not after or before a specific activity. Communicating the results should be done at intervals in order to keep members of the organisation informed of the progress of the DQ improvement project. Communicating the results can therefore be done at the discretion of the data quality assessor(s) and is not constrained by any ordering of the activities.



**Figure 6: A Model of the DQ Improvement Process**

The “model” activity, although it is linked to the “tools” activity, has no explicit input from the “tools” activity. This is also the case between the “model”-“define” and “ext”-“exe” activities. The orderings have been retained in the model because it is implicit in the COLDQ-i ITEc that the activities are carried out in this order even though there is no explicit description of inputs/outputs.

The links between certain pairs of activities have not been considered in the existing ITecs (because each ITec does not contain one of the activities in the pair) and therefore for these activities, there is no guidance on what paths may be followed to complete the DQ improvement. As well as combining the existing ITecs, this model also includes these missing links and describes how the inputs can be used. There are only two such additions (between “tools” and “trial”, and “trial” and “plan”) and these are shown with finer dotted lines between the link description and the activity links.

The model provides an indication of the activities that can be skipped by informing the data assessor of what inputs would be missing should he or she exclude an activity. For example, excluding the costs activity would mean that if the plan activity is conducted, then it would need to be based on other constraints other than cost, such as time and resources etc.

## 5. CUSTOMISING DATA QUALITY IMPROVEMENT

The model described in the previous section can be used to select the most suitable DQ improvement path that matches an organisation’s requirements. The following steps, extended from the steps in the Hybrid Approach to assessment (see [11]), are proposed to assist organisations in their attempts to determine the most suitable path.

### Step 1: State the initial motivation

The improvement process should start with a good understanding of the initial motivation for attempting to improve DQ. The initial motivation should follow from the results of a DQ assessment which provides an understanding of the current level of DQ, an indication of the current DQ problems, and the extent to which these DQ problems need to be improved. A typical statement of an initial motivation could be:

*To improve the accuracy of customer sales data, which has been identified as requiring improvement from a previous DQ assessment.*

### Step 2: Identify the company requirements related to the DQ improvement

Different companies will have different DQ improvement requirements relating to the context, such as the number of people working on the assessment, amount of data, criticality of the data etc. There may also be specific actions that the organisation has already decided are necessary and therefore wants the improvement process to include. This step requires the organisation wanting to assess DQ to identify these requirements, and a typical requirement statement could be:

*The improvement process needs to include the identification of root causes for inaccurate customer sales data*

### Step 3: Select ITec activities that meet the requirements

The aim of this step is to select the relevant activities, from the list in Table 2, which meet the requirements that were identified in the previous step. Note that in some cases it may be necessary to select more than one activity for a single requirement. The model of DQ improvement (Figure 6) can be used to plan the path through the relevant activities allowing the organisation to choose the most effective route. The general aim of this step is to select a series of activities that fulfil the needs of the organisation with respect to DQ improvement. The DQ improvement model then shows how these should be arranged.

### Step 4: Select specific methods for each activity

The final step involves selecting the concrete methods to be used to carry out each activity along the improvement path. Unfortunately, there is no single source that describes all the methods. One means of doing this is to use the original source (paper, book etc.) that described the activity and use the methods

described by the source. If more than one source describes an activity, the most useful/appropriate method can be used. The data assessor may also wish to use other methods for the activity, if desired.

## 6. CREATING A NEW ITEC

The following example uses the above steps to illustrate how to create a new ITec based on an industrial DQ improvement related scenario concerning an organisation that needs to improve maintenance related information. The organisation needs to maintain its assets (such as machine tools) that manufacture the products that the organisation sells. A DQ assessment found that data collected on machine tool vibration was not *complete* and poor maintenance decisions were being made causing the machine tools to produce defective products despite the data indicating that a break-down was not likely. The initial motivation for the company to start a DQ improvement exercise (step 1) is therefore to improve the completeness of vibration data.

The organisation wants to improve the current data in the information system containing vibration data and also to ensure that data will be entered into the system properly in the future. The information system is very large and it will not be feasible to modify, insert or update the values manually. An automated solution is therefore required. Furthermore, no existing personnel have DQ as their remit since the people who conducted the DQ assessment are no longer available. The DQ improvement requirements (step 2) therefore include determining what software, such as data cleansing and database synchronisation tools, are needed (either to be purchased or developed in-house) to fix the existing data, to identify and fix the root causes of the problems so that the future data will be correct when it is entered into the system, and to assign DQ related tasks to relevant personnel in the organisation, so that they can be responsible for DQ improvements.

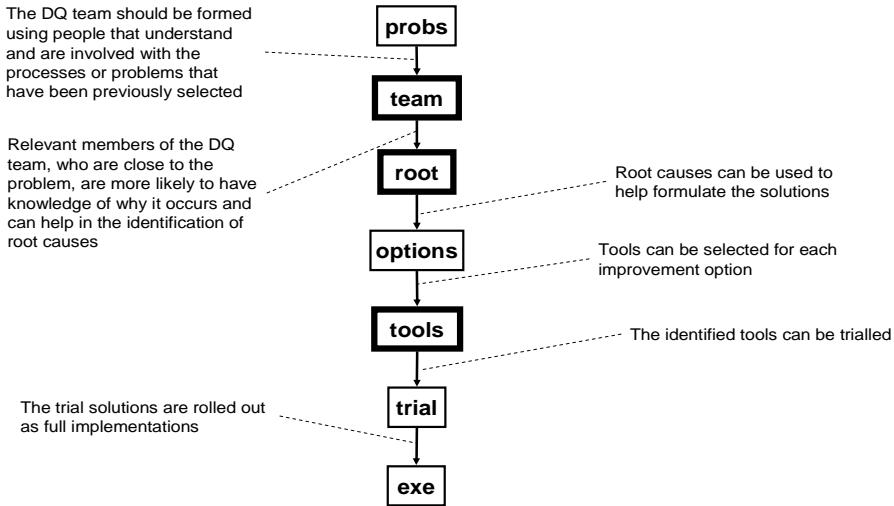
For step 3, the activities shown in Table 20 have been matched to these three requirements.

Requirement	Matching activity
Determine what DQ related software tools are needed	Select tools for the improvement
Determine the root causes of the DQ problems	Identify root causes of DQ problems
Assign DQ related tasks to relevant personnel in the organisation	Build a DQ team

Table 20: Mapping of Requirements to Activities

The DQ improvement model is then used to identify a suitable path that includes these three required activities and this is shown in Figure 7 (the required activities have been given thicker borders).

After the tools activity there are a number of alternative options, and the organisation decided that the best approach for them would be to trial the tools before rolling out the full implementation on the entire system. As the organisation conducts the improvement exercise they may, of course, use the model to identify other activities and decide that identifying the costs, which is another possible alternative in this case, is needed as well. During the improvement exercise, the organisation may use the model of DQ improvement so see what other paths/options are available given the results they have so far. It can therefore aid the decision making of what activities should be carried out on the basis of the progress so far.



**Figure 7: Path Taken by the Organisation**

Finally, for step 4, the actual concrete methods that should be used for each of these activities are extracted from the original sources describing the activities. The list of methods is shown in Table 21. In some cases, no method has been developed and there is only advice regarding how best to carry out the activity.

Ref	Activity	Selected method [and source ITec]
Probs	Select processes or problems to focus on	Advice: Select problems that have a noticeable impact, result in measureable cost savings, few political issues need to be addressed, there is senior management support, access to the problem space is open, and the problem can be feasibly solved [COLDQ-i]
Team	Build a DQ team	Advice: Appoint a DQ team containing the following roles: Project manager, system architect, domain expert, rules software engineer, and quality assurance and root cause analysis engineer [COLDQ-i]
Root	Identify root causes of DQ problems	Method: use the fishbone diagram/Ishikawa chart [TQdm-i][EDQP-i]
Options	Develop and select alternative data quality improvement options	Method: use the benefit versus cost matrix and “recommendations for action” template in EDQP-i. [EDQP-i]
Tools	Select tools for the improvement	Advice: A DQ project is likely to need the following tools: data cleansing, data standardisation, database checking/validation, rules definition system, rules execution system, approximate matching system. Estimate the number of times the application is likely to execute in order to determine what is cost-effective. [COLDQ-i]
Trial	Trial simple solutions to the DQ problem	Advice: put the solution as close to the problem as possible [DQFG-i]
Exe	Execute the improvement	Execute the tools from activity “tools” and carry out the recommendations for action from activity “options”.

**Table 21: Selection of Methods for each Activity**

## 7. LIMITATIONS

One of the main limitations of this work is that some ITeCs describe in detail how the activities are linked together, whereas others do not explicitly state why one activity should follow another. This means that it was not always possible to identify a clear link between activities in the model of DQ improvement (such as between the ‘tools’ and the ‘model’ activities); future work could address this point. Another limitation, which is inherent in the literature review method used, is the subjective nature of extracting the activities and links. To mitigate this problem we applied numerous peer review checks as far as possible in an attempt to ensure that activities were extracted as accurately as possible and that multiple opinions were considered.

## 8. CONCLUSIONS

Existing ITeCs prescribe a static set of activities which should be completed in order to improve DQ. However, each organisation wanting to initiate a DQ improvement exercise will have differing requirements related to the improvement, such as the need to carry out certain activities whilst omitting others due to time and budgetary constraints. The DQ model proposed in this paper shows how the DQ improvement process can be dynamic and can take different paths depending on what the data assessor requires. The model gives an indication of how the activities, used in a DQ improvement exercise, are linked together and therefore can be used to identify how one should progress from each activity and which activities can be avoided as required. The four step process described in this paper, coupled with the model, can be used to assist the organisation in this endeavour.

There have been numerous proposals of ITeCs that each give their own perspective on the problem of improving DQ. Whilst it is useful to have numerous options available to the data professional, the problem is that most are of a similar nature, and it is not clear which activities are fundamental and which are specific to a particular application. This work is therefore also a step towards identifying the fundamental activities required for DQ improvement. DQ ITeCs specialised to particular contexts/applications can therefore be clear about what activities are special to that context/application whilst retaining a familiar base of DQ improvement activities. Any future developments of ITeCs should therefore explicitly consider the extent to which they are different and how they overlap with the existing techniques. The model of DQ improvement proposed in this paper can help the researcher integrate any new ITeCs into the existing work.

## ACKNOWLEDGEMENTS

We gratefully acknowledge EPSRC who funded this research, project reference number EP/G038171/1.

## REFERENCES

- [1] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A., “Methodologies for Data Quality Assessment and Improvement,” *ACM Computing Surveys*, 41 (3), 2009, pp.1–52.
- [2] Batini, C., and Scannapieco, M., *Data Quality: Concepts, Methodologies and Techniques*. 1st ed., Springer, 2006.
- [3] English, L., *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. John Wiley & Sons, 1999.
- [4] Fisher, C., and Kingma, B., “Criticality of Data Quality as Exemplified in Two Disasters,” *Information & Management*, 39 (2), 2001, pp.109–116.
- [5] Loshin, D., *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann Pub, 2001.
- [6] McGilvray, D., *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Informa-*

- tion.*Morgan Kaufmann, 2008.
- [7] Pipino, L.L., Lee, Y.W., and Wang, R.Y., “Data Quality Assessment,” *Communications of the ACM*, 45 (4), 2002, pp.211–218.
- [8] Redman, T.C., *Data Quality: The Field Guide*.Digital Press,Boston, [MA.] 2001.
- [9] Wang, R.Y., “A Product Perspective on Total Data Quality Management,” *Communications of the ACM*, 41 (2), 1998, pp.58–65.
- [10] Wang, R.Y., and Strong, D.M., “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, 12 (4), 1996, pp.5–34.
- [11] Woodall, P., and Parlakad, A., “A Hybrid Approach to Assessing Data Quality,” Proceedings of the 2010 International Conference on Information Quality, 2010, .

# CHECKING AND REPAIRING THE QUALITY OF INFORMATION IN DATABASES BY INCONSISTENCY METRICS

(*Research-in-Progress*)

**Hendrik Decker**

Instituto Tecnológico de Informática, UPV, Ciudad Politécnica de la Innovación, 46022 Valencia, Spain  
[hendrik@iti.upv.es](mailto:hendrik@iti.upv.es)

**Abstract:** To some extent, the quality of the data stored in information systems can be modeled by constraints that postulate conditions of consistency. Thus, quality corresponds to the absence of inconsistency. Inconsistency can be measured by metrics that size or count the violated instances of constraints, and also by metrics that size or count the causes of inconsistency, i.e., the database facts that are responsible for constraint violations. Maintaining the consistency of databases is usually done in two ways: constraint violations are either prevented by checking constraints upon updates, or eliminated by repairing inconsistencies. Hence, the quality of the information provided by a database can be maintained by checking or repairing constraints. We show how both checking and repairing can be realized by inconsistency metrics. As opposed to conventional methods for checking and repairing, such metrics enable the tolerance of inconsistency, i.e., of impaired data quality, which is necessary in databases that may contain information that lacks quality. Inconsistency tolerance also enables an extension of quality control by metric-based quality management to concurrent multi-user databases and distributed systems.

**Key Words:** Data Quality, Inconsistency Metrics, Quality Checking, Quality Repairing, Concurrent Multi-Users

## 1 INTRODUCTION

In earlier work [6], it has already been shown that the quality of the data stored in databases and information systems can be modelled, measured and monitored by semantic constraints and suitable constraint checking methods. Such constraints are expressions in the data description language (usually SQL). They formalize conditions of consistency that are required to be satisfied in each database state. Thus, violations of such constraints correspond to a lack of data quality.

In order to control the quality of information across database updates, constraint violations should be quantifiable by inconsistency metrics. In [6], the essential idea of quantifying the quality of information was to size the set of violated instances of constraints. If, by any update  $U$ , that set would be discovered to increase, then  $U$  should be rejected. Also a simple count of all violated instances of constraints may serve as an inconsistency metric: if  $U$  would increase that count, then  $U$  should be rejected. Not only the (cardinality of) sets of violated cases, but also the database facts that cause the violations can be sized or counted for quantifying the amount of inconsistency in the information provided by a database. That idea was elaborated in [9].

In this paper, we go beyond the achievements of [6, 9] in three ways. Firstly, we generalize the case-based approach of [6] and the cause-based approach of [9] to a generic concept based on arbitrary inconsistency metrics, including a thorough axiomatization of such metrics. Also, we consider new metrics that refine the mentioned case- and cause-based metrics by application-specific weights assigned to cases or causes. Secondly, we enlarge the scope of quality maintenance, which so far has consisted only in constraint checking upon updates, by metric-based quality repairing. Repairs are updates that eliminate extant quality impairments, i.e., constraint violations. In particular, the inconsistency tolerance of our approach enables a concept of partial repairs that avoids the side effects of repairs that are not

quality-preserving. Thirdly, we recall that currently, the support for quality maintenance and hence of quality control in concurrent transactions is largely left to the user. However, our inconsistency-tolerant approach to quality maintenance also enables an automatization of quality control in distributed systems with concurrent transactions.

In Section 2, we outline formal preliminaries for the remainder. In Section 3, we define a generic concept of inconsistency metrics. In Section 4, we apply this concept to inconsistency-tolerant quality checking and repairing. In Section 5, we shall see that, as opposed to conventional constraint maintenance, the metric-based approach to quality maintenance can be extended without efforts to concurrent transactions. In Section 6, we conclude.

## 2 PRELIMINARIES

### 2.1 Database Issues

Unless specified otherwise, we use notations and terminology that are common for *datalog* [1, 4] and first-order predicate logic [10]. For an update  $U$  of a database  $D$ , we denote the updated database by  $D^U$ . An *update request* is a sentence  $R$  required to become true by updating  $D$ . *Repairs* are updates that satisfy update requests by eliminating constraint violations.

*Quality constraints* (in short, *constraints*) usually are represented by *denial clauses*, i.e., universal sentences of the form  $\leftarrow B$ , where  $B$  is a conjunction of literals that asserts what *should not hold* in any database state. A *quality theory* is a finite set of quality constraints.

Let symbols such as  $D, Q, I, U$  always stand for a database, a quality theory, a constraint and, resp., an update. For each sentence  $F$ , we may write  $D(F) = \text{true}$  (resp.,  $D(F) = \text{false}$ ) if  $F$  evaluates to *true* (resp., *false*) in  $D$ . Similarly, we write  $D(Q) = \text{true}$  (resp.,  $D(Q) = \text{false}$ ) if each constraint in  $Q$  is satisfied in  $D$  (resp., at least one constraint in  $Q$  is violated in  $D$ ). Let  $\text{vioCon}(D, Q)$  denote the set of violated constraints in  $D$ .

Let  $H$  be a universal Herbrand base and  $\mathcal{N}$  a universal set of constants, represented w.l.o.g. by natural numbers, in the language of each database. We may use ';' to delimit elements of sets since ';' also denotes conjunction in the body of clauses. Symbols  $\models, \Rightarrow$  and  $\Leftrightarrow$  denote logical consequence (i.e., truth in all Herbrand models), meta-implication and, resp., meta-equivalence. By overloading, we use  $=$  as identity, assignment in substitutions, or meta-level equality. Negations of  $\models$  and  $=$  are denoted by  $\not\models$  and, resp.,  $\neq$ .

### 2.2 Cases and Causes

Similar to  $\text{vioCon}$ , cases and causes are the basis of the inconsistency metrics that are presented in 3.2.

A *case* is a ground instance of a constraint. Let  $\text{Cas}(Q)$  be the set of all cases of constraints in  $Q$ , and  $\text{vioCas}(D, Q) = \{C \mid C \in \text{Cas}(Q) \mid D(C) = \text{false}\}$  the set of all violated cases of  $Q$  in  $D$ .

Causes have been introduced in [9]. Below, we recapitulate their definition.

The well-known *completion* of a database  $D$  be denoted by  $\text{comp}(D)$  [5]. It essentially consists of the if-and-only-if completions (in short, *completions*) of all predicates in the underlying language. For a predicate  $p$ , let  $pD$  denote the completion of  $p$  in  $D$ .

**Definition 1.**

Let  $D$  be a database,  $p$  a predicate,  $n$  the arity of  $p$ ,  $x_1, \dots, x_n$  the  $\forall$ -quantified variables in  $pD$  and  $\varphi$  a substitution of  $x_1, \dots, x_n$ . For  $A = p(x_1, \dots, x_n)\varphi$ , the *completion* of  $A$  in  $D$  is obtained by applying  $\varphi$  to  $pD$  and is denoted by  $AD$ . Further, let  $\underline{\text{comp}}(D) = \{AD \mid A \text{ in } H\}$ , and  $\text{if}(D)$  and  $\text{only-if}(D)$  be obtained by replacing  $\leftrightarrow$  in each  $AD$  in  $\underline{\text{comp}}(D)$  by  $\leftarrow$  and, resp.,  $\rightarrow$ . Finally, let  $\text{iff}(D)$  be the union of  $\text{if}(D)$  and  $\text{only-if}(D)$ . The usual equality axioms of  $\text{comp}(D)$  be associated by default also to  $\text{iff}(D)$ .

**Definition 2.**

Let  $D$  be a database and  $I = \leftarrow B$  a constraint such that  $D(\exists B) = \text{true}$ . A subset  $E$  of  $\text{iff}(D)$  is called a *cause of the violation* of  $I$  in  $D$  if  $E \models \exists B$ , and for each proper subset  $E'$  of  $E$ ,  $E' \not\models \leftarrow B$ . We also say:  $E$  is a *cause* of  $\exists B$  in  $D$  if  $E$  is a cause of the violation of  $B$  in  $D$ ;  $E$  is a *cause of the violation* of  $Q$  in  $D$  if  $E$  is a cause of the violation of a denial form of the conjunction of all  $I$  in  $2^Q$ . Let  $\text{vioCau}(D, Q)$  be the set of all causes of the violation of  $Q$  in  $D$ .

### 3 INCONSISTENCY METRICS

Inconsistency metrics are a special kind of measures for quantifying the amount of quality violation in databases. In 3.1, we axiomatize inconsistency metrics. In 3.2, we illustrate that by several examples. In 3.3, we discuss the desirability of some more axioms that are commonly associated to measures.

#### 3.1 Axiomatizing Inconsistency Metrics

An inconsistency metric is a mapping  $v$  from pairs  $(D, Q)$  to a metric space  $\Sigma$  that is structured by a partial order  $\leq$  with an infimum  $o$ , a distance  $\delta$  and an addition  $\oplus$  with neutral element  $o$ . The partial order allows to compare the amount of inconsistency in consecutive states  $(D, Q)$  and  $(D^U, Q)$ . With the distance  $\delta$ , the difference, i.e., the increase or decrease of inconsistency between  $D$  and  $D^U$ , can be sized. The addition  $\oplus$  allows to state a standard metric axiom for  $\delta$ , and  $o$  is, at a time, the smallest element of  $(\Sigma, \leq)$  and the neutral element of  $\oplus$ .

The purpose of  $v(D, Q)$  is to size the amount of inconsistency in  $(D, Q)$ . Definitions 3 and 4 below specialize conventional axioms of metric spaces and measures. In a sense, these definitions also generalize conventional axiomatizations, since they allow to size and compare different amounts of inconsistency without necessarily quantifying them numerically. With  $S = 2^{\text{Cas}(Q)}$ , for instance,  $\leq = \subseteq$ ,  $\delta = | - |$  (symmetric difference),  $\oplus = \cup$  and  $o = \emptyset$ , it is possible to measure the inconsistency of  $(D, Q)$  by sizing  $\text{vioCas}(D, Q)$ .

**Definition 3.**

A structure  $(\Sigma, \leq, \delta, \oplus, o)$  is called a *metric space for quality violation* (in short, a *metric space*) if  $(\Sigma, \oplus)$  is a commutative monoid with neutral element  $o$ ,  $\leq$  is a partial order on  $\Sigma$  with infimum  $o$ , and  $\delta$  is a distance on  $\Sigma$ . More precisely, for each  $m, m', m'' \in \Sigma$ , axioms (1)–(4) hold for  $\leq$ , (5)–(8) for  $\oplus$ , and (9)–(11) for  $\delta$ .

$$m \leq m \quad (\text{reflexivity}) \quad (1)$$

$$m \leq m', \quad m' \leq m \quad \Rightarrow \quad m = m' \quad (\text{antisymmetry}) \quad (2)$$

$$m \leq m', \quad m' \leq m'' \quad \Rightarrow \quad m \leq m'' \quad (\text{transitivity}) \quad (3)$$

$$o \leq m \quad (\text{infimum}) \quad (4)$$

$$m \oplus (m' \oplus m'') = (m \oplus m') \oplus m'' \text{ (associativity)} \quad (5)$$

$$m \oplus m' = m' \oplus m \quad (\text{commutativity}) \quad (6)$$

$$m \oplus o = m \quad (\text{neutrality}) \quad (7)$$

$$m \leq m \oplus m' \quad (\oplus\text{-monotonicity}) \quad (8)$$

$$\delta(m, m') = \delta(m', m) \quad (\text{symmetry}) \quad (9)$$

$$\delta(m, m) = o \quad (\text{identity}) \quad (10)$$

$$\delta(m, m'') \leq \delta(m, m') \oplus \delta(m', m'') \quad (\text{triangle inequality}) \quad (11)$$

Let  $m < m'$  denote that  $m \leq m'$  and  $m \neq m'$ .

#### **Example 1.**

$(\mathbb{N}_0, \leq, | - |, +, 0)$  is a metric space for quality violation, where  $\mathbb{N}_0$  is the set of non-negative integers. In this space,  $\text{vioCon}(D, Q)$ ,  $\text{vioCas}(D, Q)$  or  $\text{vioCau}(D, Q)$  can be counted and compared. As already indicated, these three sets may also be sized and compared in the metric spaces  $(2^X, \subseteq, \ominus, \cup, \emptyset)$ , where  $X$  stands for  $Q$ ,  $\text{Cas}(Q)$  or  $\text{iff}(D)$ , respectively, and  $\ominus$  is the symmetric set difference.

#### **Definition 4.**

We say that  $v$  is an *inconsistency metric* (in short, a *metric*) if  $v$  is a mapping of pairs  $(D, Q)$  to a metric space  $(\Sigma, \leq, \delta, \oplus, o)$  for quality violation.

### **3.2 Examples of Inconsistency Metrics**

#### **Example 2.**

A coarse, simple metric  $\beta$  is defined by the equation  $\beta(D, Q) = D(Q)$ , where the range of  $\beta$  is the binary metric space  $(\{\text{true}, \text{false}\}, \leq, \tau, \wedge, \text{true})$ . In this space,  $\leq$  and  $\tau$  are defined by stipulating  $\text{true} \leq \text{false}$  (i.e., satisfaction means lower inconsistency than violation), and, resp.,  $\tau(w, w') = \text{true}$  if  $w = w'$ , otherwise  $\tau(w, w') = \text{false}$ , for  $w, w' \in \{\text{true}, \text{false}\}$ .

Clearly,  $\beta$  and its metric space reflect the classical logic distinction that a set of formulas is either consistent or inconsistent, without any further differentiation of different degrees of quality. The meaning of  $\tau$  is that each consistent pair  $(D, Q)$  is equally good, and each inconsistent pair  $(D, Q)$  is equally bad.

#### **Example 3.**

The metrics  $\iota$  and  $|\iota|$  compare or, resp., count the set of violated constraints in  $Q$ . They are defined by the equations  $\iota(D, Q) = \text{vioCon}(Q, D)$  and  $|\iota|(D, Q) = |\iota(D, Q)|$ , where  $|\cdot|$  is the cardinality operator, with metric spaces  $(2^Q, \subseteq, \ominus, \cup, \emptyset)$  and, resp.,  $(\mathbb{N}_0, \leq, | - |, +, 0)$ . Two more fine-grained metrics are given by  $\zeta(D, Q) = \text{vioCas}(Q, D)$  and  $|\zeta|(D, Q) = |\zeta(D, Q)|$ , with metric spaces  $(2^{\text{Cas}(Q)}, \subseteq, \ominus, \cup, \emptyset)$  and, resp.,  $(\mathbb{N}_0, \leq, | - |, +, 0)$ . Similarly,  $\kappa(D, Q) = \text{vioCau}(Q, D)$  and  $|\kappa|(D, Q) = |\kappa(D, Q)|$ , define cause-based metrics, with metric spaces  $(2^{\text{iff}(D)}, \subseteq, \ominus, \cup, \emptyset)$  and, resp., again  $(\mathbb{N}_0, \leq, | - |, +, 0)$ . Other metrics are addressed in subsections 3.3.1 and 4.1.

### 3.3 More Axioms for Inconsistency Metrics

In 3.3.1, we argue that the standard axiom of positive definiteness of measures is not cogent for inconsistency metrics. In 3.3.2, we show that the standard axiom of additivity of measures is invalid for inconsistency metrics. In 3.3.3, we also dismiss the standard axiom of monotonicity of measures for inconsistency metrics, and propose a valuable variant.

#### 3.3.1 Definiteness

For conventional measures  $\mu$ , definiteness means that  $\mu(S) = 0$  if and only if  $S = \emptyset$ , for  $S \in \Sigma$ . For inconsistency metrics  $v$ , that takes, for each  $(D, Q)$ , the form

$$v(D, Q) = o \Leftrightarrow D(Q) = \text{true} \quad (\text{definiteness}) \quad (12)$$

Clearly, (12) assigns the least inconsistency value  $o$  precisely to those databases that totally satisfy all constraints in  $Q$ . It is easy to show the following result.

#### Theorem 1

Each of the metrics  $\beta, \iota, |\iota|, \zeta, |\zeta|, \kappa, |\kappa|$  in 3.2 fulfills (12).

Axioms corresponding to (12) are standard in the literature on measure theory [2]. Yet, (12) is not cogent for inconsistency metrics. That is shown by the following modification  $\zeta'$  of  $|\zeta|$ . Let  $\zeta'(D, Q) = 0$  if  $|\zeta|(D, Q) \in \{0, 1\}$ , otherwise  $\zeta'(D, Q) = |\zeta|(D, Q)$ . Thus,  $\zeta'$  considers each inconsistency that consists of just a single violated ground case as insignificant. Hence,  $\zeta'$  does not obey (12) but arguably is a very reasonable inconsistency metric that tolerates negligible amounts of inconsistency.

#### 3.3.2 Additivity and Monotony

For conventional measures  $\mu$ , additivity means  $\mu(S \cup S') = \mu(S) + \mu(S')$ , for disjoint sets  $S, S'$  in  $\Sigma$ . For inconsistency metrics  $v$ , additivity takes the form

$$v(D \cup D', Q \cup Q') = v(D, Q) \oplus (D', Q') \quad (\text{additivity}) \quad (13)$$

for each  $(D, Q), (D', Q')$  such that  $D$  and  $D'$  as well as  $Q$  and  $Q'$  are disjoint.

Additivity is standard for traditional measures. However, (13) is invalid for inconsistency metrics, as shown by the following example.

#### Example 4.

Let  $D = \{p\}$ ,  $Q = \emptyset$ ,  $D' = \emptyset$ ,  $Q' = \{\neg p\}$ . Obviously,  $D(Q) = \text{true}$  and  $D'(Q') = \text{true}$ . Thus, it follows that  $|\zeta|(D, Q) + |\zeta|(D', Q') = 0$ , but  $|\zeta|(D \cup D', Q \cup Q') = 1$ .

For conventional measures  $\mu$ , monotonicity means  $S \subseteq S' \Rightarrow \mu(S) \leq \mu(S')$ , for each pair of sets  $S, S'$  in  $\Sigma$ . For inconsistency metrics  $v$ , monotonicity takes the form

$$D \subseteq D'; Q \subseteq Q' \Rightarrow v(D, Q) \leq (D', Q') \quad (v\text{-monotonicity}) \quad (14)$$

for each pair of pairs  $(D, Q), (D', Q')$ .

An axiom corresponding to (14) is postulated for inconsistency metrics in [11]. For definite databases and quality theories (i.e., the bodies of clauses do not contain any negative literal), it is easy to show the following result.

**Theorem 2.**

For definite databases  $D, D'$  and only definite constraints in  $Q, Q'$ , each of the metrics  $\beta, \iota, |\iota|, \zeta, |\zeta|, \kappa, |\kappa|$  in 3.2 fulfills (14).

However, due to the non-monotonicity of negation in the body of clauses, (14) is not valid for non-definite databases or non-definite constraints, as shown by Example 5, in which the foreign key constraint  $\forall x (q(x, y) \rightarrow \exists z s(x, z))$  on the  $x$ -column of  $q$  referencing the  $x$ -column of  $s$  is rewritten into de-nial form (we ignore the primary key constraint on the  $x$ -column of  $s$  since it is not relevant).

**Example 5.**

Let  $D = \{p(x) \leftarrow q(x, y), \neg r(x); r(x) \leftarrow s(x, z); q(1, 2); s(2, 1)\}$  and  $Q = \{\leftarrow p(x)\}$ . Clearly,  $D(Q) = \text{false}$  and  $|\zeta|(D, Q) = 1$ . For  $D' = D \cup \{s(1; 1)\}$  and  $Q' = Q$ , we have  $D'(Q') = \text{true}$ , hence  $|\zeta|(D', Q') = 0$ .

A variant of (14) that holds also for non-definite databases and constraints, requires that the measured amount of inconsistency in databases that violate quality is never lower than the measured inconsistency in databases that satisfy all constraints. Thus, for each pair of pairs  $(D, Q), (D', Q')$ , the following axiom is asked to hold.

$$D(Q) = \text{true}, D'(Q') = \text{false} \Rightarrow v(D, Q) \leq v(D', Q') \quad (15)$$

It is easy to show the following result.

**Theorem 3.**

Each of the metrics  $\beta, \iota, |\iota|, \zeta, |\zeta|, \kappa, |\kappa|$  in 3.2 fulfills (15).

## 4 QUALITY MAINTENANCE

To maintain the quality of data, constraint violations should be prevented or repaired. However, it may be impractical or unfeasible to totally avoid inconsistency, or to repair all violated constraints at once. Thus, inconsistency tolerance is needed. That can be achieved by inconsistency metrics.

In 4.1, we revisit metric-based inconsistency-tolerant quality checking of updates (abbr. ITQC). Also, we show how to confine inconsistency by assigning weights to violated cases of constraints. Moreover, we show how to generalize metric-based ITQC by allowing for certain increases of inconsistency that are bounded by some thresholds. In 4.2, we outline how metric-based inconsistency-tolerant quality checking can be used also for making quality repairing inconsistency-tolerant.

### 4.1 Metric-based Inconsistency-tolerant Quality Checking

Definition 5, below, characterizes quality checking methods that may accept updates if there is no increase of inconsistency, no matter if there is any extant constraint violation or not. It abstractly captures metric-based ITQC methods as black boxes, of which nothing but their i/o interface is observable. More precisely, each method  $M$  is described as a mapping from triples  $(D, Q, U)$  to  $\{ok, ko\}$ . Intuitively,  $ok$  means that  $U$  does not increase the amount of measured inconsistency, and  $ko$  that it may.

**Definition 5.** (*Inconsistency-tolerant Quality Checking; abbr. ITQC*)

A *quality checking method* maps triples  $(D, Q, U)$  to  $\{ok, ko\}$ . For a metric  $v$ , the range of which is structured by a partial order  $\leq$ , a method  $M$  is called *sound* (resp., *complete*) for  $v$ -based ITQC if, for each triple  $(D, Q, U)$ , (16) (resp., (17)) holds.

$$M(D, Q, U) = ok \Rightarrow v(D^U, Q) \leq v(D, Q) \quad (16)$$

$$v(D^U, Q) \leq v(D, Q) \Rightarrow M(D, Q, U) = ok \quad (17)$$

Each method  $M$  that is sound for  $v$ -based ITQC is also called a *v-based method*.

Intuitively, (16) says:  $M$  is sound if, whenever it outputs *ok*, the amount of violation of  $Q$  in  $D$  as measured by  $v$  is not increased by  $U$ . Conversely, (17) says:  $M$  is complete if it outputs *ok* whenever the update  $U$  that is checked by  $M$  does not increase the amount of quality violation.

As opposed to ITQC, traditional constraint checking (abbr. TCC) imposes the *total consistency requirement*. That is, TCC additionally requires  $D(Q) = true$  in the premises of (16) and (17). The metric used in TCC is  $\beta$  (cf. Example 2). Since ITQC is defined not just for  $\beta$  but for any inconsistency metric  $v$ , and since TCC is not applicable if  $D(Q) = false$ , while ITQC is, Definition 5 generalizes TCC. Moreover, the definition of ITQC in [7] is equivalent to Definition 5 for  $v = \zeta$ . Hence, the latter also generalizes ITQC as defined in [7].

In [7], we have shown that the total consistency requirement is dispensable for most TCC approaches. Similar to corresponding proofs in [7], it can be shown that not all, but most TCC methods, including built-in constraint checks in common DBMSs (e.g., for primary or foreign keys), are  $v$ -based, for each  $v$  in  $\{\iota, |\iota|, \zeta, |\zeta|, \kappa, |\kappa|\}$ . The following results are easily shown by applying the definitions.

**Theorem 4.**

If a method  $M$  is  $v$ -based, then it is  $|v|$ -based, for each  $v$  in  $\{\iota, \zeta, \kappa\}$ . If  $M$  is  $\kappa$ -based, then it is  $\zeta$ -based. If  $M$  is  $\zeta$ -based, then it is  $\iota$ -based. The converse of none of these implications holds.

Each of the metrics assigns the same significance to each case or cause of quality violation. However, depending on the application, certain cases or causes may well have more or less impact with regard to the degree of damage they inflict on the quality of the stored information. Example 6, below, illustrates how the metrics  $|\iota|, |\zeta|, |\kappa|$  that count violated constraints, cases or causes thereof can be generalized by assigning weight factors to the counted entities. Instead of indiscriminately giving the same importance to each case or cause, such weights are useful for modeling application-specific degrees of violated quality. A simple variant of such an assignment comes into effect whenever 'soft' constraints that *ought to* be satisfied are distinguished from 'hard' constraints that *must* be satisfied.

**Example 6.**

Let  $lr$  and  $hr$  be two predicates that model a low, resp., high risk. Further,  $I1 = \leftarrow lr(x)$ ,  $I2 = \leftarrow hr(x)$ , be a soft, resp., hard constraint for protecting against low and, resp., high risks, where  $lr$  and  $hr$  are defined by the database view clauses  $lr(x) \leftarrow p(y, z), x = y + z, x > th, y \leq z$  and  $hr(x) \leftarrow p(y, z), x = y + z, x > th, z < y$ , resp., where  $th$  is a threshold value that should not be exceeded. and  $p(8, 3)$  be the only cause of quality violation in some database  $D$ . Now, for each  $v$  in  $\{\iota, \zeta, \kappa\}$ , no  $v$ -based method would accept the

update  $U = \{\text{delete } p(8, 3), \text{insert } p(3, 8)\}$ , although the high risk provoked by  $p(8; 3)$  is diminished to the low risk produced by  $p(3, 8)$ . However, metrics that assign weights to cases of  $I_2$  that are higher than those of  $I_1$  can avoid that problem. For instance, consider the metric  $\omega$  that counts the numbers  $n_1$  and  $n_2$  of violated cases of  $I_1$  and, resp.,  $I_2$  in  $D$ , and assigns  $f_1 n_1 + f_2 n_2$  to  $(D, \{I_1, I_2\})$ , where  $0 < f_1 < f_2$ . Clearly, each  $\omega$ -based method will accept  $U$ .

## 4.2 Quality Repairs

Roughly, repairing means to compute and execute an update in order to eliminate quality violation. Thus, each repair can be identified with an update. In 4.2.1, we formalize repairs and illustrate them by examples. In 4.2.2, we outline how to compute repairs.

### 4.2.1 Formalizing Repairs

In [7], we have distinguished total and partial quality repairs. The former eliminate all inconsistencies, the latter only some. Partial repairs tolerate inconsistency, since violated constraints may persist, as illustrated by Example 7.

#### Example 7.

Let  $D = \{p(a, b, c), p(b, b, c), p(c, b, c), q(a, c), q(c, b), q(c, c)\}$  and  $Q = \{\leftarrow p(x, y, z), \sim q(x, z); \leftarrow q(x, x)\}$ . Clearly, the violated cases of  $Q$  in  $D$  are  $\leftarrow p(b, b, c), \sim q(b, c)$  and  $\leftarrow q(c, c)$ . Each of the updates  $U_1 = \{\text{insert } q(b, c)\}$  and  $U_2 = \{\text{delete } p(b, b, c)\}$  is a partial repair of  $(D, Q)$ , since both fix the violation of  $\leftarrow p(b, b, c), \sim q(b, c)\}$  in  $D$ . Similarly,  $U_3 = \{\text{delete } q(c, c)\}$  is a partial repair that fixes the violation of the violation of  $\{\leftarrow q(c, c)\}$  in  $D$ .

Sadly, partial repairs may cause new violations, as shown in Example 8.

#### Example 8.

Consider again Example 7. As opposed to  $U_1$  and  $U_2$ ,  $U_3$  causes a new violation:  $\leftarrow p(c, b, c), \sim q(c, c)$  is satisfied in  $D$  but not in  $D^{U_3}$ . Thus, the partial repair  $U_4 = \{\text{delete } q(c, c), \text{delete } p(c, b, c)\}$  is needed to eliminate the violation of  $q(c, c)$  in  $D$  without causing any violation that did not exist before executing the partial repair.

Definition 6, below, generalizes the definition of partial repairs by requiring that each repair must decrease the measured amount of quality violation.

#### Definition 6. (Repair)

Let  $D$  be a database,  $Q$  a quality theory such that  $D(Q) = \text{false}$ ,  $v$  an inconsistency metric and  $U$  an update.

- $U$  is said to *preserve quality* wrt.  $v$  if  $v(D^U, Q) \leq v(D, Q)$  holds.
- For each proper subset  $S$  of  $\text{Cas}(Q)$  such that  $D(S) = \text{false}$  and  $D^U(S) = \text{true}$ ,  $U$  is called a *partial repair* of  $(D, Q)$ .
- $U$  is called a  *$v$ -based repair* of  $(D, Q)$  if  $v(D^U, Q) < v(D, Q)$  holds. If, additionally,  $D^U(Q) = \text{false}$ ,  $U$  is also called a  *$v$ -based patch* of  $(D, Q)$ . Else, if  $D^U(Q) = \text{true}$ ,  $U$  is called a *total repair* of  $(D, Q)$ .

Definition 6c could be slightly modified by replacing  $D^U(Q) = \text{false}$  and  $D^U(Q) = \text{true}$  by  $o < v(D, Q)$  and  $v(D, Q) = o$ , respectively. For each  $v$  in  $\{\iota, |\iota|, \zeta, |\zeta|, \kappa, |\kappa|\}$ , that replacement yields a definition that is equivalent to Definition 6. Moreover, it is easy to show the following result.

**Theorem 5.**

For each pair  $(D, Q)$  and each  $v$  in  $\{\iota, |\iota|, \zeta, |\zeta|\}$ , each  $v$ -based patch of  $(D, Q)$  is a partial repair of  $(D, Q)$ .

Note that the converse of Theorem 5 does not hold, as seen in Example 8. Theorem 5 also does not hold for  $v$  in  $\{\kappa, |\kappa|\}$ , since the violation of some case  $C$  may have  $n$  causes,  $n > 0$ , in some database  $D$ , and a repair  $U$  may just eliminate one of the causes that violate  $C$ . Then, for  $v$  in  $\{\kappa, |\kappa|\}$ ,  $v(D^U, Q) < v(D, Q)$ , i.e.,  $U$  is a  $v$ -based patch but not a partial repair of  $(D, Q)$ , since  $\text{vioCas}(D, Q) = \text{vioCas}(D^U, Q)$ , hence  $D(S) = D^U(S) = \text{false}$ .

In the literature, repairs usually are required to be total and, in some sense, minimal. Mostly, subset-minimality is opted for. Definition 6 does not involve any notion of minimality, although each repair in Example 7 is subset-minimal.

Unpleasant side effects of repairs such as  $U_3$  can be avoided by checking if a given partial repair is a patch with any convenient metric-based method, as expressed in the following result. It follows from Definitions 5 and 6.

**Theorem 6.**

For each  $(D, Q)$ , each partial repair  $U$  of  $(D, Q)$ , each metric  $v$  and each  $v$ -based method  $M$ ,  $U$  is a  $v$ -based patch if  $M(D, Q, U) = \text{ok}$ .

#### 4.2.2 Computing Repairs

Quality repairs can be computed by off-the-shelf update methods, defined as follows.

**Definition 7.**

An update method is an algorithm that, for each database  $D$  and each update request  $R$ , computes candidate updates  $U_1, \dots, U_n$  ( $n \geq 0$ ) such that  $D^{U_i}(R) = \text{true}$  ( $1 \leq i \leq n$ ). For a metric  $v$ , an update method  $UM$  is *quality-preserving* wrt.  $v$  if each  $U_i$  computed by  $UM$  preserves quality wrt.  $v$ .

Quality-preserving update methods can be used to compute patches and repairs wrt. any metric  $v$ , as shown in [7] for the special case of  $v = \zeta$ . Theorem 7 below generalizes that result.

Several update methods in the literature work in two phases. First, they compute a candidate update  $U$  such that  $D^U(R) = \text{true}$ . Then, they check  $U$  for consistency preservation by some TCC method. If that check is positive,  $U$  is accepted. Else,  $U$  is rejected and another candidate update, if any, is computed and checked. Hence, Theorem 7, below, follows from Definition 7 and Theorem 6.

**Theorem 7.**

For each metric  $v$ , each update method that uses  $v$ -based ITQC to check its computed candidate updates is quality-preserving wrt.  $v$ .

Example 9 shows what can go wrong if an update method that is not quality-preserving is used.

**Example 9.**

Let  $D = \{q(x) \leftarrow r(x), s(x); p(a, a)\}$ ,  $Q = \{\leftarrow p(x, x); \leftarrow p(a, y), q(y)\}$  and  $R$  the view update request to insert  $q(a)$ . To satisfy  $R$ , most update methods compute  $U = \{\text{insert } r(a); \text{insert } s(a)\}$  as a candidate update. To check if  $U$  preserves quality, most methods compute the simplification  $\leftarrow p(a, a)$  of the second constraint in  $Q$ . For avoiding a possibly expensive disk access for evaluating the simplified case  $\leftarrow p(a, a)$  of  $\leftarrow p(a, y), q(y)$ , TCC methods that are not inconsistency-tolerant may use the invalid premise that  $D(Q) = \text{true}$ , by reasoning as follows. The constraint  $\leftarrow p(x, x)$  in  $Q$  is not affected by  $U$  and subsumes  $\leftarrow p(a, a)$ . Hence,  $Q$  remains satisfied in  $D^U$ . Thus, such methods wrongly conclude that  $U$  preserves quality, since the case  $\leftarrow p(a, y), q(y)$  is satisfied in  $D$  but violated in  $D^U$ . By contrast, each ITQC method rejects  $U$ , so that  $U' = U \cup \{\text{delete } p(a, a)\}$  can be computed for satisfying  $R$ . Clearly,  $U'$  preserves quality, and even removes the violated case  $\leftarrow p(a, a)$ .

The following example illustrates a general approach of how patches and total repairs can be computed by update methods off the shelf.

**Example 10.**

Let  $S = \{\leftarrow B_1, \dots, \leftarrow B_n\}$  ( $n \geq 0$ ) be a set of cases of constraints in a quality theory  $Q$  of a database  $D$ . A quality-preserving repair of  $(D, S)$  (which is total if  $S = Q$ ) can be computed by each quality-preserving update method, simply by running the update request  $\sim vioS$ , where the distinguished predicate  $vioS$  be defined by the  $n$  clauses  $vioS \leftarrow Bi$  ( $1 \leq i \leq n$ ).

So far, we have said nothing about computing any metric that may be used in quality-preserving update methods. In fact, computing metrics  $\{\iota, |\iota|, \zeta, |\zeta|\}$  corresponds to the cost of searching SLDNF trees rooted at constraint denials, which can be exceedingly costly. The same correspondence holds for computing  $\kappa$  and  $|\kappa|$  in databases and quality theories without negation in the body of clauses. If negation may occur, the cost can even be higher. Fortunately, these metrics may not need to be computed explicitly. Instead of computing  $v(D, Q)$  and  $v(D^U, Q)$  entirely, it suffices to compute a superset approximation of the increment  $\delta(v(D, Q), v(D^U, Q))$ , as many TCC methods do, for  $v = \zeta$ . As attested by such methods, approximating the increment of inconsistency in consecutive states is significantly less costly than checking the inconsistency of entire databases. Moreover, for two quality-preserving partial repair candidates  $U, U'$  of  $Q$  in  $D$ ,  $U$  is preferable to  $U'$  if  $\delta(v(D, Q), v(D^{U'}, Q)) < \delta(v(D, Q), v(D^U, Q))$ , since  $U$  eliminates more damaged quality from  $D$  than  $U'$ .

## 5 Quality Management for Concurrent Transactions

Standard concurrency theory guarantees the preservation of quality only if each transaction, when executed in isolation, translates a consistent state into a consistent successor state. More precisely, a well-known standard result of concurrency theory says that, in a history  $H$  of concurrently executed transactions  $T_1, \dots, T_n$ , each  $T_i$  preserves integrity if  $T_i$  preserves integrity when executed non-concurrently and if  $H$  is *serializable*, i.e., the effects of the transactions in  $H$  are equivalent to the effects of a serial execution of  $\{T_1, \dots, T_n\}$  [3]. For convenience, let us capture this result by the following schematic rule:

$$\text{isolated integrity} + \text{serializability} \Rightarrow \text{concurrent integrity} \quad (*)$$

Now, if quality impairment corresponds to integrity violation, and each transaction is supposed to operate on a consistent input state, then (\*) does not guarantee that concurrently executed transactions on possibly inconsistent data would preserve quality, even if they would not decrease quality when executed in isolation and the history of their execution was serializable.

Fortunately, however, the approaches and results in Section 3 straightforwardly generalize to concurrent transactions without any effort, as shown for inconsistency-tolerant integrity checking in [8], which is based on the metric  $\zeta$ .

Theorem 8 below adapts Theorem 3 in [8] to metric-based ITQC in general. It asserts that a transaction  $T$  in a history  $H$  of concurrently executing transactions does not decrease quality if  $H$  is serializable and  $T$  preserves quality whenever it is executed in isolation. On one hand, Theorem 8 weakens Theorem 3 [8] by assuming *strict two-phase locking* (abbr. S2PL) [3], rather than abstracting away from any implementation of serializability. On the other hand, Theorem 8 generalizes Theorem 3 [8] by using an arbitrary inconsistency metric  $v$ , rather than the metric  $\zeta$ , as mentioned above. A full-fledged generalization that would not assume any particular realization of serializability is possible along the lines of [8], but would be out of proportion in this paper.

### **Theorem 8**

Let  $H$  be a S2PL history,  $v$  an inconsistency metric and  $T$  a transaction in  $H$  that uses a  $v$ -based ITQC method for checking the integrity preservation of its write operations. Further, let  $D$  be the committed state at which  $T$  begins in  $H$ , and  $D^T$  the committed state at which  $T$  ends in  $H$ . Then,  $v(D^T, Q) \leq v(D, Q)$ .

The essential difference between (\*) and Theorem 8 is that the latter is inconsistency-tolerant, the former is not. Thus, as opposed to (\*), Theorem 8 identifies useful sufficient conditions for quality preservation in the presence of damaged data. Another important difference is that the guarantees of quality preservation that (\*) can make for  $T$  require the integrity preservation of all other transactions that may happen to be executed concurrently with  $T$ . As opposed to that, Theorem 8 does away with the standard premise of (\*) that all transactions in  $H$  must preserve integrity in isolation; only  $T$  itself is required to have that property. Thus, the guarantees that Theorem 8 can make for individual transactions  $T$  are much better than those of (\*).

To outline a proof of Theorem 8, we distinguish the cases that  $T$  either terminates by aborting or by committing its write operations. If  $T$  aborts, then Theorem 8 holds vacuously, since, by definition, no aborted transaction could have any effect whatsoever on any committed state. So, we can suppose that  $T$  commits. Let  $M$  be the  $v$ -based method used by  $T$ , and  $WT$  be the write set of  $T$ , i.e.,  $WT$  is an update  $U$  such that  $D^T = D^U$ . Hence, since  $T$  commits, it follows that  $M(D, Q, WT) = \text{true}$ , since otherwise, the writes of  $T$  would violate integrity and thus  $T$  would abort. Since  $H$  is S2PL, it follows that there is an equivalent serialization  $H'$  of  $H$  that preserves the order of committed states in  $H$ . Thus,  $D$  and  $D^T$  are also the committed states at beginning and end of  $T$  in  $H'$ . Hence, Theorem 8 follows from  $M(D, Q, WT) = \text{true}$  and Definition 2, since  $H'$  is serial, i.e., non-concurrent.

It follows from Theorem 7 that, similar to ITQC, also quality repairing generalizes to S2PL concurrency if realized as described in 3.2, i.e., if ITQC is used to check candidate repairs for integrity preservation.

## **6 CONCLUSION**

Although quality is not synonymous to consistency, we have observed that constraints are expressive enough to model conditions that correspond to certain quality requirements in databases. To quantify the amount of constrain violations thus correspond to measuring an amount of quality damage. We have elaborated an axiomatization of metrics for quantifying that lack of quality. As opposed to inconsistency measures in the literature, our metrics are applicable also in databases with non-monotonic negation. Moreover, as opposed to common standards, our metric spaces are not necessarily numerical; rather, any partial ordering is acceptable, in general. The metric spaces provided by sets of violations of constraints, or of instances of constraints, or of the causes of such violations, are special instances of our generic concept of inconsistency metrics. Metrics that range over such spaces allow to check and accept each update if it does not increase the measured lack of quality. Similarly, each repair is acceptable if it decreases the measured inconsistency. Since quality is not a binary property such as the satisfaction or violation of constraints, but may be compromised to varying degrees, the inconsistency tolerance of metric-based constraint checking and repairing is particularly welcome for quality maintenance. Future work includes the use of metric-based quality maintenance in replicated databases, and the use of inconsistency metrics for providing quality answers to queries in databases.

## ACKNOWLEDGEMENT

The author has been supported by FEDER and the CICYT grants TIN2009-14460-C03, TIN2010-17139.

## REFERENCES

- [1] Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley, 1995.
- [2] Bauer, H.: Maß- und Integrationstheorie, 2<sup>nd</sup> edition. De Gruyter, 1992.
- [3] Bernstein, P., Hadzilacos, V., Goodman, N.: Concurrency Control and Recovery in Database Systems. Addison-Wesley, 1987.
- [4] Ceri, S., G. Gottlob, G., Tanca, L.: What you always wanted to know about Datalog (and never dared to ask). TKDE 1(1):146-166, 1989.
- [5] Clark, K.: Negation as Failure. In H. Gallaire, J. Minker (eds): Logic and Data Bases, pp. 293{322. Plenum Press, 1978.
- [6] Decker, H., Martinenghi, D: Modeling, Measuring and Monitoring the Quality of Information. Proc. 28th ER Workshops, Workshop QOIS, edited by I. Comyn-Wattiau and B. Thalheim, Springer LNCS vol. 5833, pp. 212–221, 2009.
- [7] Decker, H, Martinenghi, D.: Inconsistency-tolerant Integrity Checking. IEEE TKDE 23(2):218-234, 2011.
- [8] Decker, H., Muñoz-Escóí, F.: Revisiting and Improving a Result on Integrity Preservation by Concurrent Transactions. Proc. OTM Workshops. Springer LNCS vol. 6428, pp. 297-306, 2010.
- [9] Decker, H.: Causes of the Violation of Integrity Constraints for Supporting the Quality of Databases. Proceedings 9th ICCSA, Part V, Springer LNCS vol. 6786, pp. 283-292, 2011.
- [10] Enderton, H.: A Mathematical Introduction to Logic, 2<sup>nd</sup> edition. Academic Press, 2001.
- [11] Hunter, A., Konieczny, S.: Approaches to Measuring Inconsistent Information. In L. Bertossi et al (editors), Inconsistency Tolerance. Springer LNCS, vol 3300, pp. 191-236, 2005.

# **INTRODUCING DATA AND INFORMATION QUALITY PRINCIPLES IN TODAY'S COLLEGE CURRICULUM VIA AN INTRODUCTORY PROBABILITY AND STATISTICS COURSE**

(Research-in-Progress)

**William Rybolt**  
Babson College  
[rybolt@babson.edu](mailto:rybolt@babson.edu)

**Leo Pipino**  
University of Massachusetts Lowell  
[leo\\_pipino@uml.edu](mailto:leo_pipino@uml.edu)

**Abstract:** This paper suggests a method of introducing some key concepts associated with data quality into the college curriculum, specifically, into the undergraduate introduction to probability and statistics course. The emphasis is not primarily to educate the students on how to solve the problems of poor data quality; rather, it is to sensitize the students to the importance of data quality, to data quality issues and to make students aware that data quality is a variable associated with any data set. Not all data sets are of good quality, and one must be sensitive to the problems caused by data of poor quality and be knowledgeable about some of the factors and root causes that lead to poor data quality.

**Keywords:** Data and Information Quality, Root Causes, Statistics, Data Cleansing, IQ Curriculum

## **INTRODUCTION AND BACKGROUND**

The importance of data and information of quality is acknowledged in all disciplines from the natural sciences to business and the social sciences. Formally teaching and sensitizing students to the principles of data quality and developing awareness of the importance of data quality, however, has proven difficult to initiate and sustain. There has been some discussion in the academic literature on how and where this might be accomplished (Kahlil et al. 1999, Lee et al. 2007). One observes few implementations of data quality courses and data quality curriculum. An example of an exception is the Information Quality Graduate Program at the University of Arkansas Little Rock. This situation should not be unexpected given the competition for space in a constrained curriculum. There is always more that should be taught and tradeoffs must be made. This, of course, is not a problem solely confronted by the data quality discipline. It is a problem continuously confronted across disciplines.

An analogous situation exists in business and governmental organizations. Everyone pays homage to the need for data and information quality but, typically, substantive resource support is limited. Too often, only when a crisis attributable to poor information occurs does management begin to support initiatives. And these sometimes are short lived.

There is an explicit assumption motivating this proposal. In general, there is not going to be a data quality course that will be taken by a large number of college students. The best hope is to find one or more existing courses that can benefit from the inclusion of data quality concepts. The question addressed in this research-in-progress paper is how we might introduce data quality principles in courses that would make the students aware of the importance of data quality and also permanently sensitize them to these principles. The hope is that the student will carry these notions and ideas forward. This added perspec-

tive will be manifested both in their future decisions when using data and their future decisions regarding the support of data quality initiatives.

Our focus in this paper is on the introductory course in probability and statistics. Our suggestion is that a number of data quality concepts can be introduced in such a course and be of benefit. Perhaps, other introductory courses, such as the introduction to MIS course can also be a vehicle to promote the principles of data quality. Although some data quality concepts may be more easily introduced in such a course, our main focus is the introductory probability and statistics course for students in a School of Management or College of Business. Of course our approach could be used in any basic probability and statistics course.

This paper and its paradigms for introducing data quality into the college curriculum can best be understood in the context in which they evolved. The origin was a desire to find a simple classroom exercise to be used on the first day of an applied introduction to probability and statistics course. The assignment required students to work in groups of three to estimate the average SAT scores of a group of five thousand students and to complete the task in twenty minutes. This required the students to take a sample from the population, enter the values into Minitab on their laptops, and calculate the average. To require a modicum of thought some of the values, 0 and 8000, made no sense since SAT scores range from 200 to 800. It was assumed that students would realize that these values should not be included in the average. Instead many students simply included these anomalous values in their averages. When discussing the exercise, the need to examine the quality of the data before using it to make decisions was mentioned. A similar question was included on the first examination with the expectation that all students would get it correct. This was not the case. From this, evolved the realization of the need for introducing data quality concepts into the course and the opportunities this presented.

Unfortunately, all of the data sets included in the introductory text books are perfect. The topic of cleansing data is never mentioned. Students are not presented with dirty data sets that require cleansing. Students are implicitly led to believe that all data sets are perfect. The concept of data cleansing, is not mentioned in introductory statistics courses. This is especially puzzling in that it has been reported by many authors and practitioners that a majority of analysis time is allocated to cleansing, preparing, and organizing data for processing rather than conducting analysis.

In an introduction to probability and statistics course students are usually exposed to methods that can be used to ensure that the sample is a random sample. Although sampling techniques affect the quality of the data, students are not exposed to data sets in which the poor choice of sampling techniques affects the quality of the data and hence the quality of the decisions that result from using the statistical techniques presented in the course.

In a similar manner, students are exposed briefly to the types of errors that can occur when taking a sample from a larger population. These errors are coverage, measurement, non-response, and sampling errors. Virtually the entire focus of the course is on sampling errors and how to use statistical techniques to calculate and understand the implications of the sampling errors. None of the data sets in the books contain the examples of the other three types of errors.

The unintended lesson that students are learning is that the secret to understanding any data is to find and use the right statistical tools. The closest that the textbooks come to speaking about data quality is when they discuss the influence of extreme points or the criteria that data sets must meet for the use of a particular statistical technique. For example, we are told that in order for linear regression to be valid the residuals must be independent, be normally distributed, and have equal variance. The implicit assumption is that if the residuals meet this criterion all is well. The fact remains that given data sets of poor

quality; no amount of statistical analysis will result in better decisions than would be made if data sets of higher and better quality were used.

We believe that this neglect of the issue of data quality is a short sighted approach and it is in the best interest of both the data quality and the statistical community to begin to introduce data quality concepts into introductory probability and statistics courses. Accordingly, this paper suggests a series of simple steps that can be incorporated into statistical courses that can begin to expose students to data quality issues and some of their root causes.

To do this and to systematize the approach, we have chosen to make use of the set of root causes of data quality first enumerated by Strong et al. (1997a, 1997b) and further elaborated in Lee et al. (2006). This is not the only approach or framework that can be adopted, but it serves our purposes in that (1) it provides a guide for developing future data sets, (2) it is useful to classify student errors (perceptions and misperceptions of quality of the data) in data that has been collected over the past five or six years, and (3) it is anchored in the data quality literature.

Recall that the 10 root causes enunciated in the above works were:

1. Multiple Data Sources
2. Subjective Judgment in Data Production
3. Limited Computing Resources
4. Security Accessibility Trade-off
5. Coded Data across Disciplines
6. Complex Data Representations
7. Volume of Data
8. Input Rules Too Restrictive or Bypassed
9. Changing Data Needs
10. Distributed Heterogeneous Systems

The key idea is to construct a series of simple exercises that could be incorporated into a probability and statistics course such that each exercise would illustrate one or more of the ten root conditions. When doing these exercises, students would be implicitly learning that data quality needs consideration and it would become part of their thinking process. The choice of a particular classification scheme is secondary to the need to expose students to data lacking in quality.

A realistic expectation is that an instructor might at first find time to incorporate one or two such exercises into his or her curriculum. As instructors come to appreciate the negative implications of only including data sets of perfect quality into their teaching and the disservice this is to the students, they will begin to address the data quality issue more openly.

We recognize that the concept of introducing data quality principles in advanced courses is desirable, should be explored, and will be developed. Further, we recognize that more advanced problems and cases will be necessary in these advanced courses. As a first step in this early research-in-progress, however, we exploit the course that we consider will offer the least resistance to implementing the introduction of data quality principles. Later in the paper we briefly address the topic of more advance courses.

In the sections that follow we describe what we have done and how our approach has evolved over the last few years. We discuss each of the 10 root causes in the context of our proposal. Since this is research in progress, we only present details associated with a subset of the ten root causes above. The remainders of the root causes are mentioned at a conceptual level.

## **THE SUGGESTED APPROACH**

Each semester in our introduction to probability and statistics course, we construct a data survey of per-

haps 40 items which we ask the students to take. The students identify themselves when they take the survey so that they can be given credit for participation. Their names and any responses that would identify a specific individual are removed before a random sample is taken.

Initially, this was intended to obtain real data sets to which students could relate and which would be used in the course. After data collection, we always tried to cleanse the data set to reduce obvious data quality errors to produce a data set with no data quality issues.

It became increasingly apparent that retaining these errors in the data sets would be useful in imparting to the students how poor quality data can affect statistical results and lead to adverse decisions; these exercises could be used to impart some of the basic principles of data quality and permanently sensitize the students to the importance of data quality. This perspective was reinforced by the observation, pointed out earlier, that the data sets in introductory textbooks were almost always flawless data sets.

As an aside, we mention that the proposal that these imperfect or dirty data sets should be used or that clean data sets should be corrupted for use has been met with resistance from many instructors. We will not delve into this debate in this paper. For the moment, however, we point out the observations made in Lang's article "The Benefit of Making It Harder to Learn" (Lang 2012) which cites sources that assert that "making material harder to learn" can "improve long-term learning and retention."

### ***Multiple Data Sources***

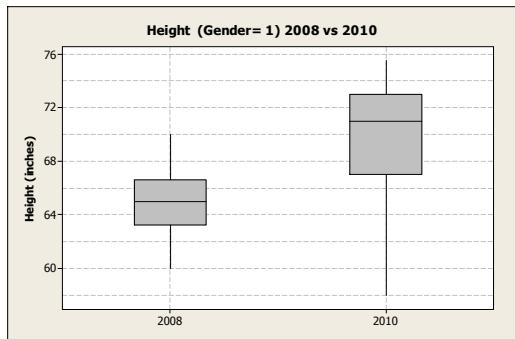
Using data obtained in over five years of surveys, one way to illustrate the effects of multiple data sources is as follows. Ask the students to compare the heights of male and female students from a previous year with the present. The students are given access to the questions used in any given year and the information on how the data was coded. A simple change in the data sets would be to code the females as 1 and the males as 2 in 2008, but to change this so as to code females as 2 and the males as 1 in 2010.

This is of course the key point: we want students to think. We do not want them to assume that all data is coded in exactly the same manner so that they get in the habit of doing analysis without thinking.

Now suppose you are illustrating the concept of a straight forward hypothesis test: is there a difference in heights between female students in 2008 and 2010 using a two sample t-test. Using the two data sets, you would obtain the following results: the 95% Confidence Interval for the difference is (-7.09, -2.56), and the value of the test statistic t is -4.32 with a p-value of 0.000.

Without thinking students would conclude that there is evidence that the average height of females in 2010 was obviously different from the average height in 2008. Most students are not sensitive to the data quality issues associated with multiple data sources. They do not question the underlying data, and, as a consequence, would conclude that females had grown significantly taller in just two years. Or they might conclude that there was a change in the population from which the female students were coming. Unfortunately, experience shows that not all students are sensitive to the nonsensical conclusion and simply report the result that females are growing taller. The goal would be to make the student realize that this made no sense. By seeking to learn the cause of this strange result, they would be made more sensitive to the data quality issues caused by multiple data sources.

Also, it is important that the results be presented visually. A Box plot of this data is given in Figure I. The Boxplot makes the strange conclusion much more obvious. This reinforces the value of visual tools for understanding data and the implications of the analysis.

**Figure 8**

A key to making this example work is that we are using data with which the students have experience. Students, *if they think about it*, have a basic understanding of the height of females and should be suspicious of the data. In contrast, if the variable were something like the time per day spent emailing, then no one would bother to challenge the quality of the data in use. It would make perfect sense that females in 2010 spend more time emailing than in 2008. This choice of variable is in effect a form of an error correcting code.

The marginal time to introduce this data quality concept into a traditional course is quite minimal, perhaps minutes not hours. We believe that this is an example of how a minor change can dramatically improve the educational value of this exercise. The exercise needs to be structured so that the students realize, on a common sense level, that the results make no sense. This ultimately leads to an examination of the nature of the data underlying the analysis.

### ***Subjective Judgment in Data Production***

A number of the questions on our data collection survey ask the students to estimate the amount of time they spend in various activities. How many hours of sleep did you get last night? During the previous semester, how many hours per week did you average viewing television? During the previous semester, how many hours per week did you average studying outside of class? How many minutes per day do you spend surfing the Web, reading on the Web, sending email, interaction on Facebook, Twitter and the like? We intentionally used two sets of units to measure the amount of time spent in different activities. Sleep is measured in hours per day and studying is measured in hours per week.

In the statistics course, then, a typical question might be how does the amount of time spent studying compare with the amount of time spent sleeping? The typical student will be focused on what type of statistical test to run before deciding that a paired t-test was needed. They would typically be given a significance level of 0.05. After doing the analysis, they would report for one of our datasets that the difference was 5.36, the t-value was 5.41, and the p-value was 0.000. They would then conclude that there was evidence that students spend more time studying than sleeping. They would judge that this made sense before hurrying to the next questions. This is not to minimize the importance of choosing the correct statistical test. It is, however, important that the proper statistical test be applied to data that is fit for use for the test.

After we correct for the different units, we find student spend about 49 hours per week sleeping and 15 hours per week studying outside of class. If you examine the data sets found in common text books you would find that for this type of question there is no need to think about the units or converting from one set of units to another before do a statistical analysis.

The subjective judgment in data production enters our data survey in a more fundamental way. What if we compare the total amount of time students spend in all communication and media actives with their time spent sleeping? The obvious approach would be to add email, web surfing, Facebook, Twitter,

television, and the like. Is it correct to add the times for the different activities to arrive at the total time? Students tend to multitask and do several activities at the same time. A problem may occur depending on how the total time was obtained. If one group is asked to estimate the total time as one number and another group is asked to estimate separate times for each activity, which are then added to obtain total time, then the two total times may not be valid for purposes of comparison. Thus, the subjective judgments associated with the data color its quality and the validity of any conclusions.

### ***Limited Computing Resources***

Thus far we have presented several examples of how students might be exposed to data that is less than perfect. In this section we will show that data quality concepts can be used by instructors in the opposite sense. We illustrate how instruction can benefit by improving the data quality associated with instruction. Instead of mentioning how limitations in computing resources can inhibit data quality, we focus on the how the improvements in computing resources can be used to improve education by focusing on improving data quality.

Within this category, we examine the information quality of the presentation and not the data itself. Although computer resources are limited, they are far less limited than in the past. It is useful to look at the implications of the vast increases in speed, accuracy, storage capacity during the last seventy years. As technology evolves, the representation, implementation, and presentation of the algorithms and paradigms for manipulation data also need to evolve. We can speak of the information quality of algorithms and paradigms. The ideal is to choose the algorithms and paradigms that have the highest information quality at a given point in time in the context of available technology. The key is to expose students to the highest quality paradigms while making them more aware of the quality of the data they are analyzing.

Consider the specific example of the algorithms used to calculate the slope of a straight line passing through a set of  $n$  values and how it is treated in a statistics course. In the table below (Table 1) are three different equations for calculating the slope of the best fit line through those points. All three are mathematically equivalent. There are, however, subtle but important differences when used in an introductory course. The first, the conceptual formula, is used to present the idea, while the second and third are minor variations of computational formulae used by different authors.

The virtue of the computational formulae is that they require only a single pass through the data to calculate the sums which can then be combined using the computational formulae. This results in faster computation and more accurate results. The conceptual formula requires two passes through the data. The first pass calculates the mean values of  $x$  and  $y$ , and the second pass calculates sums that are then combined using the conceptual formula.

With the transition from single to double precision arithmetic and from reading data from paper tape and punch cards to accessing data in high speed RAM memory, computational considerations are not as critical as in the past. Since the computational formulae yield no advantage in today's technological environment for typical datasets, they should play no role in the delivery of today's courses. This is true of many other formulae found in textbooks. Ultimately, the question becomes "which representation has the highest information quality according to empirical evidence?" That is the one which should be used for instruction. For practical reasons one would not switch to achieve only minor improvements in quality.

We have done several exploratory experiments in our introductory quantitative courses; one experiment used calculators, another Excel, and the third Minitab. In each case students were randomly assigned to one of two groups. Each group had the tasks calculating the slope of the line through a small data set.

One group used the conceptual formula and the other a computational formula. Surprisingly the conceptual group, in general, had a higher percentage of correct results. The times for the two groups differed by less than a minute. Although, there was no evidence that the computational formulae were better than then conceptual formula (Forthcoming, 2012).

The conclusion is that there are absolutely no reasons to introduce the computational formulae into the course. Eliminating these would reduce the number of formulae the students see and lessen their confusion about the role of the different formulae. Although we used slope as an example, we believe these same conclusions hold for many other formulae. We are undertaking an expanded study for future publication. Data quality should be applied to instructional algorithms and paradigms as well as conventional data sets.

Slope	Source
$\frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$	Conceptual (Berenson)
$\frac{(\sum xy) - (\sum x)(\sum y)/n}{(\sum x^2) - (\sum x)^2/n}$	Computational (Anderson)
$\frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$	Computational (Hughes)

Table 1. Comparison of Conceptual and Computational Formulae

### ***Security/Accessibility Trade-off***

When designing and collecting data there is a tradeoff between accessibility and restricting access to the target audience and the information collected to an acceptable range of values. We allowed anyone who knew the URL to take our surveys, but required that they identify themselves. Our problem was how to treat multiple responses by the same person. We had a classroom exercise in which treating the multiple responses differently led to different answers to questions raised. Do you include all responses, none, the first, the last, or an average of the multiple responses?

Exercises of this type make students sensitive to this issue. While it is tempting to restrict the entry process to only one entry, this may result in partially completed forms. To counter this one may require that the survey be completed before submission. This may lead to thoughtless entries in some of the questions. Attempts to restrict the entries to reasonable values make sense, but this may lead to frustration when attempting to enter information for unusual situations. Since incomplete forms cannot be submitted, the students quits and no information is collected.

We believe that it is of value to have data sets for student use that make explicit these types of issues. An examination of these issues can be a minor component of almost any data set. For example after we have collected about 150 survey responses, we spend a day or so cleaning the data to make it appropriate for class room use. As we cleanse the data, we allow certain of the data quality issues discussed in the paper to remain. This provides a data set what we believe to be more valuable than a perfect one with no data quality issues.

### ***Coded Data across Disciplines***

The supply and demand curves in economics are an example of where the data is coded and presented very differently. Economists place the price on the vertical axis and the quantity on the horizontal axis, whereas mathematicians tend to place price on the horizontal axis and quantity on the vertical axis. When given a set of data and a graph in one discipline, students are often confused as to how to translate the data into the representation common in the other discipline. This is especially true when the graphs are labeled with x and y with the meaning of x and y defined in text describing the problems.

We have discovered that students often do not take the time to read the details. They have a propensity to skim the information and assume that they are given exactly the information they need to solve the problem no more and no less. The assumption is the information is structured in exactly the optimum way so as to solve the problem. These habits become obvious when we started inserting phrases like “if you read this draw a circle around ..” or “draw a box around ...”. It was surprising how many students failed to follow these instructions. When one of the students was asked why, “He said when he was studying for the SAT exams, they were told to ignore the instructions.” A simple exercise might be to give the students the same data presented differently. A discussion of their results can then be used to explore this issue.

As a specific example, the convention in the USA of using a period to denote the decimal point and commas to group digits into multiples of a thousand is not universal. These conventions are reversed in many countries of the world. Similarly “month/day/year” becomes “day/month year” in other parts of the world. A good example of a data set designed to sensitize students to this issue would be financial records. The data set should contain hypothetical financial information for companies in different parts of the world. Before students did a naïve analysis, the goal would be to have the students aware of the different standards and the need to convert to a common representation before doing the comparison. Failing to do so would alert the students to the need to understand that the conventions for data representations are not universal and care must be used to address this data quality issue before using the data in the decision making process.

### ***Complex Data Representations***

A good example of the problems presented by complex data representations is exhibited by questions which require students to enter text. Text is difficult to classify and process. It typically requires human intervention to translate into a canonical representation. Examples include: favorite television shows, your goals for the course, interesting facts about yourself. This is an area which we are beginning to explore.

### ***Volume of Data***

There is a vast amount of data available for analysis. Historically statistical analysis grew up in an era where the collection of data was very resource intensive. Thus, the need for techniques to collect samples which were faithful random representations of the underlying population. Today the situation is almost the opposite. It is common to collect data sets so large that there are difficulties in communicating, storing and analyzing that information. For example some weather satellites produce more than a terabyte of data every day, day after day. It is literally necessary to buy a new hard disk every day just to store the data.

The RITA, Research and Innovative Technology Administration of the Bureau of Transportation Statistics has a web site containing Airline On-Time Statistics available from January 1995 through April 2012 for all domestic scheduled-service for US air carriers that have at least 1 percent of the passenger revenue. Information available includes departure and arrival statistics (scheduled departure time, actual departure time, scheduled elapse time, departure delay, wheels-off time and taxi-out

time) by airport and airline; airborne time, cancellation and diversion by airport and airline. However, we are warned “Due to the large amount of data to be searched, time period should be limited to a maximum total of 31 days for any combination of Month, Date and Year.” (RITA, Research and Innovative Technology Administration, 2012)

Our computer and communication resources limit our ability to make full use of freely available information. This implies that the quality of the data we use for the analysis may not be representative of the totality of the data even if the underlying data is perfect. To use this in a course to illustrate the implications of vast amount of data we suggest giving an assignment using the RITA data. The point would be that the problem could not be solved because of the size of the data that would be needed to do a complete solution involving the population. The solution would of course be to take an appropriate sample from the data. This brings us full circle to the reason that statistical methods were developed in the first place. Originally, it was too complicated and resource intensive to acquire all the data. Now it is too easy to acquire all the data so we need to use sampling to reduce the amount of data we are dealing with.

### ***Input Rules Too Restrictive or Bypassed***

When constructing an instrument to collect data, there is a tradeoff between restricting the values entered according to certain well defined rules, and allowing what might appear to be completely unreasonable values. As an example, we once had a question how many computers do you have in your dorm room. The idea was to learn if students had both a laptop and a desktop computer in their rooms or only the laptops issued by the school. The temptation was to dismiss an answer of seven and code it as missing. Discussion with the student revealed, he was running a server farm in his dorm as a business and seven was a valid response. Unusual values can provide interesting insights.

The way we recommend to address this issue, is through variables such as the SAT scores that are used as part of the college admission process in the USA. When students enter their SAT scores in a survey, we do not restrict the values to the actual range of 200 to 800. As a group exercise for an exam question, the students are given data sets with values well outside of this range, for example zero and eight thousand. Whenever they are asked to calculate the mean SAT score, it is quite common for students to simplify average all the data and report the result. Even after we had a classroom exercise and explained the need to code these values as missing data before doing any further, we still find that students will make the same error on exams. Because of their experience with the textbook, they assume all data is valid and do not consider that there may be quality issues with the data.

Other variables we have used this approach with include: desired temperature, weight of airline luggage on the last trip, height of student, and height of parent. While most students enter values assuming units of Fahrenheit, pounds, and inches; others use Celsius, kilograms, and centimeters. Again, given a data set with mixed units students do not take the time to think about quality of data issues, they simply calculate with all the numbers. When the mixed units are far apart, it is eventually easy to separate the values and correct for the disparity in units. For example, ideal temperatures fall into two groups clustered around 20 and 68, and it is easy for students to eventually spot the disparity by using a histogram. This technique does not work well for the airline baggage question, but does work for height.

### ***Changing data needs***

When we first started, a number of years ago, collecting data from the students to use in the course, there was no Facebook, Twitter, or blogging. Originally we assumed that students were either surfing the web or using email. During recent semesters we have added questions about these three activities, but we did not catch their emergence as quickly as we would have liked. An analysis of our longitudinal data would thus give a misleading history of their development. To make issues such as this more obvious one might

pose the question “how has as the student use of the internet changed over the last five years?” If students are given only the usage times for email and surfing the web they might mistakenly conclude that it was being used less. Hopefully the students would realize that something was missing. The key would to realize that the data sets did not include all relevant variables.

### **Distributed Heterogeneous Systems**

When one moves data from one environment to another environment, unexpected changes may take place. Table 2 below illustrates some of the differences that may take place when the same text data is entered into an Excel spreadsheet. Someone who is not aware of these possibilities and does not consider the quality of the data will encounter some strange results. As an exercise, students might be given a text file containing a column of values representing the fraction of the time spent on various parts of a project, and asked to do a an analysis. If several of the values are fractions, then the types of conversions displayed in Table 2 may occur. This type of exercise requires a little more thought than the standard perfect data set exercise. We believe, however, it provides added value at a minor cost of time.

	<b>Excel</b>	<b>Excel</b>	<b>Excel</b>
<b>Text</b>	<b>General</b>	<b>Number</b>	<b>Fraction</b>
3 to 5	3 to 5	3 to 5	3 to 5
3-5	5-Mar	40973.00	5-Mar
3:5	3:05	0.13	3:05
3/5	5-Mar	40973.00	3/5
3 5	3 5	3 5	8
more than 3	more than 3	more than 3	
3 1/2	3 ½	3.50	3 1/2
5/7	7-May	41036.00	5/7
none	None	none	none
34.5	34.5	34.50	34 1/2
=====	=====	=====	=====
0	123020 1/8	123020 1/8	41020 4/9

Table 2. Effect of Formatting on Data in Excel

## **CONCLUSIONS**

The statistical and the data quality communities need to work together to make the concept of data and information quality part of the topics in the introduction to probability and statistics courses at both high school and college levels. There will be initial resistance since many instructors feel that there are already too many topics being covered. The examples in this paper are designed to show how by making modest changes in data sets and their production, that quality topics emerge as an intrinsic part of the curriculum. Instead of this decreasing the students understanding of the primary topics, this change will enhance their understanding. The analysis and interpretation will become less routine and require more thought.

We do not hold the unrealistic belief that all the principles and issues will be used in a course. The hope is that instructors find at least one or two that they can embrace. A reasonable goal is to expect that perhaps ten percent of the data sets found in the textbooks and used in the courses have some type of data quality issues. The same goes for exam and homework problems. The data qualities issues will likely

require only minor modifications in the data used and its presentation. These issues need only be one part of the problems.

These data quality issues will be most helpful if they are embedded in data sets that the students can relate to and have some experience with. When the results of ignoring the data quality issues lead to situations that make no sense, then students are forced to ask “what happened?” This will not occur if the data sets involve variables and values that students have no feel for.

The next step in our research will be a variant of the capture-tag-release approach used in ecological studies of animal populations. We will begin with a clean data set for which the students have an intuitive feel. Next, we will use a computer program to randomly modify a small portion of the data set by using the issues mentioned in this paper. After the students use the modified data of reduced quality, we will determine which data quality issues they were able to find and resolve as part of their analysis. This will allow us to continue to sensitize students to the need to be aware of data quality issues.

The important aspect is that students be disabused of the experience that all the data they use is of perfect quality, and all they need to do is find an appropriate statistical technique to solve the problem. The approach they need to become accustomed to is to first ask, “Are there quality issues with this data that may impact my analysis and conclusions?”. These issues will need to be addressed either before or during the analysis, but definitely before drawing any meaningful conclusions. Our belief is that not only is this possible using the techniques presented in this paper, but that this approach will result in a better and more lively course for both students and instructors.

The primary focus of this paper was on data and data sets and their use in an introductory probability and statistics course. If, however, this approach were extended to other courses, such as an introductory MIS course as well as advanced courses, the students could be further instructed in and sensitized to the principles and issues of data and information quality. Over time this would manifest itself in the student’s future decisions regarding data quality issues and data quality initiatives.

## BIBLIOGRAPHY

- [1] Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2004). *Quantitative Methods for Business* (9e ed.). Mason, OH: South-Western.
- [2] Berenson, M. L., Levine, D. M., & Krehbiel, T. C. (2006). *Basic Business Statistics Concepts and Applications* (10th ed.). Upper Saddle River, NJ, USA: Pearson Education.
- [3] Berenson, M. L., Levine, D. M., & Krehbiel, T. C. (2009). *Basic Business Statistics Concepts and Applications* (11th ed.). Upper Saddle River, NJ: Pearson Education, Inc.
- [4] Garfield, J. (2003). *Guidelines for Assessment and Instruction in Statistical Education (GAISE Report)*. American Statistical Association.
- [5] Forthcoming, Title of Article, Name of Journal, 2012
- [6] Kellar, G. (2012). *Statistics for Management and Economics* (9 ed.). Mason, OH: South-Western Cengage Learning.
- [7] Khalil, O., Strong, D., Kahn, B., and Pipino, L. Teaching Information Quality in Information Systems Undergraduate Education. *Informing Science*, Volume 2, No. 3, 1999.
- [8] Lang, J. M. (2012, June 3). *The Benefits of Making It Harder to Learn*. Retrieved June 13, 2012, from chronicle.com/article: <http://chronicle.com/article/The-Benefits-of-Making-It/132056/>
- [9] Larose, D. T. (2010). *Discovering Statistics*. New York, NY: W. H. Freeman and Company.
- [10] Lee, Y.W., Pierce, E., Talburt, J., Wang, R.Y., and Zhu, H. "A Curriculum for a Master of Science in Information Quality". *Journal of Information Systems Education*, Volume 18, No. 2, 2007.

- [11] Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to Data Quality*. Cambridge, MA: MIT Press.
- [12] Minitab, <http://www.minitab.com>, Minitab16.
- [13] RITA, Research and Innovative Technology Administration. (2012, June 26). *ontimesummarystatistics*. Retrieved June 26, 2012, from www.bts.gov: <http://www.bts.gov/xml/ontimesummarystatistics/src/index.xml>
- [14] Forthcoming, "Conceptual versus Computational Formulae in Calculus and Statistics Courses". *The International Journal of Technology, Knowledge and Society*, NA.
- [15] Strong, D., Lee, Y., and Wang, R., "Data Quality in Context". *Communications of the ACM*, Volume 40, No. 5, May 1997, 103-110.
- [16] Strong, D., Lee, Y., and Wang, R., "Ten Potholes in the Road to Information Quality". *IEEE Computer*, Volume 30, No. 8, August 1997, 38-46.
- [17] University of Arkansas Little Rock, Information Quality Graduate Program, <http://ualr.edu/informationquality/>
- [18] Weiers, R. W. (1998). *Introduction to Business Statistics* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

# **TOWARDS THE USE OF MODEL CHECKING FOR PERFORMING DATA CONSISTENCY EVALUATION AND CLEANSING**

(Completed-paper)

**Mario Mezzananza**

Department of Statistics and Quantitative Methods – CRISP Research Centre – University of Milano-Bicocca

[Mario.Mezzananza@unimib.it](mailto:Mario.Mezzananza@unimib.it)

**Mirko Cesarini**

Department of Statistics and Quantitative Methods – CRISP Research Centre – University of Milano-Bicocca

[Mirko.Cesarini@unimib.it](mailto:Mirko.Cesarini@unimib.it)

**Fabio Mercurio**

CRISP Research Centre – University of Milano-Bicocca

[Fabio.Mercurio@unimib.it](mailto:Fabio.Mercurio@unimib.it)

**Roberto Boselli**

Department of Statistics and Quantitative Methods – CRISP Research Centre – University of Milano-Bicocca

[Roberto.Boselli@unimib.it](mailto:Roberto.Boselli@unimib.it)

**Abstract:** This paper explores the application of formal methods (specifically, model checking) to the field of data quality. A model expressing the consistency of longitudinal data is derived from the domain knowledge. This model is used (1) to automatically verify the consistency of the data stored on a database and (2) to automatically generate a universal cleanser, i.e. a cleanser which summarises all the feasible corrections for any kind of inconsistency which may affect the data (as far as they can be guessed from the formal consistency model). The universal cleanser represents a repository of corrective interventions useful to develop cleansing routines. We applied our approach to a real world scenario: a formal verification has been performed on labour market data evaluating the consistency of people working careers. The results show that the proposed approach can improve the data quality evaluation and the development of cleansing activities.

**Key Words:** Data Consistency, Data Cleansing, Model Checking.

## **1 INTRODUCTION AND CONTRIBUTION**

The ongoing relations between citizens and public administrations generate a lot of data and the administrative archives store a relevant portion thereof. Such data can be very valuable for supporting the decision making processes in several contexts: design, implementation, and evaluation of active policies, service design and improvement, etc. Some archives record also data along time, therefore they can be considered a source of longitudinal data (also called panel data), i.e. a set of (repeated) observations of

the same subjects along the time. For more details on longitudinal data see [14]. Several studies report that the *data quality* of enterprise and public administration databases is very low, e.g. [12, 2]. The organisations are getting more and more aware of the consequences and costs, therefore several plans, strategies, and actions have been implemented, e.g. as described in [15]. *Data quality* is a broad concept (a complete survey can be found in [2]). Here we focus on the *consistency* dimension which refers to the violation of semantic rules defined over a set of data items.

In this paper, a data consistency model is built from the domain knowledge, then a model checker can be used for verifying the consistency of longitudinal data and for generating the possible cleansing actions. An example is provided: the dataset in Tab. 1 shows a cruise ship travel plan. The ship usually travels by sea and stops at the port of calls (intermediate destinations), making a *checkin* when entering a harbour and a *checkout* when exiting. The reader will notice that the departure date from Lisbon is missing, since a *checkout* is necessary before entering the subsequent harbour (Barcelona). In this respect, the dataset is inconsistent.

<b>EventId</b>	<b>ShipID</b>	<b>City</b>	<b>Date</b>	<b>Event Type</b>
e <sub>1</sub>	S01	Venice	12th April 2011	checkin
e <sub>2</sub>	S01	Venice	15st April 2011	checkout
e <sub>3</sub>	S01	Lisbon	30th April 2011	checkin
e <sub>4</sub>	S01	Barcelona	5th May 2011	checkin
e <sub>5</sub>	S01	Barcelona	8nd May 2011	checkout
...	...	...	...	...

**Table 1: Travel Plan of a Cruise Ship**

Data cleansing can be performed in several ways, nevertheless when no different (and more trusted) data source is available, the only feasible solution is to exploit business rules, i.e. to implement cleansing algorithms fixing inconsistencies using domain derived knowledge. The uncertainty affecting the data can impact on the aggregate data and on the information derived for decision making purposes, therefore the inconsistencies should be appropriately managed.

The comparison among archive contents and real data is often an unfeasible or very expensive option (e.g. due to the lack of alternative data sources, the cost of collecting the real data, etc.). On the contrary data assessment and cleansing based on business rules is frequently an effective and valuable solution.

In this paper we show how longitudinal data consistency can be modelled and verified through explicit model checking techniques. Once a model has been defined, a model checker can be used for deriving the set of possible errors and the set of possible corrective actions. These can be exploited: (1) for verifying the data consistency of real world archives and (2) as a foundation to partially automate the development of cleansing routines. It is worth to note that the approach presented in this paper bounds the effort of consistency checking to the formalisation of a suitable consistency model. Then, the task of performing the consistency check and the cleansing activities can be automatically executed.

We successfully applied model-checking-based techniques to assess the quality of an administrative archive.

The paper is organised as follows: in Sec. 2 the related works are surveyed; in Sec. 3 we shortly introduce model checking on finite state systems and how model checking can be used for verifying data consistency; Sec. 4 introduces the concept of the universal cleanser and provides an algorithm to compute it; in Sec. 5 we show some experimental results obtained working on a big administrative archive managing labour market information; finally, in Sec. 6 we report the conclusions and the future work.

## 2 RELATED WORK

Data quality has been addressed in different research domains including statistics, management, and computer science as reported in [27, 4]. For the sake of clarity, the works surveyed in this section have been classified into three groups according to the (main) goal pursued: *record linkage*, *error localisation and correction*, and *consistent query answering*. The classification adopted is not strict since several works could be classified in several groups.

**Record linkage** (known as *object identification*, *record matching*, *merge-purge problem*) aims to bring together corresponding records from two or more data sources or finding duplicates within the same one. The record linkage problem falls outside the scope of this paper, therefore it is not further investigated.

**Error localisation and correction** works can be further classified in: 1) those exploiting machine learning methods and 2) those exploiting data dependencies (formalised by domain experts) to detect and correct errors. Considering the latter, the effort of domain experts is required to formalise rules.

1) *Machine learning methods* can be used for error localisation and correction. Possible techniques and approaches are: unsupervised learning, statistical methods, data profiling, range and threshold checking, pattern recognition, clustering methodologies [23]. It is well known that these methods can improve their performance in response to human feedbacks, however the model resulting from the training phase can't be easily accessed and interpreted by domain experts. In this paper we explore a different approach where the consistency models are explicitly built and validated by domain experts.

2) *Dependencies based methods*. Several approaches focus on integrity constraints for identifying errors, however they cannot address complex errors or several inconsistencies commonly found in real data [18, 21].

Other constraint types have been identified in the literature: multivalued dependencies, embedded multivalued dependencies, and conditional functional dependencies. Nevertheless, according to Vardi in [33] there are still semantic constraints that cannot be described.

In [3] a context-free-grammar based framework is used to specify production rules (e.g., Univ. → University), to reconcile the different representations of the same concept. Such approach mainly focuses on the attribute level, whilst the work presented in this paper focuses on set-of-records consistency.

Works on *database repair* focus on finding a consistent and *minimally different* database from the original one, however the authors of [11] state that computational issues affect the algorithms used for performing minimal-change integrity maintenance.

Deductive databases [25] add logic programming features to relational systems and can be used for managing consistency constraints. To the best of our knowledge, few works in the literature focus on deductive databases and data quality: [29, 19]. Furthermore, scalability issues have to be investigated when dealing with large sets of data.

In [10] database triggers are derived from dynamic constraints expressed in a time (first-order) logic variant. However triggers can raise computational issues when processing large datasets.

**Consistent query answering** works, e.g. [6], focus on techniques for finding out *consistent answers* from inconsistent data, i.e. the focus is on automatic query modifications and not on fixing the source data. An answer is considered consistent when it appears in every possible repair of the original

database. Semantic constraints are expressed using functional dependencies. Basically already with two Functional Dependencies the problem of computing Consistent Query Answers involving aggregate queries becomes NP-complete [6].

**Other works** and tools not included in the previous categories are now briefly surveyed. The application of automata theory for inference purposes was deeply investigated in [34] in the database domain. The problem of checking (and repairing) several integrity constraint types has been analyzed in [1]. Unfortunately most of the approaches adopted can lead to hard computational problems. Formal verification techniques were applied to databases, to formally prove the termination of triggers [9], for semistructured data retrieval [24], and to solve queries on semistructured data [17].

Many data cleansing toolkits have been proposed for implementing, filtering, and transforming rules over data. A detailed survey of those tools is outside the scope of the paper. The interested reader can refer to [21].

### 3 FROM DATA CONSISTENCY VERIFICATION TO MODEL CHECKING

Model checking [6] is a hardware/software verification technique to verify the correctness of a suitably modelled system. The model is described in terms of *state variables*, whose evaluation determines a state, and *transition relations* between states, which specify how the system can move from a state to the next one as a consequence of a given input action. Focusing on *explicit* model checking techniques, a model checker verifies if a state transition system (i.e., the model) satisfies a property by performing an exhaustive search in the system state-space (i.e., the set of all the possible system states). The model checker exploits techniques to reduce or compress the system state-space to be analysed, e.g. the reachability analysis: the state variable values that can be actually reached are identified, the reachable ones are analysed while the others are not (although being in the range of the admissible values).

The system model to be verified is expressed by means of a model checking language. Then the model checker generates a corresponding Finite State System (FSS) where the desired consistency properties can be evaluated. For the sake of completeness, we highlight that model checking languages can describe both an FSS and an *implicit representation* (i.e. abstract and general) of some FSSs. An implicit representation can be translated into an FSS, and the verification is always performed on the latter. Due to the space limitations, we do not formalise such implicit representation of FSSs. However, the reader can see [4] where such concept is expressed by means of *Extended Finite State Machines*.

**Definition 3.1 (Finite State System)** A Finite State System (FSS S) is a 4-tuple  $(S, I, A, F)$ , where:  $S$  is a finite set of states,  $I \subseteq S$  is a finite set of initial states,  $A$  is a finite set of actions and  $F: S \times A \rightarrow S$  is the transition function, i.e.  $F(s, a) = s'$  iff the system from state  $s$  can reach state  $s'$  via action  $a$ .

Hence, a trajectory is a sequence of state, action  $\pi = s_0 a_0 s_1 a_1 s_2 a_2 \dots s_{n-1} a_{n-1} s_n$  where,  $\forall i \in [0, n-1]$ ,  $s_i \in S$  is a state,  $a_i \in A$  is an action and  $F(s_i, a_i) = s_{i+1}$ .

Let  $S$  be an FSS according to Def. 3.1 and let  $\varphi$  be an invariant condition specifying some properties to be satisfied (called *safety properties* in the model checking domain) a state  $s_E \in E$  is an error state if the invariant formula  $\varphi$  is not satisfied. Then, we can define the set of *error states*  $E \subseteq S$  as the union of the states violating  $\varphi$ . We limit the error exploration to at most  $T$  actions (the finite horizon), i.e. only sequences reaching an error  $s_E \in E$  within the finite horizon are detected. Note that this restriction has a limited practical impact in our contexts although being theoretically quite relevant.

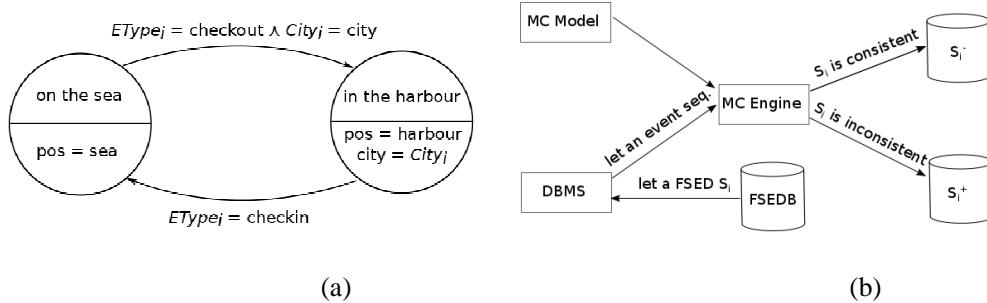
Model checking is traditionally used to explore and verify all the feasible execution paths of a system. Then, informally speaking a *model checking problem* is composed by a description of the FSS to be explored, an invariant to verify, and a finite horizon. A *feasible solution* (if any) is a trajectory leading the system from an initial state to an error one.

### 3.1 Finite State Events Database

In an event-driven architecture, an *exogenous event* represents a change that may occur in the system configuration due to an external occurrence. A connection can be established between event-driven systems and databases containing longitudinal data: a database record (or a subset thereof) can be seen as an *event* arriving from the external world, and an ordered set of records can be seen as an *event sequence* (or action sequence). More precisely:

**Definition 3.2 (Event, Event Sequence, and Finite State Event Dataset)** Let  $R=(R_1, \dots, R_n)$  be a schema relation of a database, let  $e=(r_1, \dots, r_m)$  be an event where  $r_1 \in R_1, \dots, r_m \in R_n$ , then  $e$  is a record of the projection  $(R_1, \dots, R_m)$  over  $R$  with  $m \leq n$ .

A total order relation  $\sim$  on events can be defined such that  $e_1 \sim e_2 \sim \dots \sim e_n$ . An *event sequence* (or *action sequence*) is a  $\sim$ -ordered sequence of events  $\varepsilon=e_1, \dots, e_n$ . A *Finite State Event Dataset* is a *longitudinal* dataset extracted from a database that can be expressed as an event sequence.



**Figure 1: (a) A Graphical representation of the Cruise Ship Travel Plan model. The lower part of a node describes how the system state evolves when an event happens. (b) A Graphical representation of a process where the model checker is used to verify an FSEDDB consistency.**

Intuitively, the application of model checking to data quality problems is driven by the idea that a *model* describing the consistent evolution of *feasible* event sequences can be used to verify if the *actual data* follows a consistent behaviour.

An FSS can be used to formalise the domain business rules and to check the consistency of Finite State Event Datasets. Although the whole content of a database could be checked by an FSS, it is advisable to split the database in several subsets (each being a separate FSED) and to check each of them separately.

**Definition 3.3 (Finite State Event Database)** Let  $S_i$  be an FSED, we define a Finite State Event Database (FSEDDB) as a database DB whose content is  $DB = \bigcup_{i=1}^k S_i$  where  $k \geq 1$ .

How can an actual database be verified by a model checker? A schematic representation of this approach is depicted in Fig. 1(b):

1. A domain expert codifies the evolution of the system as well as the consistency properties using the model checking tool language (i.e., the model).
2. An FSED  $S_i$  is retrieved from the database (i.e., the FSEDDB) and the model checker automatically generates an FSS representing the evolution of the model caused by  $S_i$ .
3. The model checker looks for an error trace on the FSS. A solution (if any) represents an inconsistency affecting the database event sequence  $S_i$ . Otherwise the event sequence is consistent.

Any model checker can be used to perform the verification. In our case, we used the CMurphi tool [7] which allows one to use C/C++ functions to interact with the database.

The concept of Consistency Failure Point (CFP) is now introduced: a CFP is an event of a FSED from which the sequence becomes inconsistent. The CFP event is not necessarily the responsible of the consistency failure, but it is the point where the failure emerges. The FSED is labelled as inconsistent if a CFP is discovered. The remaining of the event sequence can be hardly tested (or cannot be test at all) since the inconsistency might hinder the FSS-state-evolution identification thereafter. Considering the example of Tab. 1, the missing Lisbon departure prevents the exploitation of the FSS (Fig.1(a)) straight after the Lisbon checkin (the subsequent Barcelona checkin is the CFP), since other events could be missing, not only the Lisbon checkout. Generally speaking, the uncertainty originating after a CFP can prevent the execution of the consistency check for some or all the subsequent events. Considering again the example of Tab. 1, the uncertainty doesn't last for long time: the Barcelona harbour checkin event is enough to guess the FSS state and to resume the consistency check. In other cases, e.g. the one presented in Sec. 5, the uncertainty can last longer. The question is how to detect the points where the consistency check can be safely resumed. For this reason we introduce the *reset actions*. A *reset action* is an action so that the FSS state can be determined with certainty thereafter, even though the previous history is unknown. It can be observed that a reset action leads the FSS always to the same state, independently of the previous history. More formally:

**Definition 3.4 (Reset Action)** Let  $S(S, I, A, F)$  be a Finite State System according to Def. 3.1, an action  $a \in A$  is a reset action iff  $\exists s_a \in S$  s.t.  $\forall s \in S$  either  $F(s, a) = s_a$  or  $F(s, a)$  is not defined.

Since events can be mapped to actions, the *reset event* can be defined in a similar way: it is the event that lead the FSS always to the same state, independently of the previous history. The reset events can be used for partitioning a dataset into small event segments whose consistency can be evaluated independently. An example is showed in Sec. 5. In this way, a CFP found within a segment does not prevent the consistency evaluation of the subsequent segments.

**Running Example.** The following example should clarify the matter. Let us consider the Cruise Ship example as introduced in Tab. 1.

The whole dataset is the FSEDDB whilst a FSED is the travel plan of a single ship. An *event*  $e_i$  is composed by attributes *ShipID*, *City*, *Date*, and *Event Type*, namely  $e_i = (ShipID_i, City_i, Date_i, EventType_i)$ . Moreover, the total-order operator  $\sim$  could be the binary operator  $\leq$  defined over the event's attribute *Date*, hence  $\forall e_i, e_j \in E, e_i \leq e_j$  iff  $Date_{e_i} \leq Date_{e_j}$ . Finally, a simply consistency property could be “*if a ship checks in to the harbour A, then it must check out from A before checking in to the next harbour*”. We can model this consistency property as a model checking problem. An implicit representation of the domain is given in Fig. 1(a). In our settings, the system state is composed by (1) the variable *pos*, which describes the ship's position, and (2) the variable *city* describing the city where the ship is harboured. The consistency property of a database events sequence, e.g., the travel plan of Tab. 1, can be expressed as a model checking problem. In such a case, a solution (i.e., the error trace) is represented by the event sequence  $e_1, e_2, e_3, e_4$  which generates an inconsistent trajectory on the corresponding FSS.

## 4 DATA CLEANSING VIA MODEL CHECKING

In the previous sections we described how the consistency of a database event sequence can be modelled and verified through model checking. Looking forward, one can wonder if the consistency model can be used as the basis to identify cleansing activities. Namely, once the FSS describing the dataset consistency is generated, can the FSS be exploited to identify the corrective actions that can make such dataset consistent? Let us consider an inconsistent event sequence having an action  $a_i$  that applied on a (reachable) state  $s_i$  leads to an inconsistent state  $s_j$ . Intuitively, a corrective action sequence represents an alternative route leading the system from state  $s_i$  to a state when the action  $a_i$  can be applied (without violating the consistency rules). In other words, a *cleansing action sequence* (if any) is a sequence of actions that, starting from  $s_i$ , makes the system able to reach a new state on which the action  $a_i$  can be applied and results in a consistent state. More formally we can define the following.

**Definition 4.1 (Cleansing Action Sequence)** Let  $S = (S, I, A, F)$  be an FSS,  $E$  be the set of errors states (i.e. inconsistent states) and  $T$  be the finite horizon. Moreover,

- Let  $\Omega = \bigcup_{i \in E} \text{Reach}(s_i)$  be the set of all the states reachable from the initial ones;
- Let  $\pi = s_0 a_0 \dots s_i a_i s_j$  be an *inconsistent trajectory* where  $s_j \in \Omega$  is an inconsistent state (i.e.,  $s_j \in E$  and  $s_0, \dots, s_i \notin E$ ).

Then, a *T-cleansing action sequence* for the pair  $(s_i, a_i)$  is a non-empty sequence of actions  $A^c = c_0, \dots, c_n \in A$ , such that exists a trajectory  $\pi_c = s_0 a_0 \dots s_{i-1} a_{i-1} s_i c_0 s_{i+1} c_1 \dots s_{i+n} c_n s_k a_i$  on  $S$  with  $|A^c| \leq T$ , where all the states  $s_0, \dots, s_k$  are consistent.

In the AI Planning field a *Universal Plan* is a set of policy, computed off-line, able to bring the system to the goal from any feasible state (the reader can see [13, 5, 9] for details). Similarly, we are interested in the synthesis of an object, which we call *Universal Cleanser* (UC), which summarises for each *pair* (state, action) leading to an inconsistent state, the set  $A'$  of *all* the feasible cleansing action sequences. This UC is computed only once and then applied as an oracle to cleanse any kind of FSEDB.

To this aim, we proceed as follows:

**Step 1** A consistency model of the system is formalised by means of a model checking language as described in Sec. 3.

**Step 2** A database domain model is formalised, describing the attribute domains from which all the possible record subsets (i.e. event subsequences) composed by at most  $T$  events can be guessed (both the consistent and the inconsistent ones). The set of possible subsets will be called *worst case FSEDB* hereafter. E.g., for the cruise ship example an extract of the model is:  $\text{city}=\{\text{City}_x, \text{City}_y\}$   $\text{EType}_i=\{\text{checkin}, \text{checkout}\}$ . Note that the City attribute cardinality (although potentially unbounded) can be limited by a finite and small number thanks to the number of state variables and to the FSS *diameter*<sup>10</sup>.

**Step3** The model checker is used to generate the FSS representing all the inconsistent sequences, starting from the database domain model (step 2) and the consistency model (step 1), the whole process is shown in Fig. 1(b).

---

<sup>10</sup> Due to the limited space we provide only the intuition about how this task can be accomplished. The value is computed by the model checker as the *diameter* of the FSS, i.e. the largest number of states which must be visited in order to travel from one state to another excluding trajectories which backtracks or loops.

**Step4** Explore the FSS to synthesise the Universal Cleanser.

More formally, we define the Universal Cleansing Problem (UCP) and its solution.

**Definition 4.2 (Universal Cleansing Problem and Solution)** A Universal Cleansing Problem (UCP) is a triple  $D = \{S, E, T\}$  where  $S$  ( $S, I, A, F$ ) is an FSS,  $E$  be the set of error (or inconsistent) states computed by the model checker, and  $T$  is the finite horizon.

A solution for  $D$ , or a Universal Cleanser for  $D$  is a map  $K$  from the set  $\mathcal{Q} \times A$  to a subset  $A'$  of the power set of  $A$ , namely  $A' \subseteq 2^A$ , where for each inconsistent trajectory  $\pi = s_0 a_0 \dots s_i a_i s_j$  if  $A' \neq \emptyset$  then  $A'$  must contain all the possible  $T$ -cleansing action sequences for the pair  $(s_i, a_i)$ .

It is worth to highlight that, while on the one hand the UC generated is *domain-dependent*, i.e. it can deal only with event sequences conforming to the model that generated it, on the other hand it is *data-independent* since, once the UC is computed on a worst-case FSEDB, it can be used to cleanse *any* FSEDB. The pseudo code of the algorithm generating a Universal Cleanser is given in Procedures 1 and 2. It has been implemented on top of the UPMurphi tool [8]. The Procedure 1 takes as input the FSS of the domain, the set of error states given by the model checker (to identify inconsistent trajectories) and a finite horizon  $T$ . Then, it looks for a cleansing action sequence (according to Def. 4.1) for each inconsistent (state, action) pair. This work is recursively accomplished by the Procedure 2 which explores the FSS through a Depth First visit collecting and returning all the cleansing solutions.

**Running Example.** Consider again the Cruise Ship example of Tab. 1. We recall that an event  $e_i$  is  $e_i = (\text{ShipID}_i, \text{City}_i, \text{Date}_i, \text{EType}_i)$  and each event sequence and subsequence is ordered with respect to the event dates. It is worth to note that the finite horizon  $T = 2$  is enough to guarantee that any kind of inconsistency will be generated and then corrected using no more than 2 actions. Note that the cardinality of the city attribute can be potentially unbounded, but since a state can store only one city information at a time, we can use two elements ( $\text{City}_x$  and  $\text{City}_y$ ) to represent any feasible  $\text{City}_i$  value in the system. Consider that the main elements of an event are  $\text{EType}_i \in \{\text{checkin}, \text{checkout}\}$ ,  $\text{City}_i \in \{\text{City}_x, \text{City}_y\}$ , i.e., 4 possible events. Then, we represent the *worst-case* FSEDB by considering into our model all the possible 2-step event subsequences (i.e., simply enrich each node of the graph in Fig. 1(a) with all the possible edges).

Table 2 shows the Universal Cleansing for our example, which is *minimal* with respect to the number of event variable assignments, i.e., the missing pair  $([\text{pos}=\text{sea}], (\text{checkout}, \text{City}_y))$  fits on  $([\text{pos}=\text{sea}], (\text{checkout}, \text{City}_x))$ . The UC, once generated, is able to cleanse any kind of FSEDB compliant with the model from which it has been generated.

([state],(action))	list of corrective actions
([pos=sea], (checkout, City <sub>x</sub> ))	(checkin, City <sub>x</sub> )
([pos=harbour $\wedge$ city=City <sub>x</sub> ], (checkout, City <sub>y</sub> ))	(checkout, City <sub>x</sub> ),(checkin, City <sub>y</sub> )
([pos=harbour $\wedge$ city=City <sub>x</sub> ], (checkin, City <sub>y</sub> ))	(checkout, City <sub>x</sub> )
([pos=harbour $\wedge$ city=City <sub>x</sub> ], (checkin, City <sub>x</sub> ))	(checkout, City <sub>x</sub> )

**Table 2: Universal Cleanser for the Cruise Ship Example.**

---

**Procedure 1 UNIVERSALCLEANSING**

---

**Input:**  $FSS S$  ,  
set of error states  $E$ ,  
finite horizon  $T$   
**Output:** Universal Cleanser  $K$

- 1:  $level \leftarrow 0$ ; //to stop when  $T$  is reached
- 2: for all  $s^{err} \in E$  do
- 3:   for all  $s \in S, a \in A$  s.t.  $F(s, a) = s^{err}$
- 4:     do
- 5:        $K[s, a] \leftarrow AUXUC(s, a, s^{err}, level)$
- 6: return  $K$

---



---

**Procedure 2 AUXUC**

---

**Input:**  $s, a, s^{err}, level$   
**Output:** list of correction sequences  $cs[]$

- 1:  $cs[] \leftarrow \emptyset$  //list of correction sequences
- 2:  $cs_{aux}[] \leftarrow \emptyset$  //aux list of correction sequences
- 3:  $i \leftarrow 0$  //local  $cs[]$  index
- 4: if  $level < T$  then
- 5:   for all  $a' \in A$  s.t.  $F(s, a') = s'$  with  $s' \notin E$  do
- 6:     if  $F(s', a) = s''$  s.t.  $s'' \notin E$  then
- 7:        $cs[i] \leftarrow a'$
- 8:        $i \leftarrow i + 1$
- 9:     else
- 10:       $cs_{aux}[] \leftarrow AUXUC(s', a, s^{err}, level + 1)$
- 11:      for all  $seq \in cs_{aux}$  do
- 12:        $cs[i] \leftarrow a' \cup seq$
- 13:        $i \leftarrow i + 1$
- 14: return  $cs[]$

---

## 5 THE CASE OF “THE WORKERS CAREER ADMINISTRATIVE ARCHIVE”

The Italian Law No. 264 of 1949 requires the employers to notify the public administration whenever an employee is hired, dismissed, or her/his working contract is modified. Those notifications are called *Mandatory Communications* (“Comunicazioni Obbligatorie” in Italian). Since the 1997, the Ministry developed an ICT infrastructure, called the “CO System” [16], for recording data concerning mandatory communications, employment, and active labour market policies. Some administrative archives useful for studying the labour market dynamics [11] are generated and called “CO Archives” or “Job Registries”. Extracting the longitudinal data by the CO archives allows one to observe the overall flow of the labour market for a given observation period, obtaining insightful information about worker career paths, patterns and trends, facilitating the decision making processes of civil servants and policy makers [10]. Unfortunately the archive quality is very low, therefore cleansing is required before deriving information for decision making purposes (see, e.g. [3]). The approach presented in this paper has been used to perform data consistency evaluation and cleansing on the real data extracted from the CO archive of an Italian Area.

### 5.1 Domain Modelling

This subsection will provide some domain knowledge useful to achieve an overview of the administrative archives analysed in this paper. Every time an employer hires or dismisses an employee, or an employment contract is modified (e.g. from part-time to full-time, or from fixed-term to unlimited-term), a Mandatory Communication is notified to the CO System and stored into a job registry. The registries are managed at “provincial level” for several administrative tasks, every Italian province has its own job registry recording the working history of its inhabitants (as a side effect). For each worker, a mandatory notification (an *event* in our context) is composed by:

- w\_id:** it represents an id identifying the person involved in the event;
- e\_id:** it represents an id identifying the communication;
- e\_date:** it is the event occurrence date;
- e\_type:** it describes the event type occurring to the worker career. The allowed event types are: the *start* or the *cessation* of a working contract, the *extension* of a fixed-term contract, or a contract type *conversion*;
- c\_flag:** it states whether the event is related to a full-time or a part-time contract;
- c\_type:** it describes the contract type with respect to the Italian law (e.g. fixed-term or unlimited-term contract, etc.).
- empr\_id:** it uniquely identifies the employer involved in the event.

The evolution of a consistent worker's career along the time is described by a *sequence* of events ordered with respect to *e\_date*. More precisely, in this settings an FSED is the ordered set of events for a given *w\_id*, and the FSEDs union composes the FSEDB. Moreover, the representative element is given by the *w\_id*. Now we closely look to the consistency of the worker careers, where the consistency semantics is derived from the Italian labour law, from the domain knowledge, and from the common practice. Some rules can be identified:

- c1:** an employee can have no more than one full-time contract active at the same time;
- c2:** an employee cannot have more than K part-time contracts (signed by different employers); in our context we assume K = 2 i.e., employees cannot have more than two part time jobs active at the same time;
- c3:** a contract extension cannot change neither the existing contract type (*c\_type*) nor the part-time/full-time status (*c\_flag*) e.g., a part-time fixed-term contract cannot be turned into a full-time contract by an extension;
- c4:** a conversion requires either the *c\_type* or the *c\_flag* to be changed (or both).

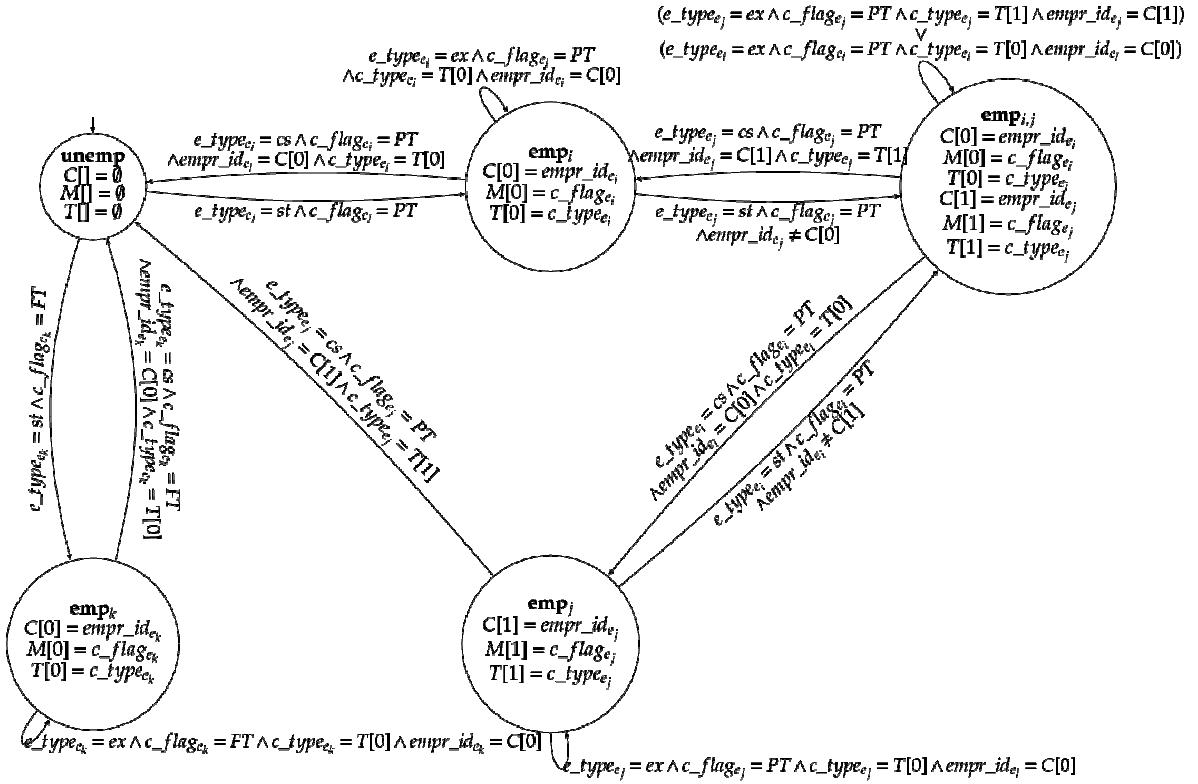
For simplicity, we omit to describe some trivial constraints e.g., an employee cannot have a *cessation* event for a company for which she/he does not work, an event cannot be recorded twice, etc.

The CMurphi model checker allows us to build an FSS which will be used to check the data consistency. The system state (i.e., a worker's career at a given time point) is composed by three elements: the list of companies for which the worker has an active contract ( $C[J]$ ), the list of modalities (part-time, full-time) for each contract ( $M[J]$ ) and the list of contract types ( $T[J]$ ).

To give an example,  $C[0]=12$ ,  $M[0]=PT$ ,  $T[0]=unlimited$  models a worker having an active unlimited part-time contract with company **12**.

The CMurphi model of the domain is showed in Figure 5.1 and it outlines a consistent career evolution. Note that, to improve readability, we omit to represent *conversion* events as well as inconsistent states/transitions (e.g., a worker activating two full-time contracts), which are handled by the FSS generated by the CMurphi model.

A valid career can evolve signing a part-time contract with company *i*, then activating a second part-time contract with company *j*, then closing the second part-time and then reactivating the latter again (i.e., *unemp*, *emp<sub>i</sub>*, *emp<sub>i,j</sub>*, *emp<sub>j</sub>*, *emp<sub>i,j</sub>*).



**Figure 2: A graphical representation of an FSS of a valid worker's career where  $st=$ start,  $cs=$ cessation,  $cn=$ conversion, and  $ex=$ extension.**

## 5.2 Data Consistency Experimental Results

We performed the consistency check using the model described in Fig. 5.1 on the “CO archive” of an Italian Area, composed by 1,248,751 mandatory communications. The CO archive ( $S$  from now on) describes how the labour market has evolved from the 1<sup>st</sup> January 2000 to the 31<sup>st</sup> December 2010, by providing CO events for 214,418 people careers. Each career has been modelled as a subset  $S_i$  where  $i \in [1, \dots, 214,418]$ . An  $S_i$  is a *FSED* while  $S$  is the *FSEDB* according to the terminology introduced in the previous section.

The consistency check computation was performed on a 32 bits 2.2Ghz CPU (connected to a MySQL server through ODBC driver) in about 20 minutes using about 50 MB of RAM.

Our results show that the 43.2% of the careers are inconsistent. More precisely, the 43.2% have *at least* one inconsistency (i.e., a CFP has been found). On the contrary, only the 56.8% of the total careers have proved to be consistent. Clearly, once an inconsistency is detected at a given time point, the remaining part of the career cannot be further evaluated since the CFP may have unpredictable effects on the consistency of the remaining part. To mitigate this effect, we exploited the consistency model of Fig. 5.1 to discover *reset events* (according to Def. 3.4) and to partition the careers into smaller segments. The following example should help to better clarify the usefulness of the reset events. Let us consider a worker career extracted from the dataset, as presented in Tab. 3(a). According to the record having  $e\_id=4$ , the worker w1 starts a new full-time contract in date 39504 without closing the on-going part-time. Due to this inconsistency, the whole career will be considered *inconsistent*, although only the first four events have been evaluated.

Focusing on the system described in Fig. 5.1, it can be observed that some events always lead the system

to a specific state regardless of the previous ones: e.g., looking at Fig. 5.1, a full time cessation always leads to the *unemp* state as well as a full time start always leads to the  $emp_k$  state. Indeed, in such cases, the state reached by the system can be guessed in spite of the previous uncertainty. These events contributing to reduce the uncertainty are the *reset events*.

w_id	id	e_date	e_t	c_flag	c_type	em_id
w1	1	38402	st	part-time	limited	empr <sub>1</sub>
w1	2	38679	st	part-time	unlimited	empr <sub>2</sub>
w1	3	39023	cs	part-time	limited	empr <sub>1</sub>
w1	4	39504	st	full-time	unlimited	empr <sub>3</sub>
w1	5	39651	cs	full-time	unlimited	empr <sub>3</sub>
w1	6	39700	st	part-time	unlimited	empr <sub>4</sub>
w1	7	40407	cs	full-time	unlimited	empr <sub>4</sub>
w1	8	40632	st	full-time	limited	empr <sub>5</sub>
w1	9	41449	ex	full-time	unlimited	empr <sub>5</sub>
w1	10	41513	cs	full-time	limited	empr <sub>5</sub>
w1	11	41726	st	full-time	limited	empr <sub>6</sub>
w1	12	42089	ex	full-time	limited	empr <sub>6</sub>

Table 3: (a) An example of a worker career (the data is not real although plausible).

Seg	w_id	id	e_date	e_t	c_flag	c_type	em_id
S1	w1 <sub>1</sub>	1	38402	st	part-time	limited	empr <sub>1</sub>
	w1 <sub>1</sub>	2	38679	st	part-time	unlimited	empr <sub>2</sub>
	w1 <sub>1</sub>	3	39023	cs	part-time	limited	empr <sub>1</sub>
	w1 <sub>1</sub>	4	39504	st	full-time	unlimited	empr <sub>3</sub>
S2	w1 <sub>2</sub>	4	39504	st	full-time	unlimited	empr <sub>3</sub>
	w1 <sub>2</sub>	5	39651	cs	full-time	unlimited	empr <sub>3</sub>
S3	w1 <sub>3</sub>	5	39651	cs	full-time	unlimited	empr <sub>3</sub>
	w1 <sub>3</sub>	6	39700	st	part-time	unlimited	empr <sub>4</sub>
	w1 <sub>3</sub>	7	40407	cs	full-time	unlimited	empr <sub>4</sub>
S4	w1 <sub>4</sub>	7	40407	cs	full-time	unlimited	empr <sub>4</sub>
	w1 <sub>4</sub>	8	40632	st	full-time	limited	empr <sub>5</sub>
S5	w1 <sub>5</sub>	8	40632	st	full-time	limited	empr <sub>5</sub>
	w1 <sub>5</sub>	9	41449	ex	full-time	unlimited	empr <sub>5</sub>
S6	w1 <sub>6</sub>	9	41449	ex	full-time	unlimited	empr <sub>5</sub>
	w1 <sub>6</sub>	10	41513	cs	full-time	limited	empr <sub>5</sub>
S7	w1 <sub>7</sub>	10	41513	cs	full-time	limited	empr <sub>5</sub>
	w1 <sub>7</sub>	11	41726	st	full-time	limited	empr <sub>6</sub>
S8	w1 <sub>8</sub>	11	41726	st	full-time	limited	empr <sub>6</sub>
	w1 <sub>8</sub>	12	42089	ex	full-time	limited	empr <sub>6</sub>

Table 3: (b) The segmented career of (a).

Using the UPMurphi tool and the model described Fig. 5.1, we verified that the *full-time* events always lead to the same state, i.e. they are *reset events*.

Given a FSED (according to Def. 3.2) describing a career composed of the events  $e_1, e_2, \dots, e_n$  the reset events  $e_{rej}$  (corresponding to full time events) are selected where  $e_{rej} \in [re_1, re_2, re_3, \dots, re_k] \subseteq [e_1, \dots, e_n]$ . The career can be splitted into segments as follows:  $[e_1, e_{re_1}], [e_{re_1}, e_{re_2}], [e_{re_2}, e_{re_3}], \dots, [e_{re_n}, e_n]$ . Excluding the last event of each segment (which is repeated as first event of the following one), the segments are non overlapping. The last event repetition is required to carry out the segment consistency check. The FSS for verifying the segment consistency has been modified by taking into account that a career segment can start from several states, not only from the *unemp* one.

Considering the example of Tab. 3(a), the career is decomposed by creating 8 segments which can be now analysed independently, as showed in Tab. 3(b). The consistency analysis on the segments shows that  $S1$ ,  $S5$ , and  $S6$  are inconsistent, whilst the remaining segments are consistent.  $S1$  is inconsistent because the job with employer  $empr_2$  is not closed before the beginning of the full-time contract with  $empr_3$ ,  $S5$  is inconsistent because the first extension event ( $e\_id=9$ ) has  $c\_type=unlimited$  and the extensions of an unlimited contract is not allowed. In  $S6$  there is a  $c\_type$  mismatch. As shown by this example, the segments can now be evaluated after the first inconsistency using the career segmentation.

We applied this approach on our administrative archive  $S$ , generating a new archive  $S^{segm}$  where each career has been decomposed into segments by using the reset events previously introduced. The consistency check has been used to evaluate the segments consistency. The results (and a comparison with the whole career results) are shown in Tab. 4. We highlight that the database  $S$  is largely composed by reset events (the full time events are about the 81% of total events) motivating the big dimension of the  $S^{segm}$  archive in terms of segments. For this reason, in  $S^{segm}$  a segment is now composed by a low average number of events, less than 2 per segment (not considering the duplicates). The number of consistent segments is the 78.3% compared to the 56.8% of the consistent careers (analysed as single entities). Thanks to the use of the reset events we obtained a more precisely evaluation of the consistency of  $S$  in terms of segments. Similarly, looking at the number of events belonging to inconsistent careers, the results show that now only the 28.3% of the total events of  $S$  belong to inconsistent segments (rather than the previous 72.2%).

Row	Dataset Analysis	$S$ (careers)	$S^{segm}$ (segments)
1	# Events	1,248,751	2,091,507
2	# Elements	214,418	1,057,090
3	#Consistent Elements	121,853 (56.8%)	828,194 (78.3%)
4	#Inconsistent Elements	92,565 (43.2%)	228,896 (21.7%)
5	#Events member of Consistent Elements	346,553 (27.8%)	895,906 (71.7%)
6	#Events member of Inconsistent Elements	902,198 (72.2%)	352,845 (28.3%)

**Table 4: A comparison between careers and segments data**

Even though the use of the reset actions has showed a more limited impact of inconsistencies in  $S$ , the analysis confirms that the original database has a low quality, motivating the need for data cleansing. The discussion about the reasons of such poor data quality is out of the scope of this paper, nevertheless it is mainly related to the data collection process (few controls, a lot of manual data entry especially before the 2005) and to some trivial errors (e.g. double entries) that can easily make the careers inconsistent.

### 5.3 Data Cleansing Experimental Results

We generated the Universal Cleanser using the model described in Fig. 4. We generated the FSS from the *worst-case* database by choosing a  $T = 5$  finite horizon, which is high enough to guarantee that any reachable inconsistent state can be considered. Then, Procedures 1 and 2 have taken as input the FSS

generated and the error states  $E$ , to detect inconsistent trajectories. Finally, Procedures 1 and 2 have been used to synthesise the Universal Cleanser. The UC contains 288 different (state, action) pairs able to make consistent any FSEDB (conforming to the model) in no more than 3 steps, avoiding looping corrective actions. We observed that  $T = 3$  is enough to guarantee that any inconsistency will be corrected, whilst using  $T = 2$  some errors cannot be fixed. To give an example, let us consider an inconsistent trajectory (i.e., a career in such a case) in which the last consistent state is  $\text{emp}_{ij}$  with ( $M:[\text{PT}, \text{PT}]$ ,  $T:[\text{Limited}, \text{Limited}]$ ,  $C:[\text{Company}_x, \text{Company}_y]$ ), then a cessation for a full-time contract with a new company arrives (i.e., an event as  $(\text{cs}, \text{FT}, \text{Limited}, \text{Company}_z)$ ). In such a case, the UC suggests to choose between two corrective interventions (similar to each other) composed by 3 actions for each. The first intervention is: to close the first part-time contract, i.e. ( $\text{cessation}, \text{PT}, \text{Limited}, \text{Company}_x$ ) then to close the second one ( $\text{cessation}, \text{PT}, \text{Limited}, \text{Company}_y$ ) and finally to start the full-time contract according to the event received ( $\text{start}, \text{FT}, \text{Limited}, \text{Company}_z$ ). The second intervention can be obtained by switching the first two cessation events.

We applied the UC generated to the dataset  $S$  to cleanse the inconsistent careers as follows. For each career  $S_i$ , when an inconsistency is found: (1) Let  $inc$  be a CFP (i.e. an inconsistency at a given sequence point) for the career  $S_i$ . (2) Look at the UC evaluating all corrective action sequences able to fix  $inc$ . (3) Select a suitable corrective action sequence (according to a given policy) and apply it. (4) Evaluate again the consistency of  $S_i$ . (5) Repeat steps 1-4 until no CFPs for the career  $S_i$  emerges.

In this work we focus on the UC synthesis. Investigating how to select corrective actions from the ones proposed by the UC is outside the scope of this paper. Nevertheless, for the sake of completeness, we detail how the UC has been used to cleanse the worker career archive. We implemented the step 2 by always selecting the corrective action sequence minimising (maximising) the (per worker) *average working days* indicator. Hence, we obtained two cleansed version of  $S$ , namely  $S^{\min}$  and  $S^{\max}$ , representing the cleansed versions of  $S$  in which inconsistent careers have been cleansed by minimising and maximising their working days respectively. In our settings, these distinct datasets allow us to perform a *sensitivity analysis* on the “working day” indicator with respect to the uncertainty due to inconsistencies. Clearly, once the UC is generated, the user can use any kind of policy for choosing a corrective action sequence. Finally, the complete UC has been made available at [1].

## 6 CONCLUSION AND FUTURE WORKS

In this paper we have shown how (longitudinal data) consistency verification tasks can be modelled as model checking problems, then we used the CMurphi verifier on some administrative archives to detect the inconsistent data. The analysed archives store the working histories of people living in an Italian area. An anonymous version of the archives has been used, according to the current law and privacy requirements. The results showed that the data quality of the source archives is very low: only about the 56% of people careers are consistent. To further investigate these results, we exploited the consistency model to partition the careers into small segments whose consistency can be analysed independently, obtaining a very fine grained evaluation of the data quality: the 78% of the segments turned out to be consistent.

Finally, we provided an algorithm working on the consistency model that can automatically build a *universal cleanser*: a cleanser *domain-dependent* (i.e., it focuses on the consistency of a specific domain) but *data-independent* (i.e., it can cleanse any kind of dataset compliant with the model).

Using model checking to evaluate a consistency model against actual data put into the hands of domain experts a powerful instrument contributing to a better comprehension of the domain aspects, of the data peculiarities, and of the cleansing issues.

As a future work we would like to explore the temporal logic to express consistency rules. Currently our research goes into the direction of comparing the universal cleanser with other approaches.

## REFERENCES

- [1] The universal cleanser of the worker career administrative archive. Public available at <http://goo.gl/OH74F>, 2012.
- [2] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer, 2006.
- [3] M. Cesarini, M. Mezzanzanica, and M. Fugini. Analysis-sensitive conversion of administrative data into statistical information systems. *Journal of Cases on Information Technology*, 9(4):57–81, 2007.
- [4] K. T. Cheng and A. S. Krishnakumar. Automatic functional test generation using the extended finite state machine model. In *Proceedings of DAC*, pages 86–91. ACM, 1993.
- [5] A. Cimatti, M. Roveri, and P. Traverso. Automatic OBDD-based generation of universal plans in non-deterministic domains. In *Proceedings of AAAI/IAAI*, pages 875–881, 1998.
- [6] E. M. Clarke, O. Grumberg, and D. A. Peled. *Model Checking*. The MIT Press, 1999.
- [7] CMurphi Web Page. <http://www.dsi.uniroma1.it/tronci/cached.murphi.html>, 2011.
- [8] G. Della Penna, B. Intrigila, D. Magazzeni, and F. Mercurio. UPMurphi: a tool for universal planning on PDDL+ problems. In *Proceedings of ICAPS 2009*, pages 106–113. AAAI Press, 2009.
- [9] G. Della Penna, D. Magazzeni, and F. Mercurio. A universal planning system for hybrid domains. *Applied Intelligence*, 36(4):932–959, 2012.
- [10] P. Lovaglio and M. Mezzanzanica. Classification of longitudinal career paths. *Quality & Quantity*, pages 1–20, 2012. 10.1007/s11135-011-9578-y.
- [11] M. Martini and M. Mezzanzanica. The Federal Observatory of the Labour Market in Lombardy: Models and Methods for the Construction of a Statistical Information System for Data Analysis. In *Information Systems for Regional Labour Market Monitoring - State of the Art and Prospectives*. Rainer Hampp Verlag, 2009.
- [12] T. C. Redman. The impact of poor data quality on the typical enterprise. *Commun. ACM*, 41:79–82, 1998.
- [13] M. Schoppers. Universal plans of reactive robots in unpredictable environments. In *Proc. IJCAI*, 1987.
- [14] J. Singer and J. Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press, USA, 2003.
- [15] S. Tee, P. Bowen, P. Doyle, and F. Rohde. Data quality initiatives: striving for continuous improvements. *International Journal of Information Quality*, 1(4):347–367, 2007.
- [16] The Italian Ministry of Labour and Welfare. Annual report about the CO system, available at [http://www.cliclavoro.gov.it/news/Documents/Rapporto\\_Annuale\\_Comunicazioni\\_Obbligatorie/executive\\_summary.pdf](http://www.cliclavoro.gov.it/news/Documents/Rapporto_Annuale_Comunicazioni_Obbligatorie/executive_summary.pdf), 2012.

# **IQ: PURPOSE AND DIMENSIONS**

(Research-in-progress paper)

**Phyllis Illari**

Department of Science and Technology Studies  
University College London  
[phyllis.illari@ucl.ac.uk](mailto:phyllis.illari@ucl.ac.uk)

**Luciano Floridi**

School of Humanities  
University of Hertfordshire and University of Oxford  
[l.floridi@herts.ac.uk](mailto:l.floridi@herts.ac.uk)

**Abstract.** This article examines the problem of categorising dimensions of information quality (IQ), against the background of a serious engagement with the hypothesis that IQ is purpose-relative. We examine some attempts to offer categories for IQ, and diagnose a specific problem that impedes convergence in such categorisations. Based on this new understanding, we suggest a new way of categorising both IQ dimensions and the metrics used in implementation of IQ improvement programmes according to what they are properties of. We conclude by outlining an initial categorisation of some IQ dimensions and metrics in standard use to illustrate the value of the approach.

## **1 INTRODUCTION**

Understanding information quality (IQ) is a pressing task. Undertaking it involves two related aspects, one conceptual and the other implementational. This is because what is needed is a settled analysis of IQ that matches definitions of IQ measures and improvement programs as well as ways to implement them. Unfortunately, current literature on IQ offers no settled agreement on answers to at least four closely related questions, all of which may be approached conceptually and implementationally:

What is a good general definition of IQ?

How should we classify the multiple dimensions of IQ?

What dimensions of IQ are there, and what do key features such as ‘timeliness’, ‘accuracy’ and so on mean?

What metrics might one use to measure the dimensions of IQ, bearing in mind that more than one metric may be required to yield an overall measure for a particular dimension?

These questions begin with the most clearly conceptual one, and descend to questions much more closely concerned with implementation. This dual nature of the problem of understanding IQ is recognised in the literature: ‘Both data dimensions and schema dimensions are usually defined in a qualitative way, referring to general properties of data and schemas, and the related definitions do not provide any facility for assigning values to dimensions themselves. Specifically, definitions do not provide quantitative measures, and one or more metrics are to be associated with dimensions as separate, distinct properties.’ (Batini & Scannapieco, 2006, p. 19)<sup>11</sup> Qualitative descriptions of the meanings of words or phrases such as ‘information quality’, or ‘timeliness’ are not the same as formal metrics required to measure them, and which are needed for implementation.

---

<sup>11</sup> Batini and Scannapieco page number references are to a manuscript copy.

In this paper, we intend to address only the conceptual aspect of the question, not the implementational one. However, since the two aspects are strongly connected, we will inevitably touch upon all four questions. Overall, it is a matter of finding a fine balance. On the one hand, a merely sequential procedure is unlikely to be fruitful: trying to answer question 1 first, then moving forward to question 2 and so forth is tempting but unlikely to succeed because, without some understanding of sensible implementable metrics and measures, it seems impossible to give a really meaningful general definition of IQ. On the other hand, it is equally unlikely to be fruitful to try to answer question 4 first, and then attempt to move backward to the others, because designing effective metrics for measuring IQ requires grasping what IQ itself is. Ultimately this set of questions needs to be answered collectively, so anyone trying to answer any of these questions is in a way concerned with all four. This might sound paradoxical, but in fact it is simply realistic. The idea is that, just as it takes two to tango, it takes both conceptual understanding and implementation, in alliance, to succeed with regard to IQ. We need to improve our conceptual understanding, then implementation measures, then back to conceptual understanding, and so on, until we get it right. With this in mind, we shall proceed in this article by developing a conceptual framework for approaching these questions, and then seek to map available metrics on to the developing conceptual picture. In this way, we hope to show that much of the task of answering the question of what IQ is indeed requires conceptual effort, and indicate what can be achieved by mapping implementable metrics to the conceptual framework we develop. In the light of this, we will not attempt in this paper to make a novel study of IQ practice, nor to extend any formal IQ metrics, although those studies must ultimately complement the conceptual study we engage in here. The ultimate test of this conceptual work is forward-looking: it will succeed if it does prove useful in moving forward the overarching project of improving IQ. We shall leave to a second stage of this project the implementational part, which will be developed in collaboration with Google UK.<sup>12</sup>

Here is a quick outline of the article. In section two, we shall briefly discuss question 1 above, noting the purpose problem for IQ. In section three, we shall examine the issue of dimensions and their classification, thus addressing questions 2 and 3 above. We shall discuss existing efforts to classify dimensions, and identify a problem that is impeding convergence of these efforts. We shall then offer our own classification, in terms of what IQ is a property of, and give an initial mapping of some IQ dimensions to that classification. In the conclusion, we shall quickly summarise the results obtained and articulate some final considerations about the so-called ‘purpose problem’ (more on this in the course of the article). A final terminological note: throughout this article we shall confine ourselves to considering ‘information quality’ or ‘IQ’. Much of the literature also writes of ‘data quality’ or ‘DQ’ and naturally we shall leave those expressions unaltered in any quotes. Yet in the following pages nothing theoretically significant depends on the distinction between IQ and DQ because, given the level of abstraction at which we are working, conceptual issues about IQ and DQ do not need to be distinguished.

## 2 PURPOSE

A major conceptual problem in the literature is the *purpose-dependence* of good information. The general idea is simple. For example, information is timely if it gets to you before you need to use it, and that depends on the purpose for which you intend to use it. Information that gets to you soon after it is gathered is not timely if it is too late to use; while information that gets to you the day before you need it is timely even if that information has been held up for a long while before it reaches you. Indeed, the obvious importance of purpose to IQ has gained so much currency that many working in, or influenced

---

<sup>12</sup> Research for this article was supported by a two-year project, entitled “Understanding Information Quality Standards and their Challenges” funded (2011-2013) by the British Arts and Humanities Research Council (AHRC), in collaboration with Google UK. Part of the project is to interact at a later stage with Google engineers to check and improve the conceptual model developed at an earlier stage.

by, the MIT group accept ‘fit for purpose’ as a general definition of IQ. For example: ‘Quality has been defined as fitness for use, or the extent to which a product successfully serves the purposes of consumers ....’ (Kahn, Strong, & Wang, 2002, p. 185). More recently, definitions of quality dimensions in the ISO/IEC 25012:2008 all make reference to a ‘specific context of use’ (ISO, 2008). One important feature, included in a specific context of use, is normal purposes in that context of use.

However, further and deeper analysis of the purpose-relativity of IQ and the connection of such analysis to implementation effectively have proven to be serious challenges: ‘While fitness for use captures the essence of quality, it is difficult to measure quality using this broad definition.’ (Kahn et al., 2002, p. 185). In particular, there is a need to understand how to lay out more specific IQ dimensions (questions 2 and 3) and specific metrics for these dimensions (question 4), against the background of a general definition of IQ (question 1) as broad as ‘fit for purpose’.

We have looked at purpose in previous work (Illari, 2012), which we summarise here as the background of the work of this paper. In that context, we argued that no IQ dimension is completely independent of purpose. This is true even though it may seem that some IQ metrics can be *defined* independently of purpose – such as tuple completeness, which measures whether there are missing values in tuples in the data – because a metric is an indicator of the dimension, and an indicator is not the dimension itself. The same view is shared by others: ‘These considerations show that even a dimension such as accuracy, which is considered only from the inherent point of view in the ISO standard, is strongly influenced by the context in which information is perceived/consumed.’ (Batini, Palmonari, & Viscusi, 2012). However, there is no need to conclude from the purpose-relativity of IQ, that IQ is *subjective*. Purpose is a *relational not a relative concept*: something has (or fails to have) a purpose for something else. Consider food, for example, it is a relation, but not a relative concept/phenomenon: something as a type (e.g., grass) is food for a specific type of eater (e.g., a cow) but not for another type (e.g., a human). Likewise, IQ does not depend merely on the opinion of the user. The purpose is chosen by the user, but how well different metrics and dimensions fit the same purpose is a matter of objective assessment; the user is constrained by the chosen purpose, and it is the purpose that determines IQ, not the user. What must be concluded instead is that what IQ means, and the best interpretations of the various IQ dimensions, are all dependent on the purpose of the information in question. We shall refer to this as the purpose problem.

In light of the purpose problem, it is worth viewing the fundamental challenge posed by IQ – in practice, rather than in the ideal case – as the request to represent and measure, as a purpose-independent feature of the information itself, something that is really a purpose-relative measure, i.e. it is a feature of the *relationship* between a purpose and the information itself. Specifically, improving IQ – of which defining it, defining and categorizing its dimensions, and designing metrics and measures for those dimensions are all a part – involves *getting metrics and so on that look purpose-independent although they aren't really*.

The metric or measure we get when we succeed is merely an estimate or indicator of IQ: ‘Although it is common in the IQ literature to talk of “measuring”, “evaluating” or “assessing” the quality of information, in practice the best we can hope for is to compute a close *estimate* of quality. ... At the end of all this, the best we can achieve is to combine the results from the various checks to make a defensible guess at the quality of the data, rather than a definitive, absolute measure of its quality.’ (Embrey, 2012). The result of making IQ indicators available to the user is to empower the user. This is in broad agreement with the following observation: ‘unless systems explicitly track their information quality, consumers of the information they provide cannot make judgments and decisions with high confidence. Information providers don't have to provide perfect IQ, but they need to be explicit about what IQ they do provide.’

(Keeton, Mehra, & Wilkes, 2009 p. 3)<sup>13</sup>

Recognising this tension between conceptual understanding of IQ and the practice of improving IQ should help to avoid misunderstanding, particularly the mistake of looking at something that has been designed to look purpose-independent, and taking it to be truly purpose independent. This is the background against which we approach the conceptual problems of categorising IQ dimensions and metrics. In particular, we will have something to say in our conclusion about the idea that the purpose problem is wholly different in kind from other IQ problems already dealt with successfully.

### **3 DIMENSIONS AND THEIR CLASSIFICATION**

We shall now try to show what can be achieved by keeping in mind that the process of improving IQ, including defining it, defining and categorizing its dimensions, and designing metrics to measure those dimensions, involves getting something that looks but is not purpose-independent.

In this section, we shall look at existing attempts to classify IQ dimensions, diagnose what may be wrong with them, and identify a fruitful approach. We shall then map some existing IQ metrics discussed by Batini and Scannapieco (2006) onto that approach. To anticipate, the main goal of this section is to show how important it is to understand IQ that we can be precise about what IQ itself and what various IQ dimensions and metrics are actually properties of. For example, are they properties of the data held by a single information producer? Or are they properties of the dynamic relationship between a whole information system, which is changing through time, and long-term users of that system?

The importance of answering such questions is a direct result of the purpose-relativity of IQ, and of the fact that a great deal of work designing and improving IQ involves trying to find a purpose-independent, intrinsic feature of the data itself to measure and use as an indicator of what is in fact a complex purpose-dependent feature of a relationship between data and user. Increased precision on these matters will help us understand how to think in a usefully clearer way about categories, dimensions and metrics. Ultimately we will argue for moving from a hierarchical organization of IQ dimensions and metrics to a relational model linking IQ dimensions and purpose.

#### **3.1 WHY EXISTING CLASSIFICATIONS OF IQ DIMENSIONS WON'T CONVERGE**

An important feature of the literature on IQ is an attempt to classify IQ dimensions. These attempts are proliferating, and there seems to be little settled convergence so far in the classifications produced. In this section, we shall examine some of the best known attempts at producing such categorisations of dimensions, and seek to diagnose the problem that is impeding a useful convergence in the debate on this issue. We begin with the categorisation of Wang (1998), which is one of the earliest and most influential categorisations of IQ dimensions, and is still frequently cited. Table 1: Wang's categorisation (Source: Wang (1998)) below is the table given in the original paper (Wang, 1998, p. 60):

<b>IQ Category</b>	<b>IQ Dimensions</b>
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Access, Security
Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational IQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

**Table 1: Wang's categorisation (Source: Wang (1998))**

---

<sup>13</sup>Page references are to a manuscript version of the paper.

This can be compared, for example, with the product and service performance model for information quality (PSP/IQ model), given in Table 22: Kahn et al.'s categorisation (Source: (Kahn et al., 2002, p. 184).. Its authors, who include Wang of the classification above, describe it as a 'two-by-two conceptual model for describing IQ. The columns capture quality as conformance to specifications and as exceeding consumer expectations, and the rows capture quality from its product and service aspects.' (Kahn et al., 2002, p. 184). Below is the table they give, mapping common IQ dimensions onto their model, using surveys of IQ practitioners (Kahn et al., 2002, p. 188):

	Conforms to Specifications	Meets or Exceeds Consumer Expectations
<b>Product Quality</b>	<u>Sound Information</u> <ul style="list-style-type: none"> <li>• Free-of-Error</li> <li>• Concise Representation</li> <li>• Completeness</li> <li>• Consistent Representation</li> </ul>	<u>Useful Information</u> <ul style="list-style-type: none"> <li>• Appropriate Amount</li> <li>• Relevancy</li> <li>• Understandability</li> <li>• Interpretability</li> <li>• Objectivity</li> </ul>
<b>Service Quality</b>	<u>Dependable Information</u> <ul style="list-style-type: none"> <li>• Timeliness</li> <li>• Security</li> </ul>	<u>Usable Information</u> <ul style="list-style-type: none"> <li>• Believability</li> <li>• Accessibility</li> <li>• Ease of Manipulation</li> <li>• Reputation</li> <li>• Value-Added</li> </ul>

**Table 22: Kahn et al.'s categorisation (Source: (Kahn et al., 2002, p. 184)).**

There are now quite a few dimension arrangements in the style of these two examples. Indeed, Lee, Strong, Kahn, and Wang (2002) even give us two comparison tables of classifications of IQ dimensions, one for academics in Table 3 (Lee et al., 2002, p. 134) and one for practitioners in Table 4 (Lee et al., 2002, p. 136), laid out according to the Wang (1998) categories:

	Intrinsic IQ	Contextual IQ	Representational IQ	Accessibility IQ
Wang and Strong [39]	Accuracy, believability, reputation, objectivity	Value-added, relevance, completeness, timeliness, appropriate amount	Understandability, interpretability, concise representation, consistent representation	Accessibility, ease of operations, security
Zmud [41]	Accurate, factual	Quantity, reliable/timely	Arrangement, readable, reasonable	
Jarke and Vassiliou [16]	Believability, accuracy, credibility, consistency, completeness	Relevance, usage, timeliness, source currency, data warehouse currency, non-volatility	Interpretability, syntax, version control, semantics, aliases, origin	Accessibility, system availability, transaction availability, privileges
Delone and McLean [11]	Accuracy, precision, reliability, freedom from bias	Importance, relevance, usefulness, informativeness, content, sufficiency, completeness, currency, timeliness	Understandability, readability, clarity, format, appearance, conciseness, uniqueness, comparability	Usableness, quantitativeness, convenience of access <sup>a</sup>
Goodhue [14]	Accuracy, reliability	Currency, level of detail	Compatibility, meaning, presentation, lack of confusion	Accessibility, assistance, ease of use (of h/w, s/w), Locatability
Ballou and Pazer [4]	Accuracy, consistency	Completeness, timeliness		
Wand and Wang [37]	Correctness, unambiguous	Completeness	Meaningfulness	

**Table 3: Classification for academics (Source (Lee et al., 2002))**

<sup>a</sup>Classified as system quality rather than information quality by Delone and McLean.

	<b>Intrinsic IQ</b>	<b>Contextual IQ</b>	<b>Representational IQ</b>	<b>Accessibility IQ</b>
DoD [10]	Accuracy, completeness, consistency, validity	Timeliness	Uniqueness	
MITRE [25]	Same as [39]	Same as [39]	Same as [39]	Same as [39]
IRWE[20]	Accuracy	Timeliness		Reliability (of delivery)
Unitech [23]	Accuracy, consistency, reliability	Completeness, timeliness		Security, privacy
Diamond Technology Partners [24]	Accuracy			Accessibility
HSBC Asset Management [13]	Correctness	Completeness, currency	Consistency	Accessibility
AT&T and Redman [29]	Accuracy, consistency	Completeness, relevance, comprehensiveness, essentialness, attribute granularity, currency/cycle time	Clarity of definition, precision of domains, naturalness, homogeneity, identifiability, minimum unnecessary redundancy, semantic consistency, structural consistency, appropriate representation, interpretability, portability, format precision, format flexibility, ability to represent null values, efficient use of storage, representation consistency	Obtainability, flexibility, robustness
Validity [8]			Metadata characteristics	

**Table 4: Classification for practitioners (Source (Lee et al., 2002))**

This is enough to illustrate the lack of convergence that should be cause for concern to those interested in the project of categorising dimensions. The problem is explicitly noted: 'In comparing these studies two differences are apparent. One is whether the viewpoint of information consumers is considered, which necessarily requires the inclusion of some subjective dimensions. The other is the difficulty in classifying dimensions, for example, completeness, and timeliness. In some cases, such as in the Ballou and Pazer study, the completeness and timeliness dimensions fall into the intrinsic IQ category, whereas in the Wang and Strong study, these dimensions fall into the contextual IQ category. As an intrinsic dimension, completeness is defined in terms of any missing value. As a contextual dimension, completeness is also defined in terms of missing values, but only for those values used or needed by information consumers.' (Lee et al., 2002, pp. 135-136). Here, they are commenting only on part of the overall comparisons they make, but the concern is clear: there is no settled agreement even on the most deeply embedded dimensions.

We suggest that there is a particular source of this problem, holding up any successful mapping of IQ dimensions onto categories. We shall develop this suggestion in the rest of this section, but must first pause to understand what creates the problem. Batini and Scannapieco (2006, p. 39) note: 'According to the definitions described in the previous section, there is no general agreement either on which set of dimensions defines data quality or on the exact meaning of each dimension. In fact, in the illustrated proposals, dimensions are not defined in a measurable and formal way. Instead, they are defined by means of descriptive sentences in which the semantics are consequently disputable.' The first important point is the descriptive, qualitative understanding of both categories such as 'intrinsic' and 'contextual',

and dimensions such as ‘timeliness’ and ‘accuracy’, however disputable, are performing a useful role in our conceptualisation of IQ. Categories such as ‘intrinsic’ and ‘representational’ and so on have an intuitive meaning, easy to understand and use, that is helpful to IQ practitioners and academics alike. The concepts of these categories are performing some kind of useful function in the academic literature, and in practice. Similarly for the concepts of IQ dimensions themselves, such as ‘accuracy’, ‘completeness’ and ‘timeliness’. They have intuitively understood meanings that are functioning usefully in the thinking of both practitioners and academics.

The importance of this role to the ultimate success of implementation of IQ improvement is also noted by Batini and Scannapieco (2006, p. 19): ‘The quality of conceptual and logical schemas is very important in database design and usage. ... Methods and techniques for assessing, evaluating, and improving conceptual schemas and logical schemas in different application domains is still a fertile research area.’ It is important that those working in IQ have both meaningful IQ dimensions and meaningful dimension categories to work with. The problem of imprecision that the intuitive meaning of category and dimension terms creates cannot be solved by eliminating all such words.

This is problematic because the IQ dimensions, defined according to the intuitive meaningful words that are generally used for dimensions, do not map onto the IQ categories, defined in turn according to the intuitive meaningful words that are commonly used for categories. We are going to spell this out in much more detail in the next subsection, by trying to offer a mapping between IQ dimensions and categories that will work, which will require adapting both categories and dimensions. Before, let us indicate the problem as briefly as possible. The heart of it is that the current meaningful dimensions have to be *split* to map properly onto existing meaningful categories. ‘Accuracy’, ‘timeliness’, ‘completeness’ and so on do not fit onto categories like ‘intrinsic’ and ‘contextual’ – only parts of these dimensions fit into each of these categories.

This is difficult to get clear, and so we shall illustrate the problem here very crudely (see Table 5: Dimensions fall into multiple categories), using the intrinsic-accessibility-contextual-representational categories of Wang (1998), and the well-known dimensions of accuracy, completeness and timeliness. The core idea is that accuracy has aspects that are intrinsic, but may also have aspects that fall under accessibility, contextual *and* representational features, as do both completeness and timeliness. Accuracy itself is not entirely intrinsic or representational, and so on, but shows aspects of all of the categories. Ultimately, as we have argued, all dimensions are purpose-relative.

<b>intrinsic</b>	<b>accessibility</b>	<b>contextual</b>	<b>representational</b>
Metrics that measure elements of accuracy, defined only on the data itself.	Information about such ‘intrinsic’ metrics, concerning availability to the user	Features of some or all of the ‘intrinsic’ metrics, relevant to the purpose for which the information will be used	Features of the presentation of the ‘intrinsic’ metrics that allow the user to use it effectively for his or her purpose
Metrics that measure elements of completeness, defined only on the data itself.	Information about such ‘intrinsic’ metrics, concerning availability to the user	Features of some or all of the ‘intrinsic’ metrics, relevant to the purpose for which the information will be used	Features of the presentation of the ‘intrinsic’ metrics that allow the user to use it effectively for his or her purpose
Metrics that measure elements of currency, defined only on the data itself.	Information about such ‘intrinsic’ metrics, concerning availability to the user	Features of some or all of the ‘intrinsic’ metrics, relevant to the purpose for which the information will be used	Features of the presentation of the ‘intrinsic’ metrics that allow the user to use it effectively for his or her purpose

**Table 5: Dimensions fall into multiple categories**

We hope the intended point is clear: aspects of *all four columns* in Table 5: Dimensions fall into multiple categories feed into an overall measure of the accuracy, the completeness, and the timeliness of the

information, in so far as these are dimensions of IQ itself.

This means that, while useful, this fourfold categorisation of dimensions does not categorise dimensions themselves, but something else. That something else is related to the kinds of category concepts that have been offered as useful up until now, but it is not related in the ways that have been assumed up until now: dimensions do not map onto categories 1-1: they do not map in such a way that each dimension can be allocated to one, and only one, category. This is what creates a problem. And although there may be other difficulties, this one by itself is already so significant to be sufficient to explain the lack of convergence in the debate on categories of IQ dimensions. Different scholars, with different intuitions about the most important *aspect* of accuracy, completeness and timeliness, will naturally allocate these dimensions to different categories.

The search for categories continues despite this problem, because there is a real need for something intervening between dimensions of IQ, and IQ itself, to give structure for thinking about IQ and its dimensions. But current approaches are not likely to succeed, since they all attempt to map each dimension to a single category. The risk is that, in order to fit square pegs in a round holes, the relations between the two are made increasingly loose, until fit is achieved only by means of irrecoverable vagueness. We shall attempt to use the insights developed here to make a positive suggestion to move the debate forward by splitting the dimensions. Initially, this will make both categories and dimensions less intuitively meaningful, but we hope to show how the overall framework ultimately recovers the meaningful aspects of both category and dimension terms currently in use, while still clearing away some of the current confusion.

### **3.2 What is IQ a property of? Towards a classification for IQ dimensions**

We shall now try to get more precise about the lesson learned from the discussion above, and begin the task of designing a classification of IQ dimensions that can generate settled agreement. We shall argue that what is vital to understanding IQ and hence being able to generate settled agreement in the different approaches to IQ is the answer to the question what *exactly* IQ itself and its dimensions are properties of. We first note the complexity of the problem. Batini and Scannapieco (2006) write: ‘definitions do not provide quantitative measures, and one or more metrics are to be associated with dimensions as separate, distinct properties. For each metric, one or more measurement methods are to be provided regarding ... (i) where the measurement is taken, (ii) what data are included, (iii) the measurement device, and (iv) the scale on which results are reported. According to the literature, at times we will distinguish between dimensions and metrics, while other times we will directly provide metrics.’ (Batini & Scannapieco, 2006, p. 19) In order to answer the four questions we began with, and so lay out a framework for consistent settled thinking about IQ, it is not just dimensions that we need to map onto the categories we have in mind: ultimately we also need to lay out the relations between dimensions, their categories, and metrics and measures.

We shall begin, then, by examining a suggested classification of IQ *metrics*, by Keeton et al. (2009 pp. 2-3).<sup>14</sup> (The construction of Table 6: Keeton, Mehra & Wilkes' classification of IQ metrics is ours):

---

<sup>14</sup>Page numbers for this paper also refer to a manuscript copy.

Standalone	'Standalone IQ metrics are independent of the use the information is put to, and can be directly measured by the information producer. They include: how recent is the data? how complete is it? how accurate is it? how representative is it (if sampled)?' (2)
Context-dependent	'Context-dependent IQ metrics can only be calculated relative to the context and needs of the information consumer. They generally cannot be evaluated by looking solely at a single information source.' (2)
Composite	'Composite IQ metrics are measures taken across multiple sources. For example: is this data source unique, or is there a duplicate copy obtainable elsewhere? Do these two sources agree (e.g., the strength of correlations or duplicate coverage between them)? Do we know the information's provenance? Is it auditable? Which source should be trusted more for the desired purpose?' (3)

**Table 6: Keeton, Mehra & Wilkes' classification of IQ metrics**

It is not difficult to see that these metrics are classified according to what they are defined on, which of course reflects what they are measures of. Standalone metrics can be measured directly by a single information producer, independently of the user, as they are defined on the data held by a single producer. Contextual metrics can only be calculated relative to the needs of the consumer, as they are features of the relationship between the data and some contextual element.

The classification above is quite specific to metrics, rather than dimensions in general, but it accords with our view of the importance of the question of what IQ is a property of. What IQ and its various dimensions and metrics can be defined on, and so what they actually track, is something well worth representing in a classification of IQ dimensions and metrics.

Consider what IQ could be a property of. Naturally, it is a property of information, but what information, exactly? There is a surprisingly large number of candidates:

- Single data item;
- Set of data about a particular worldly item;
- All data about a particular class of worldly items;
- All data in a database;
- Whole information system, even if it accesses multiple databases;
- Single data source;
- Whole information system, even if it accesses multiple databases, some or all of which use multiple sources;
- Whole dynamically evolving information system, so including IQ improvement measures which operate over time;
- Relation between entire (dynamically evolving) information system and a data consumer with a particular purpose (possibly a long-term one) in mind.

This list is probably not exhaustive. It may seem odd to count the later possibilities as possible bearers of IQ. But data is usually a collective. We do not usually worry about the quality of a datum, although we might, of course. However, clearly multiple data, or a collective of information, are legitimate bearers of information quality. As soon as that is noticed, the question of what collective we have in mind when assessing IQ is a natural one, and a question that is important for understanding IQ. It matters for what we count as, most obviously, completeness, but it also matters for other dimensions. If we think of the collective as the whole functioning information system, then dynamic properties of that system, such as correction mechanisms, become legitimate parts of the bearer of IQ.

Recall what we have indicated as the fundamental problem: that defining, modelling, and implementing good IQ requires transforming purpose-relative features of a whole information system into, as far as is possible, proxy indicators of IQ. These proxy indicators are, as far as is possible, intrinsic features qualifying only parts of the system itself, rather than properties of the relationship between the system

and its context. This means that they are features that can be defined on, and are properties of, the system itself, isolated from the world and from the purposes of any user. Now, a settled classification of standard IQ dimensions and metrics along the lines of what they are properties of would seem likely to help in the enterprise that engages with the fundamental problem.

This idea offers a way of categorising IQ dimensions that might lead to agreement and so convergence. We also hope to show that it will maintain some of the intuitive notions already in use, such as ‘intrinsic’ and ‘contextual’, which are already functioning usefully in the debate. These notions will be recoverable from the end result.

### 3.3 A new classification

The idea of a new classification is to look carefully at the information system, and identify parts of it that are different bearers of properties relevant to IQ, creating a diagram with spaces for each. Then start identifying the elements of the IQ improvement program: IQ itself, dimensions and metrics that you want to map. Then map the elements of the IQ improvement program onto the spaces representing the bearers of the property. Note that the mapping from dimension to category is not 1:1 but 1:N. Note also that there are two *kinds* of things that might be bearers of properties relevant to IQ, and the two must be distinguished:

- 1) Parts of the information system before you:
  - in which case the important thing is to get clear on which parts, as there may be several that are useful to distinguish.
- 2) Relations between the information system and something external to it, its ‘context’. This most notably includes:
  - the relation (deployment) between the information system and the purpose of the user, and,
  - the relation (reference) between the information system and the external world, particularly aspects of the world represented somewhere in your information system.

The difference between these two can no doubt be represented successfully in a myriad of ways. In our example below:

- 1) Properties of parts of the information system itself fall into columns, headed ‘Data, or the data in a particular population’, ‘a particular source of information’ ‘information in the single information system in front of you’, and ‘information across several information systems’ to discriminate different parts of an information system that may well be worth distinguishing.
- 2) Relations between the information itself and its context are represented by the ‘open’ columns on either side of the columns for the information system:
  - The left hand one ‘relation between the information system itself and the world’ allows representation of relations between the proxy indicators that can be defined on the information system, and features of the external world that are *not* the user or the purpose of use.
  - The right hand one ‘relation between information system and the purpose of the user’ allows representation of the other relational features of IQ.

We have made an initial mapping of some existing dimensions and metrics into this space. CAPITALISED words represent IQ dimensions, while words in lower case represent metrics or measures. A single row of the table contains metrics and measures that are related to the dimension also contained in that row – specifically, they are used as proxy indicators of the quality of the dimension.

This kind of mapping could usefully be done with any kind of element of IQ, including entirely new metrics, which may require more elements of the information system and its context than we illustrate below

to be identified as bearers of the properties measured. However, we will illustrate the idea of the mapping rather crudely and briefly using dimensions and metrics discussed by Batini and Scannapieco (2006), and using abstract descriptions of some of the kinds of things that we might want to identify as the bearers of the properties we are interested in when defining and constructing measures for IQ improvement. We begin with the dimension of timeliness in Table 7 below.

What is IQ a property of?					
The relation between information system and world	Data, or the data in a particular population	A particular source of information	Information in the single information system in front of you	Information across several information systems	The relation between information system and the purpose of a user
Rapidity of change in the target population	Volatility		Currency	Currency	TIMELINESS

Table 7: Timeliness and associated metrics

The idea is that timeliness is the dimension of IQ, which is relative to the purpose of use as already explained above. Currency is a metric which can be defined on the information itself, using something as simple as an update date, and it can be defined on information in one system or several, so that it falls into multiple columns. Currency does not yield timeliness, though, because whether an update date of two months ago is ‘recent’ depends on the volatility of the data in question – how rapidly the values of the data change. If your information is a house address, then 2 months ago is recent. If your information is levels of glucose within a metabolising cell, it is thoroughly obsolete. Volatility measures change in data, and of course this depends on the rapidity of change in the real-world target population.

With this simpler example in mind, we add other dimensions of usable accuracy and completeness in Table 8 below. The mapping is very far from complete or exhaustive. It is meant merely to illustrate. We suspect that this kind of mapping may be useful in many attempts to improve and better understand IQ, but that different aspects of the information system, on which different more specific metrics may be defined, will be more or less useful to identify in different cases.

What is IQ a property of?					
The relation between information system and world	Data, or the data in a particular population	A particular source of information	Information in the single information system in front of you	Information across several information systems	The relation between information system and the purpose of a user
Rapidity of change in the target population	Volatility	Sources may be characterised by usual quality	Currency	Currency	TIMELINESS
Semantic accuracy	Semantic accuracy	Sources may be characterised by usual quality	Syntactic accuracy Comparison functions Edit distance	Syntactic accuracy Comparison functions Edit distance	USABLE ACCURACY
Open World Assumption versus Closed World Assumption	Population completeness	Sources may be characterised by usual quality	Attribute completeness Entity completeness Column completeness	Attribute completeness Entity completeness Column completeness	COMPLETENESS

Table 8: Other dimensions and their associated metrics

As for timeliness, usable accuracy, and completeness with respect to purpose are the true dimensions of IQ, and, as we have argued above, they are dependent on the purpose of the user. Well-known metrics that are used as indicators of these dimensions can be defined on a single information system, and on multiple information systems. Some can be defined on a single attribute, such as attribute completeness. In both cases, again, there is also an important relation to the world. Semantic accuracy concerns whether the information in your system matches worldly values, while choosing between closed or open world assumptions involves making a big assumption – which should be marked – about the relation between the information in the system and the world. Again, useful relations between metrics as indicators of quality dimensions, the purpose of the user, and the nature of the world can be seen laid out in this manner.

The simplified mapping above was achieved conceptually, by examining the definitions and measures to pick out precisely what aspects of the information system they are defined on. Nevertheless, some quite interesting conclusions can be drawn. First, it is worth putting quite a few different elements of the information system into the columns for this mapping, and it is not difficult to think of more things that could usefully be represented. Second, many of the elements of IQ are properties of relations. Even some, such as semantic rules and integrity constraints, which can be defined on the information system itself, are properties of quite complex relationships. They remain properties of the information system itself, because those complex relationships are themselves internal to the information system. But note that semantic rules are often, if not always, constructed *successfully* using world-knowledge. Third, as expected, even though the dimensions of IQ themselves are properties of the relation between the whole information system and the user, some elements of all of them, particularly metrics used to measure them, can sensibly be defined just on the information system itself, so allowing such metrics to be properties of that system.

Finally, we note that another problem becomes very clear in the process of doing this mapping. It might be thought of as the other side of the purpose problem. Sometimes it is essential to represent explicitly relations between something in the information system, or the whole system, and the world. Some completeness and accuracy measures cannot eliminate this. There is also a further feature that receives insufficient attention: many of our design metrics, which can be defined on the data, still depend heavily on world-knowledge for their design, and for our confidence that they will work. This is true for semantic rules constructed after the fact for survey data, such as that someone who is 10 years old cannot also be a parent. It is also crucial in the choice between the open world or closed world assumptions used to design completeness measures. Any such world-knowledge is empirical and contingent, and might change. Like the purpose problem, this should also be explicitly represented, so that it cannot be forgotten in IQ improvement programmes.

### **3.4 Discussion**

The idea has been to move from a hierarchical organization of IQ dimensions and metrics to a relational model linking IQ dimensions and purpose. To this end, the previous mapping offers several advantages, including the possibility of convergence of a classification of IQ metrics and dimensions, a classification sensitive to what IQ improvement programs are really trying to do, a clear indication of potential pitfalls, and finally a valuable recovery of important concepts like ‘intrinsic’ and ‘contextual’. We shall briefly comment on each of them in turn.

First, convergence should be encouraged by this mapping, because it should be possible to map metrics and dimensions onto this kind of space, and useful in sharpening up their definition, and their interrelations. Deciding what such things are properties of – what they can be defined on – is a matter of objective assessment and should be much easier to agree on than whether entire IQ dimensions are, for example, ‘intrinsic’.

Second, this kind of mapping lays out the tools of IQ improvement in a way that is sensitive to what IQ improvement programmes try to do. It lays out the relationship between metrics that are genuinely objective measures of the data itself, and highly purpose-dependent features of the whole system. The place of such metrics as mere indicators of the relational IQ dimensions is clear. The tables give a representation of the scale of the problem, and what is being done.

Third, as a complement to the table laying out useful features of tools, it also represents the gaps. These mappings visually represent where the enterprise of finding intrinsic features of the information to act as proxy indicators of properties of relational features is forced, where the metric or dimension is a property of a relation. The forced nature of proxy indicators of the quality of the information for the purposes of the user will not be blurred or easily forgotten with such maps in mind.

Finally, this mapping allows the recovery of some important intuitive terms in the literature, but in more precise form. We suggest that intrinsic IQ metrics are those that can be defined solely on the information system itself, such as some specific completeness metrics. These are properties of the information stored, and our mapping still has the advantage of encouraging continuous attention to exactly what feature of the information stored they are properties of. Note, though, that it tends to be only metrics, and only some of them, which are intrinsic in this sense. And in so far as such metrics relate to IQ, they are always proxy indicators of a more complex relational property. Contextual features of IQ are those which attempt to measure something about the relationship between the information system and its context. We have now identified the two crucial features of that context: a) the relation between the information system and the purpose of the user, b) the relation between the information system and the world, including of course features of the world explicitly represented, such as birth dates, but also features of the world used to construct appropriate semantic rules for checking consistency. Ideas of ‘representational’ and ‘accessibility’ relations are less easy to define precisely. But we suggest they are thought of explicitly as themselves features of the relationship between the information and the user, which is an idea that requires future work.

Ultimately, our mapping has many advantages, and recovers the intuitive usability of terms that are performing a useful role in both the literature and practice.

## **4 CONCLUSION**

We have briefly summarised our reasons for thinking that the purpose problem for IQ is serious, and that much of the work on IQ responds by looking for proxy indicators of IQ that can be defined on features of the information system itself. We have offered our approach to mapping elements of all major concepts engineered for IQ improvement onto a space designed to represent what they are properties of. This is our first attempt to address the four interrelated questions with which we began:

1. What is a good general definition of IQ?
2. How should we classify the multiple dimensions of IQ?
3. What dimensions of IQ are there, and what do key features such as ‘timeliness’, ‘accuracy’ and so on mean?
4. What metrics might one use to measure the dimensions of IQ, bearing in mind that more than one metric may be required to yield an overall measure for a particular dimension?

Our mapping offers a way of seeing the problems laid out collectively, showing how much in common they have. Fitness for purpose is vital to IQ, and should inform understanding of the purpose of a classification, and also identification of dimensions and the design of metrics. It is due to the difficulty of addressing the fitness for purpose problem that metrics are used, as they are, as proxy indicators of purpose-

dependent dimensions. This research will continue by examining further metrics and adding to the mapping above, and expanding understanding of how they are designed to meet the purpose problem.

We shall now conclude this article by making a few further remarks on the purpose problem. The purpose problem looks daunting when it appears to be wholly different in kind from any other problem dealt with in designing IQ metrics. It appears different if it alone involves human subjectivity, human intention, human minds. But a final advantage of our mapping is to show that, while difficult, the purpose problem is not wholly different in kind from other problems that are dealt with very successfully. The purpose problem is just that some IQ concepts – notably its dimensions – are properties of the relation between the information system and the purpose of the user. It is the other side of the problem that some IQ metrics are properties of the relation between the information system, or aspects of it, and the external world.

Further, properties of relations are not in themselves intractable. Relational properties internal to the information system itself are frequently defined very well, such as integrity constraints. The purpose problem is just that the bearer of some features of IQ is the relation between system and purpose of user. But there is nothing here that can't be measured in principle. The relation might be imperfectly measured, perhaps, but no more imperfectly than some relational features internal to the information system itself are measured. If the purpose requires speed more than accuracy, this trade-off can be assessed, proxy measures found and implemented. If the purpose requires completeness, this too can be assessed, measures created and implemented, then tested and adjusted, and so on. From another point of view, we could track user choices, given stated purpose, and learn how to improve measures of the relation between the system and purpose that way.

To summarise: one side of the problem is just that we have to relate the information system to the world. This is probably going to mean that some measures will remain ineliminably domain-specific. The other side is that we have to relate the information system to the purpose of the user. So some measures will remain ineliminably purpose-specific. These two are both ineliminably contextual – but tractable – features of IQ.

## **ACKNOWLEDGEMENTS**

Research for this article was supported by a two-year project, entitled “[Understanding Information Quality Standards and their Challenges](#)”, currently funded (2011-2013) by the British Arts and Humanities Research Council (AHRC).

## **REFERENCES**

- [1] Batini, C., Palmonari, M., & Viscusi, G. (2012). *The Many Faces of Information and their Impact on Information Quality* Paper presented at the AISB/IACAP World Congress, University of Birmingham.  
[http://philosophyofinformation.net/IQ/AHRC\\_Information\\_Quality\\_Project/Proceedings.html](http://philosophyofinformation.net/IQ/AHRC_Information_Quality_Project/Proceedings.html)
- [2] Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*: Springer.
- [3] Embury, S. (2012). *Forget Dimensions. Define Your Information Quality Using Quality View Patterns*. Paper presented at the AISB/IACAP World Congress, University of Birmingham.  
[http://philosophyofinformation.net/IQ/AHRC\\_Information\\_Quality\\_Project/Proceedings.html](http://philosophyofinformation.net/IQ/AHRC_Information_Quality_Project/Proceedings.html)
- [4] Illari, P. (2012). *IQ and Purpose*. Paper presented at the AISB/IACAP World Congress, University

of Birmingham.

[http://philosophyofinformation.net/IQ/AHRC\\_Information\\_Quality\\_Project/Proceedings.html](http://philosophyofinformation.net/IQ/AHRC_Information_Quality_Project/Proceedings.html)

- [5] ISO. (2008). IEC FDIS *Software Engineering - Software Product Quality Requirements and Evaluation - Data Quality Model* (Vol. 25012).
- [6] Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, 45(4), 184-192.
- [7] Keeton, K., Mehra, P., & Wilkes, J. (2009 ). Do you Know your IQ? A Research Agenda for Information Quality in Systems *ACM SIGMETRICS Performance Evaluation Review*, 37(3), 26-31.
- [8] Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A Methodology for Information Quality Assessment. *Information & Management*, 40(2), 133-146. doi: 10.1016/s0378-7206(02)00043-5
- [9] Wang, R. Y. (1998). A Product Perspective on Total Data Quality Management. [Article]. *Communications of the ACM*, 41(2), 58-65. doi: 10.1145/269012.269022

# **AN INVESTIGATION INTO DATA QUALITY ROOT CAUSE ANALYSIS**

(Research-in-progress)

**Philip Woodall**

University of Cambridge

phil.woodall@eng.cam.ac.uk

**Andy Koronios**

University of South Australia

andy.koronios@unisa.edu.au

**Jing Gao**

University of South Australia

jing.gao@unisa.edu.au

**Ajith Kumar Parlakad**

University of Cambridge

ajith.parlikad@eng.cam.ac.uk

**Elaine George**

University of South Australia

Elaine.George@unisa.edu.au

**Abstract:** Data and information quality is a well-established research topic and gradually appears on the decision-makers' top concern lists. Many studies have been conducted on how to investigate the generic data/information quality issues and factors by providing a high-level abstract framework or model. As a result, a number of data and information quality methodologies and frameworks have been developed and made available to organisations. Across all examined methodologies and data quality improvement frameworks, this study found that although data quality root cause analysis (RCA) is regarded as an essential data quality improvement method, there is limited guidelines on how to conduct RCA to investigate data quality problems.

**Key Words:** Data quality, root cause analysis, root causes, data quality improvement, data quality methodology

## **INTRODUCTION**

To begin, it is noted that data and information are often used synonymously particularly when addressing quality issues. In practice, managers differentiate information from data intuitively, and describe information as data that has been processed. Unless specified otherwise, this paper will use data interchangeably with information, as well as use Data Quality (DQ) interchangeably with Information Quality (IQ).

It is generally agreed now that a lot of intellectual property is locked in enterprise data repositories. Enterprise data and information is increasingly recognized as one of the most valuable, if not the most valuable and proprietary resources that enterprises possess. It is also true however that the value of this asset needs to be unlocked for the enterprise to use it to make better decisions and to gain competitive advantage. Data and information development and management is essential for improving or developing new contexts to support the management process and make, strategic decisions([29]; [40]; [45]). For example, managers derive information from data to enable them to make operational decisions related to produc-

tion, ordering and scheduling. Accountants use data to prepare financial statements and documents including financial reports. The importance of data has been increasing. Information is an ever increasing important business resource that supports organisational decisions, and, therefore, management of the quality of this information is critical for organisational success.

As presented in existing literature and case studies, poor data quality (DQ) is often discussed as problems in organizations [19]. Regardless of the actual problem context, the typical problem solving exercise often involves Root Cause Analyses (RCA). Indeed, in industries such as manufacturing and aviation, RCA is a critical and compulsory activity for any problem solving exercises.

Although RCA had its genesis in engineering environments, many such techniques and methods have been used in a number of different contexts. Table 1 below lists the more frequently used RCA techniques and illustrates some of the limitations of each.

<b>Method/ Tool</b>	<b>Type</b>	<b>Defines Problem</b>	<b>Defines all causal relationships</b>	<b>Provides a causal path to root causes</b>	<b>Delineates evidence</b>	<b>Explains how solutions prevent recurrence</b>	<b>Easy to follow report</b>
Events and causal factors	Method	Yes	Limited	No	No	No	No
Change analysis	Tool	Yes	No	No	No	No	No
Barrier analysis	Tool	Yes	No	No	No	No	No
Tree diagrams	Method	Yes	No	No	No	No	No
Why-why chart	Method	Yes	No	Yes	No	No	No
Pareto	Tool	Yes	No	No	No	No	No
Story telling	Method	Limited	No	No	No	No	No
Fault tree	Method	Yes	Yes	Yes	No	Yes	No
Failure modes and effects analysis (FMEA)	Tool	Yes	No	Limited	No	Limited	No

**Table 1 Comparison of selected RCA methods and Tools (Modified table from [18])**

Within the data quality improvement context a number of additional RCA techniques have been used. These include Scatter diagram and Stratification/ Is/Is Not Analysis. [5] describes RCA within the DQ improvement context as a means of investigating and categorizing the root causes of IQ problems. It seems that the plethora of data quality improvement frameworks are deficient in their inclusion of RCA, yet only when we are able to determine why IQ problems have occurred, are we able to specify workable corrective measures that prevent future problems and result in sustained data quality improvement [5].

Although it is acknowledged that many DQ frameworks and methodologies are abstract in nature with little guidance on how these frameworks could be operationalized into prescriptions for data quality improvement and thus little guidance on performing RCA one of the most critical activities in preventing data quality errors from recurring. This research endeavours to address these shortcoming by investigating how RCA in data quality contexts is currently performed and how might this be improved.

## **Research Question and Design**

With the above consideration, a two-stage study has been developed in order to answer the following research question:

*To what extent is RCA, as is recommended and described in DQ improvement methodologies, sufficient to extract the root causes of DQ problems?*

The following sub-questions will be investigated in order to answer the research question

1. To what extent is RCA emphasised in DQ improvement methodologies?;
2. Are there sufficient guidelines for performing DQ RCA?; and
3. Does DQ RCA differ from generic RCA in real world practices?

This paper will focus in reporting on the findings of the first stage of this project.

The first stage of this study will examine a number of the most representative generic data quality frameworks and improvement methodologies in order to determine the role and importance of RCA. By a close investigation of these frameworks and methodologies, we also try to extract the DQ RCA guidelines which are currently available to data quality practitioners.

The second stage of this study will involve actual case studies. A number of organisations who have implemented DQ initiatives will be chosen with a particular focus on how they conducted data quality root cause analysis and the results.

This study is intended to increase understanding on how data quality root cause analysis is conducted in practice and highlight the problems and challenges in performing root cause analysis for data quality improvement. This study may also deliver a common root causes for some common data quality problems shared by organisations world-wide. More importantly, it may help the academic society enhance frameworks and methodologies for better adoption.

Within the data quality context, [5] describes RCA as a means to “Investigate and categorize the root causes of IQ problems. It is only when the reasons for each data quality error that exist have been identified that it will be possible to determine why IQ problems occurred and thus recommend appropriate corrective measures that prevent future data quality errors. This study attempts to obtain an insightful understanding on how RCA is conducted in DQ initiatives.

## **DATA QUALITY FRAMEWORKS AND IMPROVEMENT METHODOLOGIES**

As the field of data quality developed a large number of data quality frameworks and improvement methodologies have emerged (e.g., [47]; [50]; [48]; [44]; [39]; [24]; [20]; [7]; [12]; [35]; [22]; [15]). These frameworks have attempted to organize and structure important issues in data quality from a number of different perspectives. Some of these are generic frameworks whilst others are specific to particular domains, for example health or asset management.

An analysis of the most well known data quality frameworks and improvement methodologies between 1990 and 2011 was identified through a literature review (the results are listed in Table 1). These frameworks and methodologies are constructed to provide a comprehensive coverage of IQ problems, related activities, and context-driven IQ dimensions [46]. Additionally, these IQ frameworks can also be used as knowledge resources to provide guidelines on how to ensure information quality for various environments ([37]; [46]).

<b>Author (Year of Publication)</b>	<b>Area</b>
[2]	Management Information Systems
[43]	Newspapers
[28]	Corporate Communications
[49]	DQ Research
[41]	Data Bases
[32]	Data Quality Methodology
[38]	Data Quality Methodology
[33]	Information Systems
[47]	Information Systems
[10]	Information Management
[13]	Corporate Communications
[3]	Data Warehouses
[23]	Information Systems
[21]	Knowledge Management
[6]	Data Bases
[11]	Data Quality Methodology
[42]	Data Quality Methodology
[24]	Information Systems
[27]	Information Systems
[31]	Corporate House holding
[34]	Information Systems
[4]	Data Quality Improvement
[16]	Data Quality Methodology
[14]	Data Quality Assessment
[8]	Portal Data Quality
[17]	Data Quality Metrics
[26]	Sensor Data
[9]	A quality framework to evaluate E-Government service delivery
[36]	Data Quality in Data Warehousing
[25]	Enterprise Knowledge Management
[30]	Data Quality Methodology
[48]	Data Quality Methodology
[1]	Information Quality Framework for e-Learning Systems

**Table 2: Information Quality frameworks**

Some of the above frameworks and methodologies are generic and others focus on a special area or industry. Nevertheless, these DQ frameworks will help researchers and practitioners to obtain an in-depth understanding of various data quality issues.

## **IMPROVEMENT METHODOLOGIES THAT INCLUDE RCA**

From the above list, we have identified the DQ methodologies that recommend conducting RCA, and these are shown in Table 3. Interestingly, many of these are from leading experts in the DQ area who have produced methodologies that have gained wide acceptance. RCA is therefore not a peripheral topic in DQ improvement proposed by a one-off methodology. Rather, it has been independently agreed by experts as a necessary step to DQ improvement.

In the following subsections, RCA is discussed within each methodology giving attention to the sub research questions 1 and 2. In particular, in relation to sub question 1, we discuss where RCA has been included in the methodology and whether it is considered an important part. After reviewing all the RCA instances, the issue of whether RCA is used consistently is discussed.

For sub question 2, the issues of how detailed the RCA guidelines are, what methods are provided for RCA and whether the guidelines have been tailored to DQ problems are also discussed for each methodology.

Methodology reference name	Ref	Placement of RCA	RCA importance	Suggested RCA methods	Level of detail
<u>TQdM</u>	[6]	Informs the corrective improvement actions to implement	RCA results are critical to help develop improvement actions	Pareto chart analysis, Five whys, Cause-effect diagram,	Gives details on cause-effect diagram
<u>TDQM</u>	[48]	Uses the DQ measurement results to identify root causes before improving the Information Product	RCA results are critical to help develop improvement actions	SPC, Pattern recognition, Pareto chart analysis, Introduce dummy records	High level of abstraction
<u>SODQA</u>	[32]	Uses the DQ measurement results to identify root causes before informing what corrective improvement actions to implement	RCA results are critical to help develop improvement actions	none suggested	High level of abstraction
<u>DQFG</u>	[38]	Informs the corrective improvement actions to implement	RCA results are critical to help develop improvement actions	none suggested	High level of abstraction
<u>EDQP</u>	[30]	Informs the corrective improvement actions to implement	RCA is used in multiple steps of the methodology	Five whys, Cause-effect diagram, Track and trace	Most detailed
<u>CSDQ</u>	[11]	Uses the DQ measurement results to identify root causes before improving the Information Product	RCA results are critical to help develop improvement actions	Cause-effect diagram, Causes table	Some detail, but still high level
<u>AMEQ</u>	[42]	Identical to TDQM	Identical to TDQM	none suggested	Very high level

Table 3: DQ Methodologies that recommend root cause analysis

### ***RCA in TQdM***

The TQdM methodology actually provides two approaches to DQ improvement: process-driven and data-driven. The data-driven approach does not suggest doing RCA and it is the process-driven approach that includes RCA as one of its steps. The steps in the process-driven approach are shown in Table 4 and the RCA is recommended as part of the ‘Develop plan for DQ improvement’ step. This is a critical step because it leads to the implementation of the corrective actions, which aims to modify the organisation’s data, and the root causes are needed to determine what actions can and should be taken to address the problem(s). Clear guidelines and examples are given showing how to use the cause-effect diagram for DQ problems. The Pareto diagram and the five whys are also mentioned.

Step	Description of Step	Link to study [6]
Select process for DQ improvement	Identify a process where improvements can prevent business problems that cause DQ problems.	p289 step 1
Set up a DQ team	Identify a person responsible for resolving DQ problems, a project sponsor, and team members to facilitate the DQ improvement process.	p292 step 1, point 1, 2 and 3.
Develop plan for DQ improvement	Identify the root causes of a DQ problem and identify corrective actions to eliminate/minimise the causes.	p293 step 2
Implement DQ improvements	Implement improvement actions in a controlled manner to improve DQ to verify that the recommended improvements do solve the real problem.	p298 step 3
Check impact of DQ improvements	Verify the effectiveness of DQ improvement actions.	p299 step 4
Act to standardise DQ improvements	Make DQ improvements a baseline habit.	p300 step 5

**Table 4: TQdM process-driven approach**

### **RCA in TDQM**

The well known TDQM methodology suggests viewing data as an information product (IP) that is manufactured through various data changing and storage processes. The steps of TDQM are shown in Table 5, and RCA sits between the measurement and improvement steps. The paper describing the approach is pitched at a high level of abstraction and therefore there is no detailed guidance for how to conduct RCA. Statistical process control (SPC), pattern recognition, pareto chart analysis, and introduce dummy records, are recommended as possible approaches that can be used. There is also no explicit mention of how RCA needs to be tailored to determine root causes of DQ problems.

Step	Description of Step	Link to study [48]
Define IP	Define the characteristics for the information product	Define IP p61
Measure IP	Develop and measure DQ metrics	Measure IP p64
Analyse IP	From the measurement results, identify root cause(s) of DQ problems	Analyze IP p64
Improve IP	Identify key areas for improvement such as: (1) aligning information flow and work flow with the corresponding information manufacturing system, and (2) realigning the key characteristics of the IP with business needs	Improve IP p65

**Table 5: TDQM steps**

### **RCA in SODQA**

In the SODQA methodology, similar to TDQM, the RCA step is between the DQ assessment and the actual improvement of DQ (Table 6). The results from the RCA are fed into the ‘improve DQ’ step, which executes the necessary DQ improvement actions. For both TDQM and SODQA, RCA appears centrally and directs the DQ improvement actions. RCA is therefore identified as being necessary, but no

specific guidelines are given as to the best practice methods of conducting RCA or whether it needs to be tailored to DQ.

Step	Description of Step	Link to study [32]
Perform a subjective and objective data quality assessment	The subjective and objective assessments of a specific DQ dimension are compared	p215 Assessments in Practice 1st bullet point
Determine root causes of discrepancies	Comparing the results of the assessments, identifying discrepancies, and determining root causes of discrepancies	p215 Assessments in Practice 2nd bullet point
Improve DQ	Determine what the improvement options are and taking necessary actions for improvement	p215 Assessments in Practice 3rd bullet point

**Table 6: SODQA steps**

### **RCA in DQFG**

RCA in the DQFG methodology is considered an important part to inform the development of solutions to the DQ problems, similar to the previously described methodologies above. On a more detailed level, the RCA step is positioned directly after setting up a DQ team (Table 7), and this is exactly the same as in TQdM. DQFG suggests that this team should be built with people who know and are involved with the ‘problem processes’ causing DQ problems, and that people who are close to the problem will be in the best position to determine root causes.

Step	Definition of Step	Link to study [38]
Select project	Identify and select a project that the DQ improvement will focus on.	p133 fig22.2 step 1
Form and charter project team	Setup a team of people to carry out the DQ project.	p133 fig22.2 step 2
Conduct root cause analysis	Find the root causes of the DQ problems	p133 fig22.2 step 3
Identify solutions to the DQ problem	Identify simple DQ improvement solutions as a starting point.	p133 fig22.2 step 4
Trial simple solutions to the DQ problem	Trial simple solutions with the aim of demonstrating that the solutions work before rolling out a full implementation.	p133 fig22.2 step 4
Implement solution	Roll out the solution once it has been proven successful	p133 fig22.2 step 5
Hold the gains	Confirm that the solution works and ensure that the problem does not recur	p133 fig22.2 step 6

**Table 7: DQFG steps**

### **RCA in EDQP**

The steps of the EDQP methodology are shown in Table 8. RCA in EDQP comes after both the assess DQ and assess business impact steps, which can be done in parallel and provide input to the root cause analysis step. As with the other methodologies, RCA is followed by the development and execution of the improvement plans. RCA features heavily in this methodology and the results of RCA are used in multiple steps. Similar to TQdM, this methodology contrasts a process approach with a data-driven approach, and the process-driven step (prevent future data errors) includes RCA, whereas the data-driven step (correct current data errors) does not. This methodology gives the most detailed guidelines and makes sensible suggestions on how to carry out three RCA methods for DQ problems. The three suggested methods are five whys, track and trace, and cause-effect diagrams. Five whys and the cause-effect diagrams are commonly used methods. Track and trace is defined as “a way to identify the location of the problem by tracking data through the information life cycle and determining root causes where the problem first appears” [30].

<b>Step</b>	<b>Description of step</b>	<b>Link to study [30]</b>
Define business need and approach	Determine why the DQ improvement is important to the business and plan the project	p57 fig 2.14 step 1
Analyse information environment	Understand the environment so that future steps benefit from increased knowledge of the context.	p57 fig 2.14 step 2
Assess DQ	Provide a picture of the actual quality of the data using suitable DQ dimensions	p57 fig 2.14 step 3
Assess business impact	Used to determine the impact of the DQ problems on the business	p57 fig 2.14 step 4
Identify root causes	Identify root cause(s) of DQ problems	p57 fig 2.14 step 5
Develop improvement plans	Develop alternative DQ improvement options/remedies. For example, an option might be to update the company database more frequently or distribute the updates to remote sites more often. Another option could be to perform data cleansing on the database at selected time intervals. These could be data-oriented or process-oriented approaches	p57 fig 2.14 step 6
Prevent future data errors	Implement appropriate solutions that address the root causes of the DQ problems	p57 fig 2.14 step 7
Correct current data errors	Implement solutions that correct the existing data errors	p57 fig 2.14 step 8
Implement controls	Implement ongoing monitoring and metrics, and verify the improvements that were implemented	p57 fig 2.14 step 9

**Table 8: EDQP steps**

### **RCA in CSDQ and AMEQ**

CSDQ is an extension of TDQM that incorporates a user-centered approach to improving the quality of customer support data; CSDQ therefore also includes the RCA step. This step has not been moved or modified and still sits between the analyse and improve steps of TDQM (see Figure 2 in [11]). As well as suggesting the use of cause-effect diagrams, the authors of this approach suggest a new method for RCA called the causes table (see Table 9). This is a simple method that aims to be lightweight and fast to ap-

ply. In the ‘caused by’ column of Table 9, a double tick indicates the most significant cause.

AMEQ also contains the same steps as TDQM, but does not expand RCA from the TDQM methodology or describe any methods to conduct RCA.

DATA QUALITY ATTRIBUTE	PRIMARY OR SECONDARY	ATTRIBUTE IMPORTANCE	CAUSED BY		
			People	Tools	Process
Believability	Secondary				
Accuracy	Primary	High	✓	✓✓	
Reputation	Secondary				
Relevancy	Secondary				
Value-added	Secondary				
Completeness	Primary	High	✓	✓	✓✓
Appropriate amount of data	Primary	Medium	✓✓	✓	✓
Interpretability	Secondary				
Ease of understanding	Secondary				
Consistent representation	Primary	Low		✓	
Accessibility	Primary	High		✓	

**Table 9: Causes table (reproduced from [11])**

## DISCUSSION/RESULTS

The above findings are used to answer the proposed research question as presented below:

To what extent is RCA emphasised in DQ improvement methodologies and frameworks?

In total, 7 data quality methodologies were selected as examples of the most influential methodologies. Among these DQ methodologies, RCA has been emphasised as an important activity to enable the improvement actions. Typically, RCA uses the DQ measurement results to identify root causes before informing what corrective improvement actions to implement. The Cause-effect diagram is suggested across many methodologies as the most common RCA method. It must be pointed out that in the TQdM, RCA is a critical part of the process-driven approach (to uncover the real causes of DQ problems).

Sub-question 2: Are there sufficient guidelines available that describe how to conduct DQ RCA?

In addition to the most commonly suggested cause-effect, some common RCA methods are also recommended including: Pareto chart analysis, Five whys, Cause-effect diagram, Statistical process control (SPC), Pattern recognition, Introduce dummy records, Track and trace and Causes table.

TQdM and EDQP provide relatively detailed guidelines and examples that relate to DQ. All the selected methodologies have not provided clarifications on how to determine and priorities the actual DQ root causes from the potentially large number of possible data quality causes found at each organisational level and function.

It must be noted that the current guidelines as summarised in the previous section are very generic, without being tailored specifically to the DQ problems. Whether or not this is required will be investigated in future research.

## CONCLUSION

This preliminary study has identified the most influential DQ methodologies and frameworks which highlight the importance of data quality RCA and it forms a key part of the DQ methodology. However, given that it is such an important area for DQ improvement, there has been little attention focussed on some critical aspects, including:

- Determining whether RCA in the DQ context is any different to other contexts where it has been used traditionally, and what issues this raises;
- Differentiating the generic problem context and the DQ problem context, which can make the selection of the appropriate RCA difficult (e.g. control chart may not be very useful for DQ problems);
- DQ problem related elements and aspects (e.g. different DQ problem scope may lead to different root causes (e.g. do certain DQ dimensions always lead to similar root causes, or require specific methods to be used?));
- Detailed guidelines on how DQ RCA should be conducted and what the root causes should look like;
- When to apply different methods (e.g. the selection of a RCA technique may result in different root causes).

This study has found that there are limited case studies on how organisations thoroughly conduct data quality RCA in real practices. Thus, future research will seek verifications on the above mentioned issues with industry case studies. It is felt that in any future academic study on data quality frameworks and methodologies, developing sufficient guidelines for RCA will enhance the research outcome and applicability.

## ACKNOWLEDGEMENTS

This research has received supported from EPSRC, project reference number EP/G038171/1.

## REFERENCES

- [1] Alkhattabi, M., Neagu, D., Cullen "An Information quality framework for e-learning systems." *Knowledge Management & E-Learning: An International Journal*, 2 (4). 2010. pp. 340-362.
- [2] Augustin, S., Reminger, B. "Trotz Datenflut jede Menge Informationsdefizit! - Ist das erfolgreiche JIT-Konzept auch in der Info-Welt realisierbar." *Der informierte Manager*, ed. H. Bäck. TÜV Rheinland, Köln, Germany. 1990. pp. 73-82.
- [3] Ballou, D. P., Wang, R., Pazer, H., Tayi, G. K. "Modeling information manufacturing systems to determine information product quality." *Management Science*, 44(4). 1998, pp. 462- 484.
- [4] Blechar, M., Friedman, T. *Improve Application and Data Quality via Application Development Best Practices*. Gartner. 21 November, 2005, Accessed on 11 June, 2012. Available at: <http://www.gartner.com/id=486846>
- [5] Baškarada, S., Koronios, A., Gao, J. "Towards a Capability Maturity Model for Information Quality Management: A TDQM Approach." *Eleventh International Conference on Information Quality*. Cambridge, MA, USA. 2006.
- [6] Batini, C., Cappiello, C., Francalanci, C., Maurino, A. "Methodologies for Data Quality Assessment and Improvement." *ACM Computing Surveys*, 41 (3). 2009. pp.1–52.
- [7] Caballero, I., Piattini, M. "CALDEA: A Data Quality Model Based on Maturity Levels." *Third International Conference on Quality Software*. Dallas, Texas, USA. 2003. pp. 380-387.
- [8] Caro, A., Calero, C., Piattini, M. "A Portal Data Quality Model For Users And Developers." *12<sup>th</sup> International Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 2007. pp. 462-476.
- [9] Corradini, F., Hinkelmann, K., Polini, A., Polzonetti, B. "C2ST: A Quality Framework to Evaluate e-

- Government Service Delivery.” *8th International EGOV Conference*. Linz, Austria. 2009, pp. 74 – 84.
- [10] Davenport, T. *Information Ecology: Mastering the Information and Knowledge Environment*. Oxford University Press, Oxford. 1997.
- [11] English, L. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. John Wiley & Sons. 1999.
- [12] Eppler, M. J. “A Generic Framework for Information Quality in Knowledge-Intensive Processes.” *6th International Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 2001. pp.329-346.
- [13] Eppler, M. “Information oder Konfusion – Neue Kriterien für die betriebliche Kommunikation.” *io management*, 5. 1997. pp. 38-41.
- [14] Even, A., Shankaranarayanan, G. “Value-driven data quality assessment.” *10th International Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 2005. pp. 221-236.
- [15] Firth, C. P. “Data quality in practice: Experience from the frontline.” *Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 1996.
- [16] Friedman T. *Data quality methodologies: Blueprints for data quality success*. Gartner. 26 July, 2005. Accessed on 11 June, 2012. Available at: <http://www.gartner.com/id=483843>
- [17] Friedman, T. *Best Practices for Data Stewardship*. Gartner. 3 December, 2007. Accessed on 11 June, 2012. Available at: <http://www.gartner.com/id=554646>
- [18] Gano, D. L. *Apollo Root Cause Analysis: A New Way of Thinking*, 3<sup>rd</sup> Edition. Apollonian Publications, USA. 2008. Accessed on 25 June, 2012. Available at: [http://www.realitycharting.com/public/site/files/pdf/ARCA\\_Appendix.pdf](http://www.realitycharting.com/public/site/files/pdf/ARCA_Appendix.pdf)
- [19] Gao, J., Koronios, A., Kennet, S., Scott, H. “Business Rule Discovery through Data Mining Methods”. *Engineering Asset Management Review: Definitions, Concepts and Scope of Engineering Asset Management*, ed. J.E. Amadi-Echendu, K. Brown, R. Willet & J. Mathew, Springer, London. 2010. pp. 159-162.
- [20] Giannoccaro, A., Shanks, G., Darke, P. “Stakeholder perceptions of data quality in a data warehouse environment.” *Australian Computer Journal*, 31(4). 1999. pp. 110-117.
- [21] Harris, K., Fleming, M. *KM and Content Quality: What Can You Trust*. Gartner Group Research Note. 29 December, 1998.
- [22] Jarke, M., Jeusfeld, M.A., Quix, C., Vassiliadis, P. “Architecture and Quality in Data Warehouses.” *10th International Conference on Advanced Information Systems Engineering*. Pisa, Italy. 1998. pp. I93-113.
- [23] Kahn, B. K., Strong, D. M. “Product and Service Performance Model for Information Quality: An Update.” *Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 1998.
- [24] Kahn, B., Strong, D.M., Wang, R.Y. “Information Quality Benchmarks: Product and Service Performance.” *Communications of the ACM*, 45(4). 2002. pp. 184-192.
- [25] Keenan, S.L., Simmons, T. “CSDQ: A User-Centered Approach to Improving the Quality of Customer Support Data.” *10th International Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 2005.
- [26] Klein, A. “Incorporating quality aspects in sensor data streams.” *First Ph.D. Workshop in conjunction with the Sixteenth ACM Conference on Information and Knowledge Management*. Lisbon, Portugal. 2007. pp: 77-84.
- [27] Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R.Y. “AIMQ: A Methodology for Information Quality Assessment.” *Information & Management*, 40(2). 2002. pp. 133-146.
- [28] Lesca, H., Lesca, E. *Gestion de l'information, qualité de l'information et performances de l'entreprise*. Paris. 1995.
- [29] Levitin, A.V., Redman, T. C. “Data as a resource: properties, implications, and prescriptions.” *Sloan Management Review*, 40(1). 1998. pp. 89-101.

- [30] Loshin, D. *Enterprise Knowledge Management: The Data Quality Approach*. Morgan Kaufmann Pub. 2001.
- [31] Madnick, S., Wang, R. Y., Zhang, W. A. "Framework for Corporate Householding." *Seventh International Conference on Information Quality*. MIT, Cambridge, Massachusetts, USA. 2002. pp. 36-40.
- [32] McGilvray, D. *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. Morgan Kaufmann. 2008.
- [33] Miller, H. "The multiple dimensions of information quality." *Information Systems Management*, 13(2). 1996. pp.79-82.
- [34] Naumann, F., Freytag, J., Leser, U. "Completeness of Integrated Information Sources." *Information Systems*, 29 (7). 2004. pp. 583-615.
- [35] Nauman, F., Roth, M. "Information Quality: How good are off-the-shelf DBMS." *Ninth International Conference on Information Quality*. ed. S. Chengular-Smith, L. Raschid, J. Long & C. Seko , MIT, Cambridge, Massachusetts, USA. 2004. pp. 260-274.
- [36] Nemani, R. R., Konda, R. "A Framework for Data Quality in Data Warehousing." *Third International United Information Systems Conference*, 20(1). Sydney, Australia. 2009, pp. 292-297.
- [37] Parker, M.B., Moleshe, V., De la Harpe, R. Wills, G.B. "An evaluation of information quality frameworks for the world wide web." *8th Annual Conference on WWW Applications*. Bloemfontein, Free State Province, South Africa. 2006.
- [38] Pipino, L.L., Lee, Y.W., Wang, R.Y. "Data Quality Assessment." *Communications of the ACM*, 45 (4). 2002. pp.211–218.
- [39] Price R.J., Shanks, G. "A semiotic information quality framework." *IFIP International Conference on Decision Support Systems: Decision Support in an Uncertain and Complex World*. ed. R. Meredith, G. Shanks, D. Arnott and S. Carlsson, Prato, Italy. 2004. pp. 658-672.
- [40] Redman, T. C. *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business School Press, Boston. 2008.
- [41] Redman, T. C. *Data quality for the information age*, 1st Edition. Artech House, Boston, MA. 1996.
- [42] Redman, T.C. *Data Quality: The Field Guide*. Digital Press, Boston. 2001.
- [43] Russ-Mohl, S. *Der I-Faktor*. Osnabrück, Fromm. 1994.
- [44] Shanks, G., Darke, P. "Understanding Data Quality in Data Warehousing: A Semiotic Approach." *Third International Conference on Information Quality*. ed. I. Chengilar-Smith and L. Pipino, MIT, Boston. 1998. pp. 247-264.
- [45] Strong, D. M., Lee, Y. W., Wang, R. Y. "Data quality in context." *Communications of the ACM*, 40 (1). 1997. pp. 103-110.
- [46] Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C. "A Framework for Information Quality Assessment." *Journal of the American Society for Information Science and Technology*, 58(12). 2007. pp 1720 -1733.
- [47] Wand, Y., Wang, R.Y. "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM*, 39 (11). 1996. pp. 86-95.
- [48] Wang, R.Y. "A Product Perspective on Total Data Quality Management." *Communications of the ACM*, 41 (2). 1998. pp.58–65.
- [49] Wang, R.Y., Storey, V., Firth, C. "A framework for analysis of data quality research." *IEEE Transactions on Knowledge and Data Engineering*, 7(4). 1995. pp. 623-640.
- [50] Wang, R. Y., Strong, D. M. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems*, 12 (4). 1996. pp. 5-34.

# IMPACT OF CONCEPTUAL MODELING APPROACHES ON INFORMATION QUALITY: THEORY AND EMPIRICAL EVIDENCE

(Poster)

**Roman Lukyanenko**

Memorial University of Newfoundland, St. John's Canada

[roman.lukyanenko@mun.ca](mailto:roman.lukyanenko@mun.ca)

**Jeffrey Parsons**

Memorial University of Newfoundland, St. John's Canada

[jeffreyp@mun.ca](mailto:jeffreyp@mun.ca)

**Abstract:** The current work investigates the impact of conceptual modeling on information quality. We develop a theoretical model of the impact conceptual modeling approaches make on quality dimensions, including accuracy, completeness and timeliness. We then propose quality-driven principles of instance-based conceptual modeling. The advanced propositions will be evaluated using design science methodology. Empirical evidence is expected to produce a compelling argument for incorporating information quality considerations into conceptual modeling.

**Key Words:** Data Quality, Information Quality, Conceptual Modeling, Ontology, Cognition, Design Science.

## INTRODUCTION

Representing reality, as perceived by focal individuals or collectives, is considered “the primary purpose of information systems” [21, p. 208]. Information systems (IS) then make it possible to *draw inferences* about reality by querying the information base of the IS as opposed to having to directly observe objects in the domain [21].

With the proliferation of IS, much of what we know about reality is based on IS-mediated information as opposed to direct observation and evidence. In this context, we consider *information quality (IQ)* to be the degree to which an IS affords *valid inferences* about the underlying world [see 20]. Considering that direct access to reality is missing, IQ can be further treated as the difference between user perceptions and models of reality and those stored in an IS. Thinking about IQ from *data creator* or *representational* perspective leads to important practical questions: What is the impact on IQ of the process by which an IS representation of reality is created? How can the inferences we draw from an IS be more faithful to the underlying reality? Motivated by these questions, we examine the impact of the process of representing reality, or *conceptual modeling*, on IQ.

Conceptual models are informal or formal, and usually diagrammatic, representations of domain semantics. Conceptual models document system requirements, promote domain understanding and support communication between developers and users [11, 12, 22]. They also guide database and application design, and have a strong impact on information collection and storage. It is widely contended that conceptual models increase the effectiveness and efficiency of IS development [9, 22].

There is a growing awareness that many important IQ issues depend upon, and can be resolved by, changing approaches to conceptual modeling [6, 7]. Nevertheless, little is known about specific ways in

which conceptual modeling decisions translate into dimensions of IQ [7]. Although research has stressed the vital relationship between analysis and design of an IS and quality of stored data [5, 20], typically discussions of the impact of modeling on IQ examine best practices of requirements engineering and conceptual modeling, such as implementing integrity constraints and business rules [5, 12, 17]. Relatively few studies, however, have explored these issues using a theoretical perspective, and a theoretical understanding of *how* and *why* conceptual modeling approaches impact IQ is limited.

In a seminal theoretical work on IQ, Wand and Wang [20] drawn upon ontological theory to examine the extent to which an IS permits mapping of lawful states of reality to states of the IS. Wand and Wang, however, do not specifically consider conceptual modeling grammars or methods. Recently, Lukyanenko and Parsons [6, 7] employed ontological and cognitive theories to derive negative consequences of class-based conceptual models on IQ. They argued *property loss* necessarily arises from the prevalent practice to model instances as members of classes. This line of work challenges the assumption that modeling IS following prevalent conceptual modeling approaches promotes higher IQ.

The current work aims to increase our understanding of the impact of conceptual modeling on IQ. We develop a theoretical model of the relationship between conceptual modeling approaches and IQ dimensions. We extend previous research to examine the broader impact of abstraction-based representations (including class-based modeling [6, 7]) on accuracy, completeness and timeliness. We then derive an alternative instance-based representation that avoids negative consequences of abstraction-based representations. Finally, we evaluate the advanced propositions using a real IS artifact following design science methodology.

## IMPACT OF CONCEPTUAL MODELING ON IQ

Traditionally, conceptual modeling research assumes a corporate environment, where corporate users or customers are important sources of subject-matter expertise and system requirements. In such a setting, close contact with users provides an opportunity to resolve conflicts in individual views and conceptualizations: any “conflict must be solved through communication among people” [16, p. 250]. A final conceptual model, therefore, typically represents a global, integrated view of a domain but often does not represent the view of any individual user [13]. The global conceptual model then serves as the basis to establish understanding and attach consistent meaning to domain phenomena. The prevailing representation method in conceptual modeling is abstraction [11]. Abstraction enables analysts to deliberately ignore the many individual differences among domain phenomena and represent only *relevant* information for *specific functionalities* of intended systems. The popular Entity-Relationship grammar, for example, uses classes (entity types), relationships, and attributes (or properties) to represent reality [2]. Classes (e.g., *customer*) abstract from differences among instances (e.g., a *particular customer*) and capture perceived equivalencies among them. While abstraction-based conceptual models promote efficient domain representation, they also engender IQ deficiencies. Despite all efforts to reconcile individual user perspectives, each individual user may continue to maintain unique conceptualization of reality even after the process of view discovery and reconciliation is finished. Moreover, in nascent domains that encourage broad user participation, such as social media and crowdsourcing (engaging general public to work on specific tasks, see [3]), discovering and reconciling individual user conceptualizations appears unrealistic. Furthermore, new experiences alter existing user models forcing them to evolve, sometimes considerably. For example, with the emergence of online publishing, new attributes of *books* become pertinent (e.g., *digital size*, *encoding format*). According to cognitive psychology, differences in prior experiences, domain expertise, conceptualizations, and ad hoc utility, result in different abstractions of the same domain between contributors and for the same contributor over time [8, 10, 19]. In a sufficiently rich domain (e.g., a typical business environment) achieving a universal agreement among all users on how to organize knowledge in a domain is infeasible. Below, we examine the impact of abstraction-based

conceptual models on central dimensions of IQ, including accuracy, completeness, and timeliness [17].

**Accuracy.** Typically a data contributor is unable to change the way information is collected and stored. A potential mismatch between abstractions maintained by a user and those defined in the IS may lead to an incorrect data entry. For example, when forced to accept a predefined class (e.g., type of product), a contributor may choose the “wrong one.” Since databases typically do not store details on user deliberation process, a data consumer may take records at face value and make decision based on inaccurate data.

**Information loss (completeness).** Abstraction-based models engender *information loss* resulting from the failure to capture all pertinent properties of instances in reality. Ontologically, every instance is unique by the virtue of having unique properties [1]. This means, for example, storing instances in terms of classes (which abstract instance similarities) may preclude some potentially valuable properties from being stored [6, 7]. An extreme, but common, example of information loss is selecting *other* when classifying phenomena or reporting attributes: many rich and potentially useful inferences are lost when dissimilar objects are not properly differentiated. Information loss constrains many business intelligence opportunities and precludes discoveries of unanticipated phenomena.

**Completeness** is undermined for another important reason. A mismatch between abstract models of contributors and those defined in the IS may force some users to avoid contributing information. Users may be apprehensive to accept potentially incorrect data (e.g., an unfamiliar attribute), or even be disappointed with the gulf between own model and the IS one and avoid contributing. While this may appear relevant to volitional use (e.g., social media), many non-discretionary corporate IS contain optional sections that may be underutilized for the same underlying reasons.

**Timeliness.** Abstract models impact *data timeliness* due to the requirement to satisfy abstraction-driven constraints necessary to commit a transaction. For example, a class is typically modeled as a set of attributes [14, 15]. Users possessing a potentially valid, but incomplete (as per class definition) set of attributes, may not be able to contribute *until* all mandatory attributes are available. Such information may be considered invalid according to an IS, but may be perceived valid according to a user, or be valid for a different purpose (e.g., a person without a SSN and *Drivers License* may not qualify to be a *customer*, but still be of interest to marketing and sales).

A key realization is any abstraction is exogenous to the underlying reality and the reality can never conform fully to generalized models. Imposing such models *a priori* may undermine the ability to convey relevant aspects about domains accurately, completely and timely. This does not imply that the imposition of rules is to be proscribed. Data consumers may require certain information for safety, legal, accounting or other valid considerations, but the representational aspect of IQ may suffer. A number of factors moderate the impact of modeling on IQ. One important variable is domain complexity (e.g., number of unique discernable features of objects in a domain). The greater the domain complexity, the more likely a model misalignment will occur. The degree of consensus among users on how to organize reality is also pertinent. Some tightly-knit collectives (e.g., groups of likeminded employees) may be able to maintain sufficient levels of shared domain understanding to minimize the negative impacts described above. In contrast, in many emerging crowdsourcing or social media projects there are no constraints on who can participate, and many users may have incongruent views or conceptualizations of issues in a domain.

Different types of abstractions (e.g., classes, attributes, relationships, hierarchies) have varying impact on IQ. Classes, for example, have a profound impact being the primary mechanisms “for imposing a structure on the data requirements for an information system” [15, p. 840]. Other constructs (e.g., attributes

and relationships) are typically defined with respect to classes [7, 14, 15]. Other moderating factors may include the degree of discretion in IS use, the ability of users to interact with original designers, and the availability of documentation that users can consult to interpret IS models.

Given the inherent negative impact of prevalent conceptual modeling approaches on IQ, a critical question is how to mitigate these consequences. Here we consider an approach to conceptual modeling that uses instance-based representation. To derive this representation, we turn to fundamental theories about what exists in reality (ontology). We use the general ontology of Bunge [1] to specify fundamental elements of existence assumed to be *observer independent*.

According to Bunge [1] the world consists of “things” (or *instances*), elementary and observer independent ontological constructs. Every instance is unique as it has distinct properties. Properties are attached to instances and cannot exist without them. Properties can be *intrinsic* to things (e.g., *age*) or *mutual* if they belong to multiple things (e.g., *date hired* is a joint property of a person and a company). The change of instance properties over time allows to model system dynamics (e.g., events, transactions).

Bunge’s ontology can be used to develop quality-driven conceptual modeling principles and grammars. Instances can be modeled directly by allowing different potential users to report attributes of instances in a domain. This conceptual model can guide instance-based database design [14] and a data collection interface that allows users to report instances and attributes free of abstraction-based constraints. We contend that this modeling approach should promote capturing original user input (and hence, perceived reality) more faithfully.

## PROPOSED EMPIRICAL WORK

To empirically evaluate the impact of conceptual modeling on IQ, we develop hypotheses for each IQ dimension above (e.g., Completeness Hypothesis: An IS consistent with a predefined abstraction-based model will capture significantly fewer user contributions than an IS implementing the instance-based principles). We plan to build a real IS artifact – an online natural history *citizen science* website. Online citizen science is a type of crowdsourcing that engages general public in scientific research [4, 18]. The scientific community is increasingly leveraging contributions from citizens to expand the scope of available information and reduce research costs. IQ in citizen science projects is critical for any meaningful use of citizen science data in research. The objective of the project will be mapping biodiversity of a region in North America (a territory of over 150,000 square miles). The project will be carried out in partnership with biology experts and wildlife authorities. Their expertise will be leveraged in IS development (e.g., building and verifying conceptual models) and in evaluating quality of user contributions. Upon the launch of the project, information contributors (citizen scientists) will be randomly assigned to two data collection interfaces (and underlying conceptual models). In the abstraction-driven interface, users will be reporting sightings of plants and animals using a prevalent class-based approach to data collection. In the instance-based condition, users will be asked to report attributes of an observed instance. We intend to collect one year of observations and expect to have several thousand data points in each condition (a series of pre-tests and a laboratory experiment have been completed signalling viability and informing design strategy for the project).

The results of data collection from both conditions will be compared along the accuracy, information loss, timeliness and completeness dimensions. We also plan to evaluate the *fitness* of the instance and attribute-based data *for use* in biology and biodiversity management. Analyzing IQ in a real setting should enhance validity and afford a robust test of the advanced conceptualizations.

## CONCLUSIONS AND EXPECTED CONTRIBUTIONS

This paper describes and explains the relationship between conceptual modeling approaches and IQ. We examine the negative impact of prevalent modeling approaches on accuracy, completeness, and timeliness of data. We then use fundamental theories of ontology to derive quality-driven principles of instance-based conceptual modeling. Empirical evidence supporting the proposed principles is expected to produce a compelling argument for incorporating quality principles into conceptual modeling.

This work is expected to have important implications for research and practice. It is widely contended that deficient conceptual models lead to unnecessary costs and failures [22]. Similarly, IQ deficiencies entail significant societal and economic losses [17, 23]. By deeply grounding IQ management in conceptual modeling, we hope to improve the quality of inferences about reality drawn from information systems.

## REFERENCES

- [1] Bunge, M. *Treatise on basic philosophy: Ontology I: the furniture of the world*. Reidel. Boston, MA, 1977.
- [2] Chen, P. "The entity-relationship model - toward a unified view of data." *ACM Transactions on Database Systems*, 1 (1). 1976. pp. 9-36.
- [3] Doan, A., R. Ramakrishnan and A. Y. Halevy "Crowdsourcing systems on the World-Wide Web." *Communications of the ACM*, 54 (4). 2011. pp. 86-96.
- [4] Goodchild, M. "Citizens as sensors: the world of volunteered geography." *GeoJournal*, 69 (4). 2007. pp. 211-221.
- [5] Hoxmeier, J. A. "Typology of database quality factors." *Software Quality Journal*, 7 (3). 1998. pp. 179-193.
- [6] Lukyanenko, R. and J. Parsons. Information Loss in the Era of User-Generated Data. In *Proceedings of the pre-ICIS SIG IQ*. (Shanghai, China). 2011.
- [7] Lukyanenko, R. and J. Parsons. Rethinking data quality as an outcome of conceptual modeling choices. In *Proceedings of the 16th International Conference on Information Quality*. (Adelaide, Australia). 2011.
- [8] McCloskey, M. and S. Glucksberg "Natural categories: Well defined or fuzzy sets?" *Memory & Cognition*, 6 (4). 1978. pp. 462-472.
- [9] Moody, D. L. "Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions." *Data & Knowledge Engineering*, 55 (3). 2005. pp. 243-276.
- [10] Murphy, G. L. *The big book of concepts*. MIT Press. Cambridge, Mass., 2004.
- [11] Mylopoulos, J. "Information modeling in the time of the revolution." *Information Systems*, 23 (3-4). 1998. pp. 127-155.
- [12] Olivé, A. *Conceptual modeling of information systems*. Springer. Berlin Heidelberg New York, 2007.
- [13] Parsons, J. "Effects of Local Versus Global Schema Diagrams on Verification and Communication in Conceptual Data Modeling." *Journal of Management Information Systems*, 19 (3). 2003. pp. 155 - 184.
- [14] Parsons, J. and Y. Wand "Emancipating Instances from the Tyranny of Classes in Information Modeling." *ACM Transactions on Database Systems*, 25 (2). 2000. pp. 228-268.
- [15] Parsons, J. and Y. Wand "Using cognitive principles to guide classification in information systems modeling." *MIS Quarterly*, 32 (4). 2008. pp. 839-868.
- [16] Pohl, K. "The three dimensions of requirements engineering: a framework and its applications." *Information Systems*, 19 (3). 1994. pp. 243-258.
- [17] Redman, T. C. *Data quality for the information age*. Artech House. Norwood, MA, 1996.
- [18] Silvertown, J. "A new dawn for citizen science." *Trends in Ecology & Evolution*, 24 (9). 2009. pp.

467-471.

- [19] Smith, L. B. *Emerging ideas about categories*. L. Erlbaum Associates, 2005.
- [20] Wand, Y. and R. Y. Wang "Anchoring data quality dimensions in ontological foundations." *Communications of the ACM*, 39 (11). 1996. pp. 86-95.
- [21] Wand, Y. and R. Weber "On the Deep-Structure of Information-Systems." *Information Systems Journal*, 5 (3). 1995. pp. 203-223.
- [22] Wand, Y. and R. Weber "Research commentary: Information systems and conceptual modeling - A research agenda." *Information Systems Research*, 13 (4). 2002. pp. 363-376.
- [23] Wang, R. Y. and D. M. Strong "Beyond accuracy: what data quality means to data consumers." *Journal of Management Information Systems*, 12 (4). 1996. pp. 5-33.

# **THE MANY FACES OF INFORMATION AND THEIR IMPACT ON INFORMATION QUALITY**

(Research paper)

**Carlo Batini**

**Matteo Palmonari**

**Gianluigi Viscusi**

University of Milano-Bicocca

Department of Informatics, Systems and Communication (DISCo), Italy

{baltini, palmonari, viscusi}@disco.unimib.it

**Abstract:** In this paper we discuss the main issues considered in data and information quality and several factors influencing these issues. The main goal of the paper is exploratory, aiming to identify basic or key issues characterizing information quality (IQ) research and their impact on future information quality research perspectives in a context where information is increasingly diverse and represented according to several data models. The investigation considers several relevant topics related to data and information representation, access, and usage. We conclude the paper by discussing how philosophical studies on knowledge and truth can contribute to a better understanding of some key foundational problems that emerged in our analysis.

**Key Words:** Data Quality, Information Quality, Philosophy

## **INTRODUCTION**

In the last decades, information systems of public or private organizations have been migrating from a hierarchical/monolithic to a network-based structure, where the potential information sources that single organizations or networks of cooperating organizations can use for the purpose of their activity is dramatically increased in size and scope. At the same time data representations have evolved from structured data, to semi-structured and unstructured text, to maps, images, videos and sounds. The data & information quality issue, which concerns the capability to define, model, measure and improve the quality of data and information that are exchanged and used in everyday life, in business processes of firms and administrative processes of public administrations, is becoming critical for human beings and organizations all over the word. Despite the relevance of the quality of information assets, the growing literature on data and information quality, and early conceptualizations of the main constructs and dimensions of the data and information quality research fields [42], it is our believe that a further clarification and formalization of their main concepts are required, as also pointed out at by the authors at the Information quality symposium at AISB/IACAP World Congress (Birmingham, UK, 2nd-6th July 2012).

As for these issues, the main goal of this paper is exploratory, aiming to discuss key issues in information quality (IQ) in the context of data represented according to different and very heterogeneous data models and formats. The discussion we present in this paper emerged after several studies on topics relevant to IQ and related to data and information representation, access, and usage, including: the quality of data (see [6]), the quality of information (see [7,9]) the quality of conceptual schemas and ontologies (see [10]), the quality of scientific data (big data) [11]; conceptual dependencies between data quality dimensions [3]; conception and comparison of methodologies for data quality assessment and improvement (see again [6]) and conception of methodologies for joint data and information quality assessment and improvement [9]; conceptual modelling for data base design (see [5]).

We organize the discussion as follows. First, we introduce three basic issues that have been addressed in

IQ research so far - namely, definitions of IQ, specific quality dimensions, and IQ dimension classifications - and several factors that significantly influence IQ. Then we discuss the impact of the considered influencing factors on each basic issue. Finally, we discuss an issue that emerges throughout our analysis and deserves a particular attention when we consider information systems dealing with an increasing amount of non-structured information: focusing on ontologies as particularly flexible information organization structures, we discuss the impact that the flexibility characterizing an information representation model has on IQ. A discussion of the results concludes the paper.

## **IQ RESEARCH: BASIC ISSUES AND INFLUENCING FACTORS**

For sake of clarity, we adopt the following convention in this paper: when we refer to *data quality*, we refer to quality of structured data, when we refer to *information quality*, we consider types of data represented according to different heterogeneous models, such as semi-structured data, texts, drawings, maps, images, videos, sounds, etc. This pragmatic distinction also reflects a common use of these terms in the technical literature.

When attempting to formalize the concept of data quality, the first issue concerns the concepts of data and quality. Traditionally, international standard bodies are authoritative and knowledgeable institutions when definitional and classification issues are considered. ISO has enacted in 2008 the standard ISO/IEC 25012:2008 (see [23]), that defines data quality as the “*degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions*”, and provides “*a general data quality model for data retained in a structured format within a computer system*”. When we look at the definitions of data and information proposed in the document, we discover that i) *data* is defined as “*reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing*”; ii) *information* is defined as “*information-processing knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that within a certain context have a particular meaning*”.

This choice is specular to the usual one in textbooks and scientific papers, where information is defined in terms of data (see e.g. [18]), and knowledge in terms of information in some definitions (e.g. in [30]). The ISO effort shows severe limitations, such as the flat classification adopted among characteristics, that contradicts e.g. the classification provided in the document “ISO/IEC 9126 Software engineering — Product quality, an international standard for the evaluation of software quality”, where quality characteristics are expressed in terms of sub-characteristics; furthermore, several characteristics (e.g., completeness) depend on the model adopted for data representation, even though this dependence is not explicitly discussed, and data are organized in models that neatly distinguish between instances and schemas are considered, e.g. the relational model, while schemaless data, such as e.g. textual documents, are ignored; finally, there is no attempt to distinguish between different types of data and information, from structured data to texts and images.

Furthermore, when attempting to formalize the concept of data and information quality (IQ), it is of primary importance to define a set of research coordinates. We distinguish in the following between basic issues and influencing factors. Basic issues are:

*BI1. Definitions of IQ - How many different definitions exist of information quality?*

*BI2. IQ Dimensions - How many dimensions are considered to capture the multifaceted character of the concept of IQ?*

*BI3. IQ dimension classifications – In how many ways dimensions can be classified?*

We now introduce a non-exhaustive list of factors influencing IQ:

*IF1. Type of information representation* - Types of information representation investigated in [7, 8, 10], have been: graphical representations of conceptual schemas, maps, images, from one side, emphasizing the visual perceptual character of information, and structured, semi-structured, unstructured type of text, emphasizing the linguistic character of information; a specific type of semi-structured text has been

considered, laws.

*IF2. Life cycle of information* – Information has usually a life cycle, made of acquisition (or imaging), validation, processing, exchange, rendering and diffusion. Does the life cycle of the different types of information representations influence IQ?

*IF3. Type of information system* – Information system architectures have evolved from hierarchical systems, where the information is highly controlled, to distributed, cooperative, peer to peer, web based information, where information flows are anarchic and undisciplined. How this evolution has influenced IQ?

*IF4. Level of semantic constraints: binding vs freedom in coupling data and schemas and open vs closed world assumption* – Data can undergo different levels of semantic constraints. In databases data and schemas are tightly coupled, while other data, e.g. RDF data, can be loosely coupled with schema level constraints by means of metadata. Moreover, in data bases the closed world assumption (CWA) usually holds, meaning that any statement that is not known to be true is false. In knowledge bases, the open world assumption (OWA) states that any statement that is not known, cannot be predicated neither true nor false. Do the binding/freedom in coupling schemas and data and CWA/OWA influence IQ?

*IF5. Syntax vs semantics* – How the syntax vs the semantics of information play a role in IQ?

*IF6. Objective vs subjective assessment of IQ* – With the term subjective we mean “evaluated by human beings”, while the term objective means “evaluated by a measurement performed on real world phenomena”. How the objective vs subjective quality evaluation is related with IQ?

*IF7. Influence of the observer* - How IQ is influenced by the observer/receiver, human being vs machine?

*IF8. Influence of the task* - IQ is intrinsic to information or it is influenced by the application/task/context in which information is used?

*IF9. Topological/geometrical/metric space in visually perceived information* – How the different spaces influence IQ?

*IF10. Level of abstraction of information represented* – The same real world phenomenon can be represented at different levels of abstraction (see [4] where levels of abstractions are defined for conceptual database schemas).

IQ is a relatively new discipline in information sciences. As a consequence, a discussion on above basic issues and influencing factors can be made at the state of the art in terms of examples and counterexamples leading to observations, statements, conjectures that cannot be formally stated and validated. Conscious of these limitations and immaturity, in the rest of the paper we proceed discussing (some) basic issues, influencing factors and relevant relationships between them.

## DEFINITIONS OF IQ

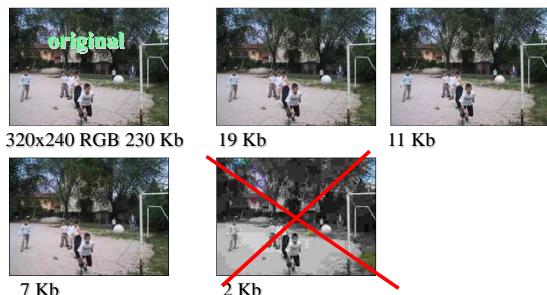
We first deal with one of the most controversial questions around IQ: is there an *intrinsic information quality*? Look at Figure 1. Before reading the next paragraph, reply to this question: which is the most accurate/ faithful image of Mars?



Figure 1: Two pictures of Mars.

The first image has been downloaded from a blog, the second was downloaded from the National

Aeronautics and Space Administration (NASA) site. Your judgments were probably based on your own model of Mars. Now that you have some ancillary data you could change your opinion. So, may we come to the conclusion that an intrinsic information quality does not exist? As another example, Figure 2 shows five different version of a photo, that make use of a decreasing number of dots per inch; looking at the 7Kb version, we consider acceptable the rendering of the image with respect to the original, while in the 2K case the resolution is not perceived as acceptable. So, we can conceive a concept of minimal amount of data needed to represent a piece of information over a threshold of reasonable quality. However we also observe that the context of use plays a role in defining this threshold; as an example, an image used as a web thumbnail is expected to be displayed at lower size (dpis and pixels) than the same image as a picture in a newspaper. The examples show that to predicate the quality of a piece of information, sometimes we need a reference version of the information, other times we evaluate the quality according to perceptual and/or technological characteristics of information, which depends on the type of information representation (IF1) (e.g., the image resolution, which can be measured subjectively or in terms of a metrics based on dots per inch). We want now to investigate more in depth (see Table 1) the relationship between definitions of IQ in the literature and corresponding influencing factors shown in column 1 of the table.



**Figure 2: Several representation of the same photo with decreasing amount of dots.**

Looking at columns, three different information representations are considered, a. structured data, b. images and c. a specific type of semi-structured text, laws.

IF1 Type of InfoR → Related issues/factors	Structured data	Images	Structured text: Laws
IF2/IF6 Absence of defects Adherence to the original		A perfect image should be free from all visible defects arising from digitalization and processing processes	
BI2 - Quality as a list of properties	1. High quality data is accurate, timely, meaningful, and complete 2. The degree of excellence of data. Factors contributing to data quality include: the data is stored according to their data types, the data is consistent, the data is not redundant, the data follows business rules, the data corresponds to established domains, the data is timely, the data is well understood		
IF6/IF7 Impression of the observer		Impression of its merit or excellence as perceived by an observer neither associated with the act of photography, nor closely involved with the subject matter depicted [III Association 2007]	
IF8 Fitness for use/ Adequacy to the task	Data are of high quality "if they are fit for their intended uses in operations, decision making and planning.	1. The perceptually weighted combination of significant attributes (contrast, graininess,...) of an image when considered in its marketplace or application 2. Degree of adequacy to its function/goal within a specific application field	Laws whose structure and performance approach those of "the ideal law": - It is simply stated and has a clear meaning - It is successful in achieving its objective - It interacts synergistically with other laws - It produces no harmful side effects - It imposes the least possible burdens on the people
Conformance...	...to requirements	of match of the acquired/reproduced image with IF2 the original → Fidelity IF7 viewer's internal references. → Naturalness	

Table 1: Definitions of IQ and related issues and factors mentioned in definition.

As an example of the absence of defects definition look at Figure 3. We can define the quality of the image as the lack of distortions or artifacts that reduce the accessibility of its information contents. Some of the most frequent artifacts considered are: blurriness, graininess, blackness, lack of contrast and lack of saturation. The definition referring to quality as a list of properties (BI2) is inspired by former contributions from the conceptual modeling research area [26]. Whereas the overall framework in Table 1 assumes the definition of data and information quality as based on the role of an information system as a representation [43], and the consequent distinction between the internal and external views of an information system [42]. The internal view is use-independent, supporting dimensions of quality as intrinsic to the data; while the external view considered the user view of the real world system (the observer perspective), where possible data deficiencies happen [43]. Moreover, it is worth noting that most of the research effort in the literature on data quality has provided by far greatest attention to the design and production processes involved in generating the data as the main sources of quality deficiencies [43]. Notice also that the definition more closely influenced by the observer (third row) claims for a “third party” subjective evaluation, not influenced by the domain.



Figure 3: Low readability.

Coming to the fourth row of Table 1, we see that fitness for use, that corresponds to IF9, Influence of the task, is the only common driving issue, while the impression of the observer (IF6) is typical of images, that are characterized by a high prevalence of subjective measures on objective ones (IF7). According to IF9, IQ can be expressed quantifying how it influences the performance of the task that uses it. Focusing on images, for example, in *medical imaging*, an image is of good quality if the resulting diagnosis is correct, in a *biometric system*, an image of a face is of good quality if the person can be reliably recognized, in an *optical character recognition system* a scanned document has a good quality if all the words can be correctly interpreted. As another context related to the influence of the task, see Figure 4. The image on the left is the true one (there is some fog in the parking area...), the image on the right is not accurate but it is certainly the most informative (or useful) for a driver that needs to know parking rules.



Figure 4: Two images of a parking sign board.

Finally we comment the conformance definition, that in case of images may be associated to the original, focusing in such a way on possible distortions during the processing life cycle (IF2), as a consequence subsuming the possibility to access to the original, or else may be associated to viewer's internal references (IF8). More in general, this last characteristic is typical of information representations such as images that may influence emotions of human being.

## IQ DIMENSIONS

Many possible dimensions and metrics can be conceived for IQ. In [6] several examples of synonyms and homonyms existing in the literature among dimensions are shown. We first discuss two dimensions among others, *accuracy* and *completeness*. As for accuracy, at the state of the art ([6]) two types of accuracy are considered, syntactic and semantic (IF5). As an example, we consider a set of Italian first names (Maria, Mario, Valerio, Carlo, Miriam), and compares them with the item "Mrio" that does not

correspond to any of them. Semantic accuracy of a value  $v$  can be intuitively defined as closeness of the value  $v$  to the true value  $v^*$ ; for a formal definition in the context of relational databases, the first order logic interpretation of the relational model can be adopted. Since semantic accuracy can be complex to measure and improve, a second type of accuracy, syntactic accuracy, measures the minimal distance between the value  $v$  and all possible values in the domain  $D$  of  $v$ . In our case, if we consider as distance the edit distance, the minimum number of character insertions, deletions, and replacements to convert “Mrio” to a string in the domain  $D$  of  $v$ , the syntactic accuracy of “Mario”, is 1. Notice that the string corresponding to “Mrio” is “Mario”, but it could be possible that two errors have occurred so that the true value of “Mrio” is “Maria”. To recognize this, we need more knowledge on the object represented by “Mrio”, e.g. that is a female. Another intriguing relationship to be investigated concerns accuracy and level of abstraction (IF10). Here we focus on maps. In our experience of visiting a city or making a travel by car, we need maps at different levels of detail. Cartographic generalization involves symbolizing data, and applying a set of techniques that convey the salient characteristics of that data. These techniques seek to give prominence to the essential qualities of the feature portrayed, e.g. that buildings retain their anthropogenic qualities – such as their angular form. In Figure 5 we show the same geographic area around the town of Lanvollon in France represented at three abstraction levels.

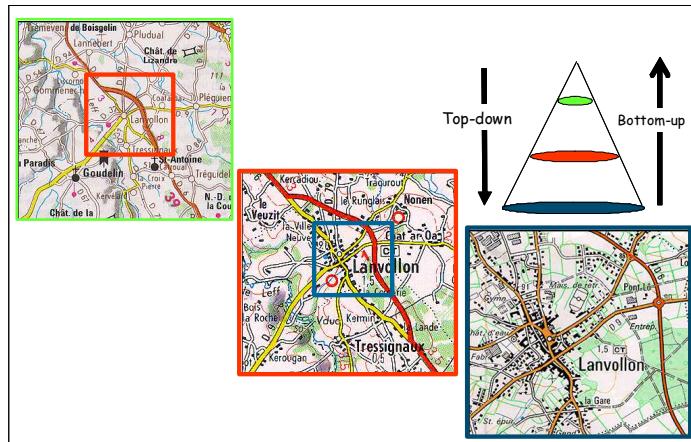


Figure 5: the same geographic area represented at three abstraction levels.

As said in [16], “Different combinations, amounts of application, and different orderings of these techniques can produce different yet aesthetically acceptable solutions. The focus is not on making changes to information contained in the database, but to solely focus upon avoiding ambiguity in the interpretation of the image. The process is one of compromise reflecting the long held view among cartographers that making maps involves telling small lies in order to tell the truth! “. These considerations show that even a dimension such as accuracy, that is considered only from the inherent point of view in the ISO standard, is strongly influenced by the context in which information is perceived/consumed. As for the completeness, its definition depends on the type of information representation (IF1), and is also influenced by the CWA/OWA (IF4). Let us consider a table reported with attributes Name, Surname, BirthDate, and Email. The table has four tuples. If the person represented by tuple 2 has no e-mail, tuple 2 is complete. If the person represented by tuple 3 has an e-mail, but its value is not known then tuple 3 presents incompleteness. Finally, if it is not known whether the person represented by tuple 4 has an e-mail or not, incompleteness may or may not occur, according to the two cases. Further, relation completeness, i.e., the number of tuples w.r.t. to the total number of individuals to be represented in the table, depends on the adoption of CWA or OWA. CWA is usually adopted in data bases; in this case, a relation is always complete. Instead semantic data are usually considered under OWA; if we adopt this assumption for our table, then we cannot compute

completeness, unless we introduce the concept of reference relation, for details see [6].

We now investigate the relationships between IQ dimensions and the evolution of types of information systems enabled by the evolution of ICT technologies. The shift from centralized and tightly coupled distributed systems to loosely coupled distributed and peer to peer systems, and from “controlled” sources to the unrestrainable web results both in bad and in good news from the point of view of IQ. From one side, the overall quality of the information that flows between networked information systems may rapidly degrade over time if both processes and their inputs are not themselves subject to quality control. On the other hand, the same networked information system offers new opportunities for IQ management, including the possibility of selecting sources with better IQ, and of comparing sources for the purpose of error localization and correction, thus facilitating the control and improvement of data quality in the system. Peer to Peer data management (P2P) Systems, typical of many application areas such as the ones found in the domain of biological databases, differently from centralized and strongly coupled distributed systems do not provide a global schema of the different sources. P2P systems are characterized by their openness, i.e. a peer can dynamically join or leave the system, and by the presence of mappings usually relating pairs of schemas. In P2P systems (and even more in the web) new quality dimensions and issues have to be considered such as *trustworthiness* and *provenance*. The evaluation of the trustworthiness (or confidence) of the data provided by a single peer is crucial because each source can in principle influence the final, integrated result. A common distinction is between the reputation of a source, which refers to the source as a whole, and the trust of provided data, e.g., the trust of the mapping that the source establishes with the other sources in a P2P system. While several trust and reputation systems have been proposed in the literature (see [21] for a survey), there is still the need to characterize the trust of a peer with respect to provided data and use such information in the query processing step. Effective methods for evaluating trust and reputation are needed, with the specific aim of supporting decisions to be taken on result selection. Information provenance describes how data is generated and evolves with time going on, which has many applications, including evaluation of quality, audit trail, replication recipes, citations, etc. Generally, the provenance could be recorded among multiple sources, or just within a single source. In other words, the derivation history of information could take place either at schema level (when defined), or at instance level. Even if significant research has been conducted, a lot of problems are still open. For the schema level, the most important are query rewriting and schema mappings including data provenance, and for the instance level, we mention relational data provenance, XML data provenance, streaming data provenance [12]. Moreover another important aspect to be investigated is dealing with uncertain information provenance for tracking the derivation of information and uncertainty.

Influencing factor IF4 deserves special attention in this context. As we anticipated in the introduction to this factor, different levels of semantic constraints can be imposed to data. In databases, data and schemas are tightly coupled; schemas pre-exist to data and control methods implemented by database management systems can enforce data to comply to the schema, which, even if poorly, defines their semantics. As an example, normal forms in relational databases are defined at the schema level, and are expressed in terms of properties of functional dependencies defined in relational schemas. A relational database whose relation schemas are in normal form, has relation instances free of redundancies and inconsistencies in updates, since every “fact” is represented only once in the database. The coupling of data and schemas in semi-structured data, e.g., data represented with languages such as XML, RDF, JASON [1], is way looser. Even when languages for semi-structured data are accompanied with languages for describing data schemas, e.g., XML-Schema for XML, RDFS and OWL2 for RDF [1], schemas are not required to pre-exist to data and the enforcement of the compliance of data to a schema at publishing time is weaker (it is left to the data publisher). Data in these cases are associated with schemas by means of annotation mechanisms. Finally, the use of metadata, e.g., based on folksonomies,

or other annotation schemes, can be seen as a way to associate data with schema-level information that provides data with semantics. However, the maximum freedom achieved by these representation approaches leads to a yet weaker coupling of data and schemas. As an example, let us focus on semantic data represented in RDF, which is also accompanied with expressive languages for the representation of schemas. A schema for RDF data can be defined by a RDFS vocabulary; however, there is no mechanism to enforce data to be compliant to the schema; even using reasoning, RDFS is not expressive enough to detect inconsistencies, because of its deductive semantics (the schema is used to make inference, not to constraint their meaning) and the lack of expressivity (concept disjointness and cardinality restrictions cannot be modeled in RDFS) [1]; although counterintuitive, inferences can be considered a measure of poor compliance between data and schemas [38], no inconsistencies can be detected, making a quality dimension such as *soundness* difficult to assess. In addition, the adoption of CWA or OWA has an influence on this discussion; OWA has an impact on the difficulty of defining and evaluating the compliance between data and schemas: a relation between two instances can hold even if the schema does not model such relation between the concepts the instances belong to; conversely, we cannot conclude that a relation between two concepts of different schemas does not hold because it is not represented in the data instances.

## IQ DIMENSION CLASSIFICATIONS

Several classifications of dimensions are considered in the literature, we shortly mention them, while their comparison is outside the scope of the paper. In [24] a two ways classification is proposed based on i) conforms to specification vs meets or exceeds consumer expectations (here we find an influence from IF6), and ii) product quality vs service quality. [39] proposes an empirical classification of data qualities, based on intrinsic, contextual, representations, accessibility qualities. The approach of [27], is based on the concept of evolutional data quality, where the data life cycle is seen as composed of four phases:

- *Collection*, data are captured using sensors, devices, etc.
- *Organization*, data are organized in a model/representation.
- *Presentation*, data are presented by means of a view/style model.
- *Application*, data are used according to an algorithm, method, heuristic, model, etc.

Qualities that in other approaches are generically attached to data, here are associated to specific phases, e.g. accuracy to collection, consistency to organization. A theory in [27] is a general designation for any technique, method, approach, or model that is employed during the data life cycle; for example, when data in the Organization phase is stored, a model is chosen, such as a relational or object-oriented model to guide the data organization. Due to the attachment of data to theories, when defining quality, we need to consider how data meet the specifications or serve the purposes of a theory. Such a concept of quality is called *theory-specific*. E.g. in the relational model, theory specific qualities are normal forms and referential integrity. In the following we adopt the classification in clusters of dimensions proposed in [7], where dimensions are empirically included in the same cluster according to perceived similarity. Clusters concern:

1. *Accuracy/correctness/precision* refer to the adherence to a given reference reality.
2. *Completeness/pertinence* refer to the capability to express all (and only) the relevant aspects of the reality of interest.
3. *Currency/volatility/timeliness* refer to the information up-to-dating.
4. *Minimality/redundancy/compactness* refer to the capability of expressing all the aspects of the reality of interest only once and with the minimal use of resources.
5. *Readability/comprehensibility/usability* refer to ease of understanding and fruition by users.
6. *Consistency/coherence* refer to the capability of the information to comply to all properties of the membership set (class, category,...) as well as to those of the sets of elements the reality of interest is in some relationship.

7. *Credibility/reputation*, information derives from an authoritative source.

In Table 2 we relate dimensions cited in the literature with dimension classifications (BI3) and the types of information representation (IF1) they are related to. Several dimensions in the table are associated to corresponding influencing criteria. We discuss some of them in the following.

1. *Accuracy* is often considered as an intrinsic IQ dimension (IF9), and its quality level is measured either by comparison with the “true” value (IF5, semantics) or else by comparison with a reference table (IF5, syntax).
2. *Accuracy* for structured data is defined both at the schema level and at the instance level, while for unstructured texts is defined at the instance level, with reference to a weaker property called *structural similarity* (IF4).
3. *Accuracy* for structured data has different metrics for different definition domains, e.g. last names of persons, usually made of one word item (e.g. Smith), or else names of businesses, that may involve several word items (e.g. AT&T Research Labs).
4. *Spatial accuracy* for maps refers to a bidimensional or tridimensional metric space (IF9).
5. Most definitions of *completeness* for structured relational data consider CWA, while completeness within the OWA is discussed in [6] (IF5).
6. *Consistency* for geographic maps is defined both in the topological space and in the geometric space (IF9).
7. *Cohesion and coherence* are proposed for unstructured texts. Both cohesion and coherence represent how words and concepts in a text are connected on particular levels of language, discourse and world knowledge. Cohesion is considered an objective property (IF6) of the explicit language/text, and is achieved by means of explicit linguistic devices that allow to express connections (relations) between words, sentences etc. These cohesive devices cue the reader on how to form a coherent representation. Coherence results from an interaction between text cohesion and the reader. The coherence relations are constructed in the mind of the reader (IF7) and depend on the skills and knowledge that the reader brings to the situation. Coherence is considered a characteristic of the reader’s mental representation, and as such is considered subjective (IF6). A particular level of cohesion may lead to a coherent mental representation from one reader but an incoherent representation for another (IF7).
8. *Diagrammatic readability* is usually expressed in terms of the achievement of several aesthetic criteria such as: a) minimize crossings; b) use only horizontal and vertical lines; c) minimize bends in lines; d) minimize the area of the diagram; e) place most important concept in the middle; f) place parent objects in generalization above child objects.

Quality Dimension Cluster	Structured data	Geographic Maps	Images	Unstructured Texts	Laws and legal frameworks
Correctness/ Accuracy/ Precision	<b>IF4</b> Schema accuracy w.r.t. requirements w.r.t. the model <b>IF4</b> Instance accuracy <b>IF5</b> Syntactic <b>IF5</b> Semantic <b>IF8</b> Domain dependent (ex. Last Names, etc.)	Instance <b>IF9</b> Spatial accuracy <ul style="list-style-type: none"> <li>- Relative/Absolute</li> <li>- Relative Inter layer</li> <li>- Locally increased r.a.</li> <li>- External/Internal</li> <li>- Neighbourhood a.</li> <li>- Vertical/Horizontal/Height</li> </ul> Attribute accuracy. <b>IF8</b> Domain dependent accuracy (ex. Traffic at critical intersections, Urban vs rural areas, etc.) Accuracy of raster representation	<b>IF8</b> Accuracy Syntactic Semantic "Reduced" semantic Genuineness Fidelity Naturalness Resolution Spatial resolution <b>IF2</b> Scan type	<b>IF8</b> Accuracy <b>IF5</b> Syntactic <b>IF5</b> Semantic <b>IF4</b> Structural similarity	Accuracy Precision Objectivity Integrity Correctness Reference accuracy
Completeness/ Pertinence	Schema Completeness Pertinence <b>IF5</b> Instance Value C., Tuple C., Column C., Relation C., Database C.	Completeness (btw different datasets) Pertinence	Completeness	Completeness	Objectivity Completeness
Temporal	Currency <b>IF8</b> Timeliness, Volatility	Recency/ Temporal accuracy/ Temporal resolution			
Minimality/ Redundancy/ Compactness/ Cost	Schema Minimality Redundancy	Redundancy	Minimality		For a law: Conciseness For a legal framework: Minimality, Redundancy
Consistency/ Coherence/ Interoperability	Instance Intrarelational Consistency Interrelational Consistency Interoperability	<b>IF9</b> Consistency Object consistency Geometric consist. Topological consist. Interoperability	Interoperability	<b>IF5</b> Cohesion Referential, Temporal, Locational, Causal, Structural <b>IF5</b> Coherence Lexical Nonlexical	Coherence Consistency among laws Consistency among legal frameworks
Readability/ Comprehensibility/Usability/ Usefulness Interpretability	Schema <b>IF7</b> - Diagrammatic Readability Compactness Normalization	Instance Readability/Legibility Clarity Aesthetics	<b>IF5</b> - Readability, Lightness, Brightness, Uniformityness, Sharpness, Hue chroma reproductionness Usefulness	<b>IF5</b> - Readability Comprehensibility <b>IF5</b> Cultural readability	<b>IF6</b> Clarity Simplicity

Table 2: Comparative analysis of quality dimensions for diverse information representations.

Notice that criteria a, b, c and d can be considered syntactic criteria, while e and f are semantic criteria (IF5). Applying such criteria to the two semantically equivalent Entity Relationship diagrams in Figure 6, we may come to the conclusion that the diagram on the right is more readable than the diagram on the left. This is not a universal conclusion, considering that one of the authors presented to visiting colleagues from another university the two diagrams, and the professors preferred the diagram on the left, claiming that they liked asymmetry and sense of movement (IF7).

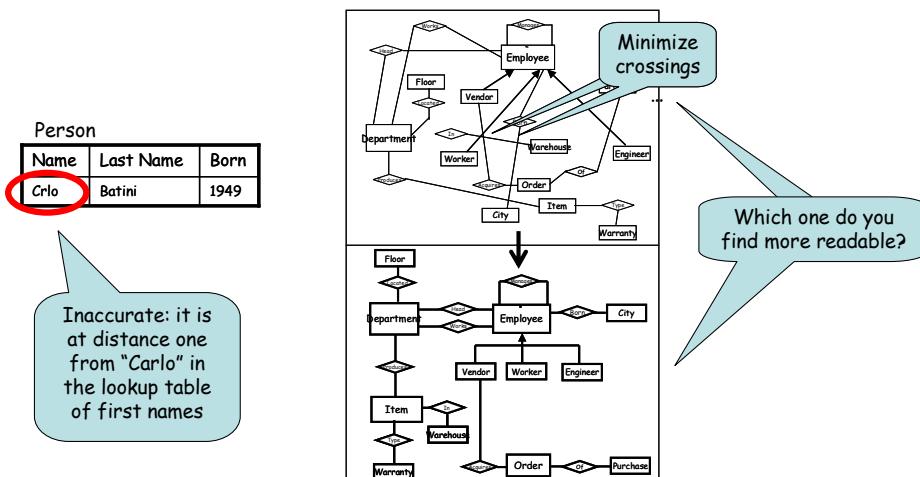


Figure 6: comparison of IQ measures for relational tables and diagrams.

9. Readability of unstructured texts and cultural accessibility refer to the readability/ comprehensibility

cluster. Readability is usually measured by using a mathematical formula that considers *syntactic features* of a given text, such as complex words and complex sentences, where e.g. complex words are evaluated on the basis of shallow syntax, such as number of syllables. *Cultural readability* refers to difficult (to understand) words, so they are related to the understanding of the word meaning, and as such can be considered more semantic oriented (IF6).

10. Concerning the relationship between IQ dimensions in the different representations vs objective/subjective measures (IF6), we have produced some figures in the past that confirm the validity of the following intuitive statement in the literature: the less the information is structured, from a restricted domain to a totally unstructured domain, the more subjective measures prevail on objective measures. Figure 6 represents two types of information representations, relational tables and diagrams, and three measures of IQ quality, respectively for *accuracy* of data for relational tables, and *readability* for diagrams addressed in previous point 8. Also in this case, objective measures can be conceived for diagrams, but to a certain extent, after that we have to deal with human being perceptions.

## THE IMPACT OF THE REPRESENTATION MODEL FLEXIBILITY ON IQ

So far, we have discussed several topics that are emerging in the IQ domain. However, the consideration of these topics appears significantly mediated by databases as main information representation technology. Historically, research on IQ in Computer Science addressed information represented in relational databases. Information in relational databases is organized in a well-structured manner and according to models with semantics that has a clear mathematical interpretation. Semantics associated to other types of (schemaless) data, e.g., maps, or tags associated to multimedia, is more difficult to define and to understand. Moreover, types of data other than structured data can be used in different ways depending on the application context. The more types of information are considered, and the more diverse and decentralized information management models and architectures are, the more we are in need of rethinking the perspective through which we look at information quality (in computer science).

Before generalizing our observation to other types of data, we can analyze some interesting IQ issues that have been considered when moving from data bases and data base schemas, to ontologies. These issues show some interesting research directions that can be applied to more types of data, and the role that diversity of information objects can play in IQ. Ontologies, and in particular Web ontologies, i.e. ontologies represented with formal languages compliant with the Web such as OWL, RDFS [1], and so on, have become increasingly popular in Computer Science with applications in areas such as bioinformatics, data integration, semantic Web, information retrieval, software engineering, service science, and many more. Ontologies are used to design conceptual models of information systems or to make the semantics of data more explicit for advanced information processing. We can see an ontology for an information system as a knowledge base consisting of a terminological and an assertional component [34]; the first one conveys general knowledge about a domain in terms of logical constraints that define the meaning of the concepts (and relations) used in the language (e.g. “every Cat is an Animal”); the second one expresses facts in terms of properties of individuals and relations holding between them (e.g. Fritz is a Black Cat; Fritz is friend of Joe); this distinction can be more or less sharp depending on the language an ontology is represented with, but can be adopted without loss of generality for our purposes. Some ontologies, e.g. an upper-level ontology such as DOLCE (<http://www.loa.istc.cnr.it/DOLCE.html>), are only defined at the terminological level; some other semantic resources, e.g. a linked open dataset such as Geonames (<http://www.geonames.org/>), are associated with such a shallow terminology, that although they can be still considered ontologies (Web ontologies, in fact), they are more similar to data bases associated with a schema. Finally, lexical resources such as Wordnet (<http://wordnet.princeton.edu/>) or vocabularies represented in SKOS are sometimes referred to as ontologies in the community [29]. Not every approach discussing IQ issues in the field of ontologies has the same type of ontology as target; however, most of the approaches

referenced here below consider ontologies as terminologies defined by a formal language. We now concentrate on three topics that we believe of interest in this context, because they highlight some peculiar perspectives on IQ that have been studied in the ontology domain. The topics are ontologies as *semiotic objects*, ontologies as *diverse knowledge objects*, and ontologies as *(reusable) computational resources*.

One of the first works that addressed the problem of evaluating (the quality of) ontologies exploited a framework based on a semiotic model [13]. A similar approach appears in a model that describes the relationship between ontologies as formal (externalized) specifications, (mental) conceptualization and the “real world” [19]. Within this cognitive-flavored semiotic approach, several quality dimensions and metrics have been defined on top of these frameworks. [19] distinguishes between quality dimensions and evaluation principles. Three types of dimensions under which it is possible to evaluate an ontology are discussed. The *structural dimension* focuses on syntax and formal semantics, i.e. on ontologies represented as graphs (context free metrics). The *functional dimension* is related to the intended use of a given ontology and of its components, i.e. their function in a context. The focus is on the conceptualization specified by an ontology. The *usability-profiling dimension* focuses on the ontology profile (annotations), which typically addresses the communication context of an ontology (i.e. its pragmatics). Then several principles (or evaluation-driven dimensions) are introduced, namely: *cognitive ergonomics, transparency, computational integrity and efficiency, meta-level integrity, flexibility, compliance to expertise, compliance to procedures for extension, integration, adaptation, generic accessibility, and organizational fitness*. Following the cognitive flavor of this point of view, a quite recent approach studied a measure of cognitive quality based on the adequacy of represented concept hierarchies w.r.t. the mental distribution of concepts into hierarchies according to a cognitive study [17]. These cognitive approaches clarify an important issue that has been central in the research about IQ in the ontology domain: ontologies are knowledge objects that are used by someone and for some specific goals; the evaluation of the quality of an ontology should consider an ontology in its semiotic context.

As it can be captured from the broad definition of ontology given at the beginning of this paragraph, ontologies are very different one from another. Some ontologies are flat, while some others consist in deep concept hierarchies; some ontologies are deeply axiomatized, while others look more like database schemas [14, 15]. Moreover, often ontologies cannot be modified but are reused and eventually extended. Some metrics defined for evaluating an ontology can be adopted to provide a value judgment about an ontology. Other metrics proposed so far are more intended as analytic dimensions to profile an ontology, and to understand its structure and its properties. As an example, one of the first unifying framework proposed to assess ontology quality distinguishes between syntactic, semantic, pragmatic and social qualities (see Table 3) [13]. Although lawfulness and interpretability clearly lead to a value judgment, metrics such as richness and history can be hard to be associated with a value judgment.

Dimension	Metrics	Definition
Syntactic quality	Lawfulness	Correctness of syntax
	Richness	Breadth of syntax used
Semantic quality	Interpretability	Meaningfulness of terms
	Consistency	Consistency of meaning of terms
	Clarity	Average number of word senses
Pragmatic quality	Comprehensiveness	Number of classes and properties
	Accuracy	Accuracy of information
	Relevance	Relevance of information for a task
Social quality	Authority	Extent to which other ontologies rely on it
	History	Number of times the ontology has been used

Table 3: Types of qualities and dimensions in [13].

In other frameworks such as the one proposed by [36, 19], which put a lot of focus on the computability of the defined metrics, most of the metrics are more aimed at profiling an ontology, rather than at assessing its quality from a value perspective. The idea is that these quality metrics can be used to summarize the main property of an ontology and their evaluation can be used by third party applications. As an example, a machine learning method that takes advantage of fine-grained ontology profiling techniques (extended from [36]) to automatically configure an ontology matching system has been recently proposed [15]. These approaches, which consider ontologies also as computational resources (see point above), differ from early works on ontology quality that were based on philosophical (metaphysical) principle to establish the quality of an ontology as a conceptual model, but whose analytical principles are more difficult to be made computable.

Finally, a key aspect of ontologies is that they are expected to be reused by other ontologies, applications, or, more generically, third party processes. It is often the case that one has to select an ontology to reuse it in a given domain. Ontologies can be used to support search or navigation. Different aspects of an ontology can be more or less amenable depending on the task an ontology is aimed to support. Approaches that evaluate ontologies on a task basis [37, 25, 35] seem to have received more attention, recently, than previous approach based on metaphysical and philosophical considerations [20], which better fit the use of ontologies as conceptual models, rather than as computational objects.

## CONCLUSION

In this paper we have discussed the main issues considered in data quality and information quality and several factors influencing these issues. According to a quite common use of the terms in the technical literature published by the data management community, we referred to data quality when structured data where addressed, and to information quality when information represented according to other data models is considered. We are aware that this pragmatic distinction is not based on a solid theoretical framework and can be questioned. However, the consideration of information digitally represented by different types of data and organized according to different data models has definitely a deep impact on the most relevant issues considered in information quality, including the definition itself. The more heterogeneous the considered information is, the more a comprehensive theoretical framework defining in a general way the mutual relationship between several crucial concepts in the definition and assessment of information quality (e.g., data, information, information carrier, observer, task, and so on) is needed. Recent works in the field of ontology evaluation framed the (information) quality problem within a broader semiotic and cognitive framework (see [19, 17]). A similar concern can be found in several works on information quality coming from the Information Systems community (see [42, 43]). These approaches can provide important contributions to a theoretical clarification of the common use of information quality core concepts and issues, in a context where the amount and the degree of complexity, diversity, and interconnection of the information managed in ICT is constantly increasing.

One problem that we believe particularly interesting is tightly related to the influencing factor IF4 addressed in this paper, which considers the impact on information quality of the degree of coupling between data and schemas (where available), and the difference in the semantics associated with structured and other types of data (e.g., schemaless data such as texts, images, sounds). An interesting research question concerns the extent to which information quality is affected by the degree of coupling between data and schemas, or, more in general, the role played by semantics defined by data models and schemas in the definition of information quality. This issue tightly relates to the relationship between data, information and *truth* in information systems. If schema-driven data can be easily interpreted as carriers of factual information and interpreted according to a semantic theory of truth [22] (e.g., through mapping to First-Order Logic), the connection between other types of information representations (e.g., maps, images, sounds) and factual information has been less investigated and results more obscure. Texts can be taken as borderline examples from this point of view: most of textual documents are clearly carriers of factual information to a human reader, but their digital representation is by no means related to

any factual interpretation (hence, investigations in the field of natural language processing, knowledge extraction, and so on). As a consequence of the above discussion, although the early effort for a theoretical foundation of data quality research [43], we point out that information quality still asks today for a general theoretical foundation of the basic key issues identified in this paper.

Considering suggestions from the above mentioned Information quality symposium at AISB/IACAP World Congress and following a practice which has guided foundational and grounding initiatives in the information systems research, we now consider potential analytical contribution from philosophy, in order to clarify and ground the results of our exploratory research on a more solid theoretical basis [41,42]. Indeed, the above research questions seem echoing the problem of knowledge of things *by acquaintance* (e.g. in the case of images) and *by description* (e.g. in the case of structured data), as stated for example by Bertrand Russel: “*there are two sorts of knowledge: knowledge of things, and knowledge of truths. [...] Knowledge of things, when it is of the kind we call knowledge by acquaintance, is essentially simpler than any knowledge of truths, and logically independent of knowledge of truths. Knowledge of things by description, on the contrary, always involves [...] some knowledge of truths as its source and ground.[...]We shall say that we have acquaintance with anything of which we are directly aware, without the intermediary of any process of inference or any knowledge of truths*” [31]. Thus, differently from knowledge by acquaintance, knowledge by description connects the truths (carried by data, in our case) with things with which we have acquaintance through our direct experience with the world (*sense-data*, in the Russel perspective). As an example of the role of factual information carried by data in information quality, observe that data and information quality pose the question of adherence of a certain representation to real world (see for example, clusters of dimensions such as *Accuracy/correctness/precision* or *Completeness/pertinence*). This question points to one of the most controversial issues discussed in philosophy so far. Significantly, Russel discusses this issue using the term *data*, and in particularly distinguishing between *hard data* and *soft data*: “this distinction is a matter of degree, and must not be pressed; but if not taken too seriously it may help to make the situation clear. I mean by ‘hard’ data those which resist the solvent influence of critical reflection, and by ‘soft’ data those which, under the operation of this process, become to our minds more or less doubtful. *The hardest of hard data are of two sorts: the particular facts of sense, and the general truths of logic [our italics]*” [33, p.56]. Indeed, from the above discussion we could ask ourselves to which extent information quality (and specific quality dimensions) may pertain to the domain of both hard and soft data. “*Our data now are primarily the facts of sense (i.e. of our own sense-data) and the laws of logic. But even the severest scrutiny will allow some additions to this slender stock. Some facts of memory—especially of recent memory—seem to have the highest degree of certainty. Some introspective facts are as certain as any facts of sense. And facts of sense themselves must, for our present purposes, be interpreted with a certain latitude. Spatial and temporal relations must sometimes be included[...] And some facts of comparison, such as the likeness or unlikeness of two shades of color, are certainly to be included among hard data*” [33, pp. 56-57]. As to this issue, the critical question is if information quality pertains to facts of sense or rather to laws of logic, which play a fundamental role both at the data model level (e.g., relational algebra for relational databases) and at the schema level (e.g., all persons are identified by their Social Security Number). And again, what can we say about data that are not straightforwardly associated with any truth-based semantics (e.g. images)? Finally, we mention that the role of the processes and tasks that are supported by an information system has to be considered when investigating the above research questions (the number of papers focusing on task-oriented evaluation of information quality is in fact increasing, e.g., see [37, 25, 35]). The above insights can be considered working constructs, with the aim of investigating whether perspectives coming from philosophical researches can bring some theoretical clarification on issues too often narrowly considered under a technical perspective in computer science.

## ACKNOWLEDGMENTS

The work presented in this paper is the mature result of the theoretical inquiries started with the first

version of this contribution presented at the AISB/IACAP World Congress 2012 - Information Quality: a special thanks to Phyllis Illari and Luciano Floridi, for the invitation and the opportunity of starting the reflections in this work. Furthermore, we acknowledge Raimondo Schettini for providing insights and some of the figures in the paper, with specific reference to image quality.

## REFERENCES

- [1] Antoniou, G., van Harmelen, F. *A Semantic Web Primer*. 2nd Ed., MIT Press, Cambridge, MA, 2008.
- [2] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (Eds.) *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.
- [3] Barone, D., Stella, F., Batini, C. “Dependency Discovery in Data Quality”. *CAiSE 2010*, 2010, pp. 53-67.
- [4] Batini, C., Di Battista, G., Santucci, G. “Structuring Primitives for a Dictionary of Entity Relationship data schemas”, *IEEE Transactions on Software Engineering*, March 1993.
- [5] Batini, C., Ceri, S., Navathe, S. B. *Conceptual Database design, An Entity Relationship Approach*. Benjamin & Cummings, 1994.
- [6] Batini, C. & Scannapieco, M. *Data quality: dimensions, techniques and methodologies*. Springer Verlag, 2006.
- [7] Batini, C., Cabitza, F., Pasi, G., Schettini, R. “Quality of Data, Textual Information and Images: a comparative survey”, Tutorial at the 27th International Conference on Conceptual Modeling (ER 2008), Barcelona, Spain, available on request to batini@disco.unimib.it.
- [8] Batini, C. “Looking for a ‘fil rouge’ among qualities in different information representations: the case of structured data, semi structured data, maps, unstructured texts, laws, and images”, Invited Speech at the QDB Workshop, International Conference on Very Large Data Bases, Lyon, France, 2009.
- [9] Batini, C., Cappiello, C., Francalanci, C., Maurino, A. “Methodologies for data quality assessment and improvement”, *ACM Computing Surveys* (41) 3, July 2009.
- [10] Batini, C., Palmonari, M. “Information Quality in the Web Era”, Tutorial at CAiSE 2010, Hammeth, Tunisia, <http://www.slideshare.net/palmonari/information-quality-in-the-web-era> .
- [11] Batini, C. “Data quality”. *Workshop on Global Scientific Data Infrastructures: the Big Data Challenges*, Capri 12-13 May 2011
- [12] Buneman, P. and Tan, W. “Provenance in databases”. In Proceedings of the 2007 ACM SIGMOD international Conference on Management of Data (Beijing, China, June 11 - 14, 2007). *SIGMOD '07*. ACM, New York, NY, 1171-1173, 2007
- [13] Burton-Jones, A., Storey, V. C., Sugumaran, V., Ahluwalia, P. “A semiotic metrics suite for assessing the quality of ontologies”. *Data Knowl. Eng.* 55 (1), pp. 84–102, 2005.
- [14] Cruz, I.F., Palmonari, M., Caimi, F., Stroe, C. “Towards ‘On the Go’ Matching of Linked Open Data Ontologies”. *LDH 2011*, pp. 37-42, 2011.
- [15] Cruz, I.F., Fabiani, A., Caimi, F., Stroe, C., Palmonari, M. “Automatic Configuration Selection Using Ontology Matching Task Profiling.” *ESWC 2012*, pp. 179-194, 2012.
- [16] Encyclopedia of Geographical Information Systems, Springer, 2008.
- [17] Evermann, J. & Fang, J. “Evaluating ontologies: Towards a cognitive measure of quality”. *Information systems*, Vol. 35 Issue 4, June 2010.
- [18] Floridi, L. “Semantic concept of Information”. *Stanford Encyclopedia of Philosophy*, 2009.
- [19] Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J. “Modelling ontology evaluation and validation”, in: Sure, Y. & Domingue, J. (Eds.), *ESWC, Vol. 4011 of Lecture Notes in Computer Science*, Springer, 2006, pp. 140–154.
- [20] Guarino, N. & Welty, C. A. “Evaluating ontological decisions with OntoClean”. *Commun. ACM* 45(2): 61-65, 2002.

- [21] Jøsang, A., Ismail, R. and Boyd, C. "A survey of trust and reputation systems for online service provision." *Decis. Support Syst.* 43, 2, March 2007.
- [22] Kirkham, R. *Theories of Truth*. Bradford Books. 1992.
- [23] ISO/IEC FDIS 25012 – Software Engineering – Software product quality requirements and evaluation – Data Quality Model, 2008.
- [24] Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. "AIMQ: A methodology for information quality assessment." *Information and Management*, 2001.
- [25] Lei, Y., Uren, V. S., Motta, E. "A framework for evaluating semantic metadata", in: Sleeman, D. H. & Barker, K. (Eds.), *K-CAP, ACM*, 2007, pp. 135–142.
- [26] Lindland, O.I., Sindre, G., Solvberg, A. "Understanding quality in conceptual modeling." *Software*, IEEE , vol.11, no.2, pp.42-49, March 1994.
- [27] Liu, L. & Chi, L. "Evolutionary Data quality", in Proceedings of the 6th International Conference on Information Quality, Boston, MA 2002.
- [28] Madnick, S. & Zhu, H. "Improving data quality through effective use of data semantics." *Data & Knowledge Engineering*, Volume 59, Issue 2, Including: Sixth ACM International Workshop on Web Information and Data Management, November 2006, 460-475.
- [29] Manaf, N.A.A., Bechhofer, S., Stevens, R. "The Current State of SKOS Vocabularies on the Web." In Proceedings of ESWC 2012: 270-284, Springer-Verlag, 2012.
- [30] Merriam Webster "Knowledge", *Merriam Webster*, 2012.
- [31] Russell, B., "Knowledge by Acquaintance and Knowledge by Description", *Proceedings of the Aristotelian Society (New Series)*, Vol.XI, (1910-1911), pp.108-128.
- [32] Russel, B. "Logic as the essence of philosophy", in Russel, B. *Our knowledge of the external world*, Routledge, 2009 (First edition 1914).
- [33] Russel, B. "Our knowledge of the external world", in Russel, B. *Our knowledge of the external world*, Routledge, 2009 (First edition 1914).
- [34] Staab, S. & Studer, R. (Eds.): *Handbook on Ontologies*. International Handbooks on Information Systems, Springer, 2004
- [35] Strasunskas, D. & Tomassen, S. L. "Empirical insights on a value of ontology quality in ontology-driven web search" in Meersman, R. & Tari, Z. (Eds.) *OTM Conferences* (2), Vol. 5332 of Lecture Notes in Computer Science, Springer, 2008, pp. 1319–1337.
- [36] Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., Aleman-Meza, B. "OntoQA: Metric- based ontology quality analysis", in Proceedings of IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources, 2005.
- [37] Yu, J., Thom, J. A., Tam, A. "Ontology evaluation using Wikipedia categories for browsing", in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, ACM, New York, NY, USA, 2007, pp. 223– 232.
- [38] Yu, Y. & Heflin, J. "Extending functional dependency to detect abnormal data in RDF graphs. In Proceedings of the 10th international conference on The semantic web" in Aroyo, L., Welty, C., Alani, H., Taylor, J. & Bernstein, A. (Eds.), *ISWC'11* Vol. Part I. Springer-Verlag, Berlin, Heidelberg, 2011, pp. 794-809.
- [39] Wang, R.Y. & Strong, D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems*, vol. 12, no. 4, 1996.
- [40] Wikipedia, "Knowledge", *Wikipedia, the free Encyclopedia*, 2012.
- [41] Wand, Y. & Weber, R. "An Ontological Model of an Information System." *IEEE Trans. Soft. Eng.* 16, 11, 1990., pp. 1282–1292.
- [42] Wand, Y. & Weber, R. "On the deep structure of information systems." *J. Info. Syst.*,1995, pp. 203–223.
- [43] Wand, Y. & Wang, R.Y. "Anchoring Data Quality Dimensions Ontological Foundations." *Communications of the ACM*, November 1996, Vol. 39, No. 11.

# THE EFFECT OF MISSING DATA ON CLASSIFICATION QUALITY

(Research in Progress)

**Michael Feldman**

Ben-Gurion University of the Negev, Israel  
[fmichael@bgu.ac.il](mailto:fmichael@bgu.ac.il)

**Adir Even**

Ben-Gurion University of the Negev, Israel  
[adireven@bgu.ac.il](mailto:adireven@bgu.ac.il)

**Yisrael Parmet**

Ben-Gurion University of the Negev, Israel  
[iparmet@bgu.ac.il](mailto:iparmet@bgu.ac.il)

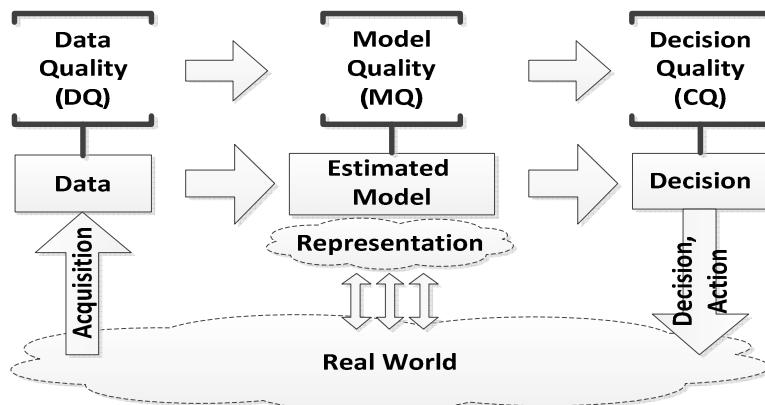
**Abstract:** The field of data quality management has long recognized the negative impact of data quality defects on decision quality. In many decision scenarios, this negative impact can be largely attributed to the mediating role played by decision-support models - with defected data, the estimation of such a model becomes less reliable and, as a result, the likelihood of flawed decisions increases. Drawing on that argument, this study presents a methodology for assessing the impact of quality defects on the likelihood of flawed decisions. The methodology is first presented at a high level, and then extended for analyzing the impact of missing values on binary Linear Discriminant Analysis (LDA) classifiers. To conclude, we discuss possible directions for extensions and future directions.

**Key Words:** Data Quality, Missing Values, Decision Making, Classification, Linear Discriminant Analysis

## INTRODUCTION AND BACKGROUND

The common saying “Garbage in Garbage Out” reflects a key concern in the field of data quality management (DQM) – the negative impact of data quality (DQ) defects on decision making (Redman, 1996; Shankaranarayanan and Cai, 2006; Liu et al., 2010). This study explores that impact through the mediating role played by decision-support models, arguing that a wrong decisions are often the result of an unreliable model that was built from low-quality data. Decision-making is often supported by a model (Shim et al., 2002) - a form of representation (e.g., theoretical, analytical, visual, statistical) that describes phenomena or behaviors in the real-world. Such a model permits prediction of future behavior to an extent and, by that, assists with the formation of decisions and actions. Following this notion, Decision-Support Systems (DSS) provide the infrastructure and the utilities for building, applying and evaluating models that aid the decision-maker.

Recent years have witnessed a major transition toward decision-making culture that is based on data collection and analysis (Davenport, 2006). This transition can be associated with the growing popularity of Business Intelligence and Data Warehousing (BI/DW) systems – DSS that rely on the collection and integrating data from diverse resources (Davenport, 2006). Data repositories, in BI/DW systems and others, are often subject to DQ defects – such as missing, inconsistent, and/or inaccurate data values. Such defects might create a biased view of the real-world and, consequently, lead to flawed decisions and actions. A plethora of studies (e.g., Redman, 1996; Heinrich et al., 2009; Even et al., 2010) have described real-world scenarios in which defected data led to wrong decisions and major damages. The goal of this study is to contribute some insights into the mechanisms that may further explain that link.



**Figure 1: A Decision Process**

Our methodology is conceptualized along three key stages of a typical data-driven decision process (Shim et al., 2002), and the associated quality assessments (Figure 1):

- **Data Quality (DQ):** Organizational data resources are built through ongoing complex processes of data acquisition, transfer and storage, during which they might become subject to DQ defects (Ballou et al., 1998; Parssian et al., 2004). Those data resources can support a variety of usages (Davenport, 2006, Even and Shankaranarayanan, 2007) – in this study we particularly observe the use of data for constructing and estimating models for decision-making support. DQ can be assessed along multiple dimensions, each reflecting a different type of data quality defects (Pipino et al., 2002, Even and Shankaranarayanan, 2007) – e.g., currency that reflects data that is not up-to-date, and accuracy that reflects incorrect values. This study addresses the impact of missing values – a common type of data quality defects, which is typically associated with the DQ dimension of completeness (Even et al., 2010). Data values may be missing due to reasons such as poorly designed data-entry screens, details that were not available (or not provided on purpose) at the time of data collection, database storage and update failures, or processing errors (Redman, 1996). This study focuses on missing completely at random (MCAR) patterns (Little, 1987), where missing data in one attribute does not depend on missing-value behavior in other attributes. Other missing-value patterns, such as missing at random (MAR) and not missing at random (NMAR), may assume some dependency between missing values. Such patterns should be further explored in future extensions to this study.
- **Model Quality (MQ):** The number of data items is often very large; hence, in many decision scenarios, data cannot be used as is. It is more common to use the data for constructing models that reflects real-world behavior in more compact and aggregated forms (e.g., formulas, charts, reports, digital dashboards, and the subject of this study – statistical classification models) that let a decision maker understand and analyze certain phenomena and behaviors. Model complexity and reliability may significantly affect decision making (Shim et al., 2002, Blake and Mangiameli, 2011). We interpret MQ as an assessment of model goodness – the extent to which our model reflects the true reality in a reliable manner. It is likely that with a higher rate of data quality defects (reduced DQ), the estimated model will provide a less reliable representation of reality (reduced MQ).
- **Decision Quality (CQ):** Models can serve as an input to decision-makers for gaining insights on how the real-world behaves, making some assessments and predictions, and act accordingly. The link between data quality and decision correctness, which has been explored in a variety of studies (e.g., Askira-Gelman, 2011, Blake and Mangiameli, 2011), is often complex and difficult to assess. We define CQ as the extent to which the decisions are correct. It is reasonable to assume that a flawed model might lead to misconceptions, flawed insights and hence wrong decisions – what motivates our claim that CQ is affected by MQ; hence, also by DQ.

In this study, we focus on classification – decision scenarios in which we associate a certain object, behavior, or situation with one category (or class) among a set of choices. Many decision scenarios, in different contexts, can be interpreted as classifications – e.g., replenishing inventory items (Davenport, 2006), assigning a customer to a segment (Even et al., 2010), or medical decisions, based on patient diagnostics (Session and Valtorta, 2009). Misclassification might damage reputation (e.g., misclassifying customers as “unimportant”), result in losses (e.g., investing in “overestimated” assets), or even threaten life (e.g., failing to detect hazardous medical conditions). Classifications often rely on models that can help associating a certain object with a certain class among a given set of choices – e.g., Distance-Based classifiers, k-Nearest-Neighbors (kNN), and Bayesian Classifiers (Duda and Hart, 2001). Classification models are often estimated (or “trained”) from a dataset. If the “training” dataset suffers from DQ defects – the estimated classifier is likely to be biased; hence, with a higher likelihood, the resulting decisions will be flawed. In this study we chose to evaluate our methodology with a relatively simple but common classifier – the binary Linear Discriminant Analysis (McLachlan, 1992). The next section introduces a methodology that links the quality levels described above – data, model, and decision - and highlights the relationships among them in the context of classifiers. The methodology is further developed for binary LDA – but some of the evaluation and measurement methods applied can be used in broader contexts. The concluding section summarizes the key contributions of our study, highlights its limitations, and proposes possible extensions and directions for future research.

## THE IMPACT OF INCOMPLETENESS ON CLASSIFIERS

This section develops a methodology for assessing the impact of data quality (DQ) on model and decision quality (MQ and CQ, respectively). The methodology (Figure 2) consists of the following components:

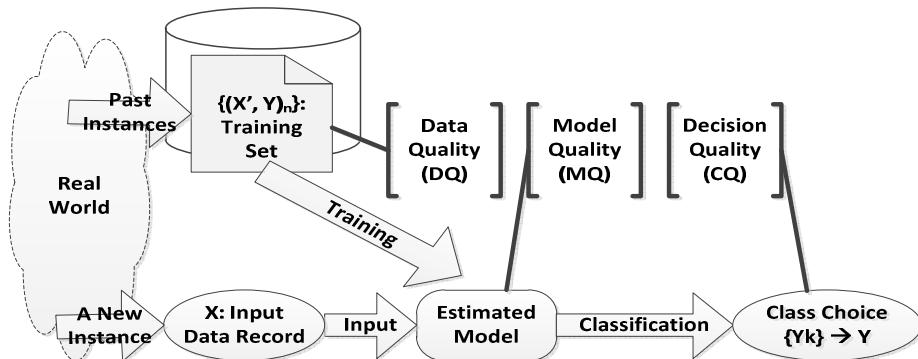


Figure 2: The General Methodology

**Training Sets and Data Quality Measurement ( $Q^D$ ):** the data stored in organizational repositories can be used for the estimation of classification models. Following common terminology (Duda and Hart, 2001), we refer to the process of estimating the model “training” and to the dataset  $\{(X, Y)_n\}$  used to estimate the model as a “training set”. The annotation reflects N records (indexed 1..N), where X is a vector of M attributes (indexed 1.. M), each reflecting a certain property of a real-world instance. The Y component is a 1..K integer that associates the record with one among K classes. Following common DQ measurement schemas (Even and Shankaranarayanan, 2007), each record is associated with a  $Q_n$  measurement of completeness - 0, if one or more attribute values (or the entire record) are missing (i.e., NULL), 1 is the record is complete. The quality of the entire dataset  $Q^D$ , in terms of completeness, is defined as the rate of non-missing values, where  $Q^D=1$  reflects a complete training set:

$$Q^D = \frac{1}{N} \sum_{n=1}^N Q_n, \quad 0 \leq Q^D \leq 1 \quad (1)$$

**Classification Models, and Model Quality Measurement ( $Q^M$ ):** A classifier can be described, in general, as a function  $M(X)=Y$  that maps an  $M$ -dimensional input vector  $X$ , which reflects a real-world instance to be classified, to an output integer  $Y=1..K$  associated with a class within a  $K$ -class set. In the decision scenarios that we discuss, the classifier parameters have to be estimated from a training set, as discussed above. With an “infinite” number of random sample (i.e., a very large  $N$ ), the estimates of model parameters are likely to be accurate and reliable. However, with a smaller number of samples, the likelihood of misestimating parameters is higher and so is the likelihood of classification errors.

The confidence interval (CI) is a common approach for assessing the reliability of estimated model parameters. For example, when estimating a certain parameter  $A$  from a training set – the estimated value  $\hat{a}$  is not necessarily the true one. CI assessment would allow us to assume that “with a confidence of  $g\%$  the true value of  $A$  resides within the CI of  $[\hat{a} - \Delta_1, \hat{a} + \Delta_2]$ ”. Obviously – the smaller are the CI’s for all parameters, the more reliable is the classification model. Further, with classification models that involve CI assessment, it can be shown that the CI gets smaller with a higher  $N$ . Adopting the CI-assessment concept - we take  $L$ , the length of the confidence interval as a measure for model quality ( i.e., if the confidence interval is defined by  $[\hat{a} - \Delta_1, \hat{a} + \Delta_2]$ , then  $L = \Delta_1 + \Delta_2$ ). The model-quality metric has to be defined for each model parameter  $A$ . It has to consider the desired target confidence level  $\rho$ , the number of samples  $N$  in the complete dataset, and the missing value rate (as reflected by  $Q^D$ ):

$$Q_A^M(\rho, N, Q^D) = L_A(\rho, N * Q^D) \quad (2)$$

Where

$A$ -	The model parameter under evaluation
$\rho$ -	The target confidence level
$N$ -	The number of samples in the complete training dataset
$Q^D$ -	The data quality level (i.e., the rate of non-missing values)
$L_A(x, y)$ -	The CI length for parameter $A$ , given target confidence level $y$ , and $x$ samples

**Confusion Matrix, and Decision Quality Measurement ( $Q^C$ ):** The classification output  $Y$  is an integer in the range of  $[1..K]$ , which reflects an association to the input record (or vector)  $X$  to one class within a  $K$ -class set. A classification is said to be correct if an instance that belongs to class  $k$  is indeed classified to class  $k$ , and incorrect otherwise. With binary classifiers (i.e.,  $K=2$ ), in which the output is either positive ( $Y=1$ ) or Negative ( $Y=0$ ), it is common to assess classification performance with the 2-way confusion matrix (Table 1) – a Positive item that was classified as Positive is considered as “True Positive” ( $TP$ ), and so on (Han and Kamber, 2006). The total number of instance per quadrant ( $N_{TP}$ ,  $N_{FP}$ ,  $N_{FN}$ ,  $N_{TN}$ , respectively, where  $N_{TP}+N_{FP}+N_{FN}+N_{TN}=N$ ), are commonly used for assessing the following classification quality metrics, and possibly others:

- **Classification Accuracy ( $Q^{CA}$ )**, reflecting the rate of items classified correctly:  $(N_{TP} + N_{TN}) / N$
- **Classification Precision ( $Q^{CP}$ )**, reflecting correctness within positive results:  $N_{TP} / (N_{TP} + N_{FP})$
- **Classification Sensitivity ( $Q^{CS}$ )**, reflecting the ability to detect positive results:  $N_{TP} / (N_{TP} + N_{FN})$
- **Classification Specificity ( $Q^{CF}$ )**, reflecting the ability to detect negative results:  $N_{TN} / (N_{TN} + N_{FP})$

Real-World Class	Classification	
	1	0
1	True Positive (TP)	False Negative (FN)
0	False Positive (FP)	True Negative (TN)

**Table 1: Binary Classification Assessment with 2-Way Confusion Matrix**

A more general formulation of classifier-performance assessment, which can also address classifications with a larger number of classes ( $K > 2$ ), uses a confusion matrix (Table 2). The a-priory probabilities  $\{V_1 \dots V_K\}$  reflect the real-world distributions of classed ( $\sum_{k=1..K} V_k = 1$ ). The matrix items  $\{W_{i,j}\}$  ( $(\sum_{j=1..K} W_{i,j} = 1)$ ) reflects the probability of a real-world instance that belongs to class  $i$  to be classified as class  $j$  (a correct classification if  $i=j$ , incorrect otherwise). Accordingly, the decision quality  $Q^C$  is defined as the overall likelihood of correct classification (similar to “classification accuracy” for the binary classification case):

$$Q^C = \sum_{k=1}^K V_k W_{k,k}, \quad 0 \leq Q^C \leq 1 \quad (3)$$

Real-World Class	A-Priory Probability	Classification			
		1	2	...	K
1	$V_1$	$W_{1,1}, U_{1,1}$	$W_{1,2}, U_{1,2}$	...	$W_{1,K}, U_{1,K}$
2	$V_2$	$W_{2,1}, U_{2,1}$	$W_{2,2}, U_{2,2}$	...	$W_{2,K}, U_{2,K}$
...	...	...	...	...	...
K	$V_K$	$W_{K,1}, U_{K,1}$	$W_{K,2}, U_{K,2}$	...	$W_{K,K}, U_{K,K}$

**Table 2: K-Class Confusion Matrix, Including Relative Costs**

An enhanced definition of  $Q^C$  may take into account the relative classification value, assuming that certain classification errors are possibly more severe than others. The parameters  $\{U_{i,j}\}$  in the weighted confusion matrix (Table 2) reflects that relative value of classifying an item that belongs to real-world class  $i$  as  $j$ . We assume that all the diagonal values are non-negative  $U_{i,i} \geq 0$  (i.e., correct classification cannot cause a damage), and that for each  $i$  and  $j$ ,  $U_{i,j} \geq U_{j,i}$ . This means that misclassification cannot have a higher value than a correct classification (otherwise, we would have adjusted the classifier to “misclassify”). However, misclassification might have a negative value – i.e., a certain costly damage to the overall performance (i.e.,  $U_{i,j}$  can be negative if  $i \neq j$ ). Following these assumptions, the decision quality  $Q^C$  definition can be adjusted to:

$$Q^C = \frac{\sum_{k=1}^K V_k \sum_{j=1}^K W_{k,j} U_{k,j}}{\sum_{k=1}^K V_k U_{k,k}} \leq 1 \quad (4)$$

Notably the denominator in that expression  $U^{\max} = \sum_{k=1..K} V_k U_{k,k}$  reflects the expected value from a single classification act, with no classification errors. Hence,  $Q^C$  reflects the ratio between the expected value with some misclassification and  $U^{\max}$ . As the value of some misclassification can be negative,  $Q^C$  might turn out to be negative too (e.g., in case that some likelihood exists for very costly misclassification). When all the diagonal values are equal  $U_{k,k} = U$ , and when all other non-diagonal values are 0 (i.e., no value, and no damage), the  $Q^C$  expression in Equation 4 becomes identical to Equation 3.

A special treatment is needed for the case where the diagonal values are all 0, but some non-diagonal

values are negative - i.e.,  $U_{i,j}=0$  for  $i=j$ ,  $U_{i,j} \leq 0$  for  $i \neq j$ . This case reflects a decision scenario in which there is no value associated with correct classification, but there is some damage associated with misclassification. In that case, instead of measuring decision quality as defined earlier, it would be more reasonable to measure the decision cost  $C^C$ :

$$C^C = \sum_{k=1}^K V_k \sum_{j=1}^K W_{k,j} U_{k,j} \leq 0 \quad (5)$$

The decision quality and cost discussed so far may rely on the number of samples  $N$  in the training set. Even with an “infinite” number of samples (i.e., a very large  $N$ ), the model may still have some classification errors due to possible overlaps between classes (as shown later for LDA classifiers). With a smaller, “finite” number of samples – the classifier’s performance is likely to degrade further. We now define the decision quality  $Q^C(N)$ , as a function of the number of samples  $N$ . The upper limit  $Q^{C*}$  reflects the best possible decision quality for a classifier that was estimated with an “infinitely large” number of sample and  $CI \rightarrow 0$ . Similarly, we define the decision cost  $C^C(N)$  as a function of the sample size. The lower limit  $C^{C*}$  reflects the lowermost decision cost for a given classifier, with very large  $N$ , and  $CI \rightarrow 0$ .

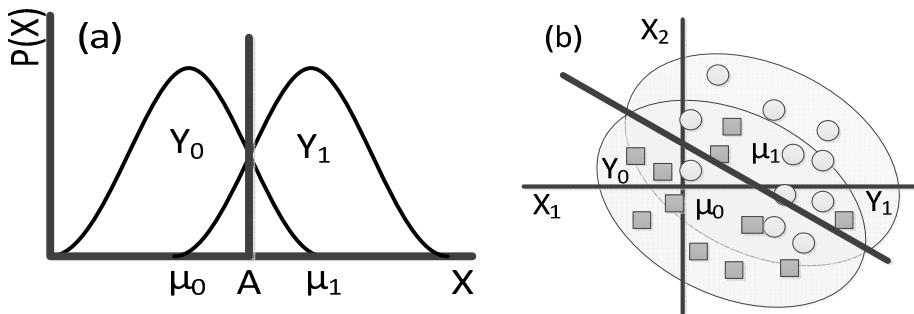
$$Q^{C*} = \lim_{N \rightarrow \infty} Q^C(N), \quad C^{C*} = \lim_{N \rightarrow \infty} C^C(N) \quad (6)$$

The metrics developed so far, within the measurement methodology introduced in this section, were defined in a general manner that permits their usage in many classification scenarios. However, we suggest that with further analytical development, such metrics can become even stronger tools for assessing and predicting DQ, MQ, and CQ behavior, and setting DQ policies accordingly. In the following section we demonstrate such an extension for the commonly-used, LDA classifiers.

## DEVELOPMENT AND EVALUATION FOR BINARY LDA CLASSIFIERS

The binary Linear Discriminant Analysis (LDA) classifier (McLachlan, 1992; Duda and Hart, 2001) assigns an input vector  $X$  to either class  $Y_0$  or class  $Y_1$ . For terminology convenience, and with no loss of generality, we term one class as “positive” and the other as “negative” and annotate them with “1” and “0” respectively. The LDA assumes that two classes reflect normally-distributed populations, with a different mean per class ( $\mu_0$  and  $\mu_1$  respectively), but with the same covariance matrix  $\Sigma$ . The LDA classifies a vector  $X$  (all attributes are continuous) to  $Y_0$  or  $Y_1$  by calculating a Cartesian product between  $X$  and a separation hyper-plane  $W$  and comparing the result to a threshold value  $A$ :

$$W \bullet X > A, \quad \text{where} \quad W = \Sigma^{-1}(\mu_1 - \mu_0) \quad (7)$$



**Figure 3: LDA Classifiers for (a) 1-dimensional space, and (b) 2-dimensional space**

Figure 3a shows a binary LDA classifier for a scalar (“1 dimensional”) input, in which case the classification rule can be simplified to:  $X$  is classified as  $Y_1$  if  $X > A$ , or classified as  $Y_0$  otherwise (Again, with no loss of generality, we assume that the class with the higher mean is the “positive”, while the class with the lower mean is “negative”). Figure 3b shows a binary LDA classifier for a 2-dimensional input vector.

Both examples highlight the fact that the binary LDA is not a perfect classifier – some misclassifications may occur, as the populations of the two classes may overlap to an extent. However, it can be shown that given the parameters of the two distributions – the LDA classifier defines the optimal linear separation in terms of minimizing the likelihood of error. To demonstrate our evaluation concept, and highlighting the potential contribution, the rest of this section develops further the scalar (1-dimensionl) case. In the concluding section we will discuss a few extensions currently under research.

As summarized in Table 3,  $Y_1$  (“positive”) and  $Y_0$  (“negative”) are with a-priory probabilities of  $V_1=V_0=0.5$ . Each class reflects a Normally-distributed population with different means  $\mu_1 > \mu_0$  but the same STDEV  $\sigma$ . We consider a case where there’s no positive value to correct classification, but some known cost  $U$  of misclassifications (The cost is identical for “False Positive” and “False Negative”). With some probability  $W_{TP}$  a “positive” item can be classified correctly as “positive”, and with some probability  $W_{FN}=1-W_{TP}$  as “negative” ( $W_{TP}+W_{FN}=1$ ). Similarly, with some probability  $W_{TN}$  a “negative” item can be classified correctly as “negative”, and with some probability  $W_{FP}=1-W_{TN}$  as “positive”.

Class	A-Priory Probabil-ity	Distribution Function	Classification	
			1 – Positive	0 – Negative
<b>1 - Positive</b>	$V_1 = 0.5$	$P_1 \sim N(\mu_1, \sigma)$	<b>True Positive:</b> $W_{TP}, 0$	<b>False Negative:</b> $W_{FN}, U$
<b>0 – Negative</b>	$V_0 = 0.5$	$P_0 \sim N(\mu_0, \sigma)$	<b>False Positive:</b> $W_{FP}, U$	<b>True Negative:</b> $W_{TN}, 0$

**Table 3: The Confusion Matrix, for the Binary LDA Case**

The LDA model, in that case, has one parameter only – the threshold  $A$  that defines the classification rule (a new instance  $x$ , with unknown classification, is classified as “positive” if  $x>A$ , or “negative” otherwise). Based on the assumptions above, it can be shown that with known distribution parameters ( $\mu_1$ ,  $\mu_0$ , and  $\sigma$ ), the optimal threshold value, in terms of maximizing classification accuracy, is  $A=0.5*(\mu_0+\mu_1)$ , with a confidence interval of  $CI_A=0$  (as the distribution parameters are known, and not estimated). The probabilities of correct classifications versus misclassification can be calculated accordingly as follows:

$$\begin{aligned}
 W_{TP} &= 1 - \Phi((A - \mu_1)/\sigma) = 1 - \Phi\left(\left(\frac{\mu_0 + \mu_1}{2} - \mu_1\right)/\sigma\right) \\
 &= 1 - \Phi((\mu_0 - \mu_1)/2\sigma) = \Phi((\mu_1 - \mu_0)/2\sigma) \\
 W_{FP} &= 1 - W_{TP} = 1 - \Phi((\mu_1 - \mu_0)/2\sigma) = \Phi((\mu_0 - \mu_1)/2\sigma)
 \end{aligned} \tag{8}$$

$$\begin{aligned}
 \text{Due to symmetry : } W_{TN} &= W_{TP} = \Phi((\mu_1 - \mu_0)/2\sigma), \\
 W_{FN} &= W_{FP} = \Phi((\mu_0 - \mu_1)/2\sigma) \\
 (\Phi - \text{Cumulative Normal Distribution})
 \end{aligned}$$

The expected decision quality (Equation 3) for this case is:

$$Q^C = Q^{C^*} = V_1 * W_{TP} + V_0 * W_{TN} = \Phi((\mu_1 - \mu_0)/2\sigma) \tag{9}$$

It can be shown that with known distribution parameters ( $\mu_1$ ,  $\mu_0$ , and  $\sigma$ ), the expression in equation 9 would be the best possible decision quality that can be obtained (hence,  $Q^{C^*}$ ). With  $\mu_1 - \mu_0 \rightarrow 0$  and/or with  $\sigma \rightarrow \infty$ ,  $Q^{C^*} \rightarrow 0.5$  (a random “flip of a coin”). With  $\mu_1 \gg \mu_0$ , and/or with  $\sigma \rightarrow 0$ ,  $Q^{C^*} \rightarrow 1$ . The expected decision cost (Equation 5), in that case, would be:

$$C^C = C^{C^*} = U * (1 - \Phi((\mu_1 - \mu_0)/2\sigma)) = U * \Phi((\mu_1 - \mu_0)/2\sigma) \tag{10}$$

Again, with known distribution parameters, this would be the lowest possible decision cost (hence,  $C^{c^*}$ ). With  $\mu_1 - \mu_0 \rightarrow 0$ , and/or with very large  $\sigma$ ,  $C^{c^*} \rightarrow 0.5U$ . With  $\mu_1 \gg \mu_0$ , and/or with  $\sigma \rightarrow 0$ ,  $C^{c^*} \rightarrow 0$ .

### **Parameter Estimation and Model Quality Metric for the Binary LDA Classifier**

So far, the development reflected classifier parameters that are known in advance – however, in the decision scenarios that we discuss, the parameters  $\mu_1$ ,  $\mu_0$ , and  $\sigma$  have to be estimated from a “training set” –  $\hat{\mu}_1$ ,  $\hat{\mu}_0$ , and  $\hat{\sigma}$ , respectively. At full size, our “training set” has  $N$  samples for each class (a total of  $2N$ ). Some values are missing from that training set, hence a data quality level of  $Q^D$ . We assume that the values are missing completely at random (MCAR); hence, the incompleteness distributes evenly between the two classes, and the training set contains  $Q^D N$  samples of the each group. We annotate the “positive” and “negative” training sets with the missing values by  $\{x_n^1\}$  and  $\{x_n^0\}$ , respectively (in both classes the index  $n$  goes between  $1..Q^D N$ ). Under the MCAR assumption, we can use unbiased estimators for the means and the variance:

$$\begin{aligned}\hat{\mu}_1 &= \frac{\sum_{n=1}^{Q^D N} x_n^1}{Q^D * N}, \quad \hat{\mu}_0 = \frac{\sum_{n=1}^{Q^D N} x_n^0}{Q^D * N} \\ \hat{\sigma}_1 &= \sqrt{\frac{\sum_{n=1}^{Q^D N} (x_n^1 - \hat{\mu}_1)^2}{Q^D * N - 1}}, \quad \hat{\sigma}_0 = \sqrt{\frac{\sum_{n=1}^{Q^D N} (x_n^0 - \hat{\mu}_0)^2}{Q^D * N - 1}} \\ \hat{\sigma} &= \sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}{2}} = \frac{\sqrt{\sum_{n=1}^{Q^D N} (x_n^1 - \hat{\mu}_1)^2} + \sqrt{\sum_{n=1}^{Q^D N} (x_n^0 - \hat{\mu}_0)^2}}{2\sqrt{Q^D * N - 1}}\end{aligned}\tag{11}$$

As mentioned earlier, if distribution parameters are known, the classification threshold can be calculated by  $A=0.5*(\mu_0+\mu_1)$ . Here, we need to estimate  $\hat{A}$ , based on the training set. As the samples in the training set are drawn from Normally-distributed populations, the estimator  $\hat{A}$  is also a normally-distributed random variable, for which we can calculate the expected value  $E[\hat{A}]$ , and the variance  $VAR[\hat{A}]$ :

$$\begin{aligned}\hat{A} &= \frac{\hat{\mu}_1 + \hat{\mu}_0}{2}, \quad E[\hat{A}] = E\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\mu_1 + \mu_0}{2} \\ VAR[\hat{A}] &= VAR\left[\frac{\hat{\mu}_1 + \hat{\mu}_0}{2}\right] = \frac{\sigma^2}{2Q^D * N}\end{aligned}\tag{12}$$

As discussed in the previous section, the rate of missing values (as reflected by data quality measurement  $Q^D$ ) may directly affect the classification rule, by increasing uncertainty about best classification threshold. As seen in equation 12 above, missing values that follow the MCAR, do not bias of expected threshold (the expression  $E[\hat{A}]$  does not depend on the data quality level  $Q^D$ ). However, missing values might affect estimation uncertainty and hence, the model quality  $Q^M$ . The estimation variance  $VAR[\hat{A}]$  and the associated confidence interval (CI), increase with a higher rate of missing values (lower  $Q^D$ ). As the estimator for the threshold parameter has a Normal distribution, the confidence interval  $CI_A$  for the estimator  $\hat{A}$ , given a desired confidence level  $\rho$ ,  $N$  samples, and a data quality level of  $Q^D$  is:

$$CI_A(\rho, N) = \left[ \hat{A} - t_{1-\rho/2, 2N-2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}}, \quad \hat{A} + t_{1-\rho/2, 2N-2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}} \right] \quad (13)$$

Where,

- $\hat{A}$  - The estimation of the LDA threshold A
- $\rho$  - The target confidence level
- $N$  - The number of samples in the complete training dataset
- $Q^D$  - The data quality level (i.e., the rate of non-missing values)
- $t_{1-\rho/2, N}$  The 1- $\rho$  quantile of Student-t distribution with N degrees of freedom

Accordingly, we can calculate the CI-length (and, with equation 2, also the MQ metric) for the LDA threshold A, given a desired confidence level  $\rho$ , N samples in the complete dataset, and a DQ level of  $Q^D$ :

$$Q_A^M(\rho, N, Q^D) = L_A(\rho, N, Q^D) = 2 * t_{1-\rho/2, 2N-2} * \sqrt{\frac{\hat{\sigma}^2}{2Q^D * N}} = \\ t_{1-\rho/2, 2N-2} * \frac{\sqrt{\sum_{n=1}^{Q^D N} (x_n^1 - \hat{\mu}_1)^2} + \sqrt{\sum_{n=1}^{Q^D N} (x_n^0 - \hat{\mu}_0)^2}}{\sqrt{Q^D * N - 1}} \quad (14)$$

Figure 4 shows the model quality ( $Q^M$  – the confidence interval length) versus the data quality (QD) for different sample sizes, and with  $\rho = 0.05$ . The samples were taken from two normally-distributed populations with  $\mu_0=2$ ,  $\mu_1=4$  and common  $\sigma=3$ .

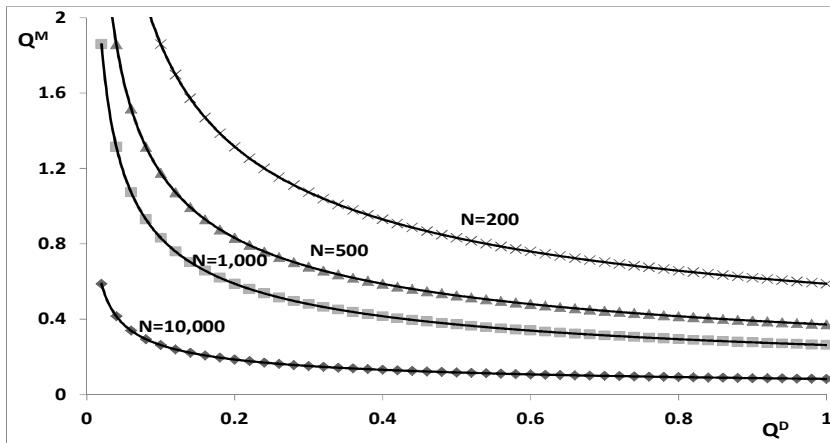
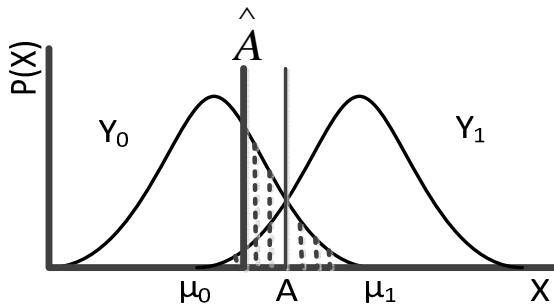


Figure 4: Model Quality (QM) versus Data Quality, with  $\rho = 0.05$

The figure highlights our earlier arguments - model quality is likely to increase (smaller confidence interval) with a higher N, and with a higher DQ level. Notably, with the highest sample-size shown (N=10000), the QM degradation is relatively minor for small QD degradation (QM (QD=1) = 0.08, versus QM (QD=0.6) = 0.1), but becomes more severe as QD reaches low rates (QM (QD=0.1) = 0.26). It can be shown that with a large N, the Student-t distribution can be approximated with a Normal distribution - e.g., with 30 or more degrees of freedom, the error of approximating the probability density function (PDF) of a Student-t distribution with a Normal distribution is less than 0.005. Accordingly, the CI-length will be approximated by  $L_A(\rho) = 2 * Z_{1-\rho/2} * \sigma$ .

### Decision Quality Metric for the Binary LDA Classifier

After showing the effect of DQ on MQ, we now show the impact of DQ and MQ on the decision quality CQ. In our decision scenario (Table 3), there is no value for correct classification, but some negative cost U for misclassification – hence, we assess decision quality in terms of lowering cost. With known distribution parameters, the lowest-possible cost (Equation 10) was shown to be  $C^{C^*} = U * \Phi((\mu_1 - \mu_0)/2\sigma)$ . In this section, we will show that when the parameters have to be estimated from a sample – the decision quality will degrade (i.e., higher negative cost) with a smaller sample size and lower DQ level.



**Figure 5: Misclassification Due to Biased Threshold Estimation**

Given a certain threshold  $\hat{A}$  that was estimated from a training set (Equation 14) - misclassification of instance X occurs when it is “positive”, but smaller than  $\hat{A}$  or “negative” but greater than  $\hat{A}$ . Given a cost parameter of U and an estimated threshold  $\hat{A}$ , the expected misclassification cost at is:

$$\begin{aligned} C^c(\hat{A}) &= U * \left( P(X < \hat{A} | X \in Y_1) + P(X > \hat{A} | X \in Y_0) \right) = \\ &= U * \left( \Phi\left(\frac{\hat{A} - \mu_1}{\sigma}\right) + 1 - \Phi\left(\frac{\hat{A} - \mu_0}{\sigma}\right) \right) = \\ &= U * \left( \Phi\left(\frac{\hat{A} - \mu_1}{\sigma}\right) + \Phi\left(\frac{\mu_0 - \hat{A}}{\sigma}\right) \right) \end{aligned} \quad (15)$$

It can be shown that  $C^c$  is minimized when  $\hat{A} = A = 0.5 * (\mu_0 + \mu_1)$  (i.e., with a sample size  $N \rightarrow \infty$ ):

$$C^{C^*} = C^c(\hat{A} = A = 0.5 * (\mu_0 + \mu_1)) = U * \Phi((\mu_1 - \mu_0)/2\sigma) \quad (16)$$

Given a finite sample-size N and a quality level  $Q^D$  (i.e., an actual sample size of  $Q^D * N$ ) – we define the expected classification cost  $C^c$  as the mean of  $C^c(\hat{A})$  for all possible values of the estimated threshold  $\hat{A}$ .

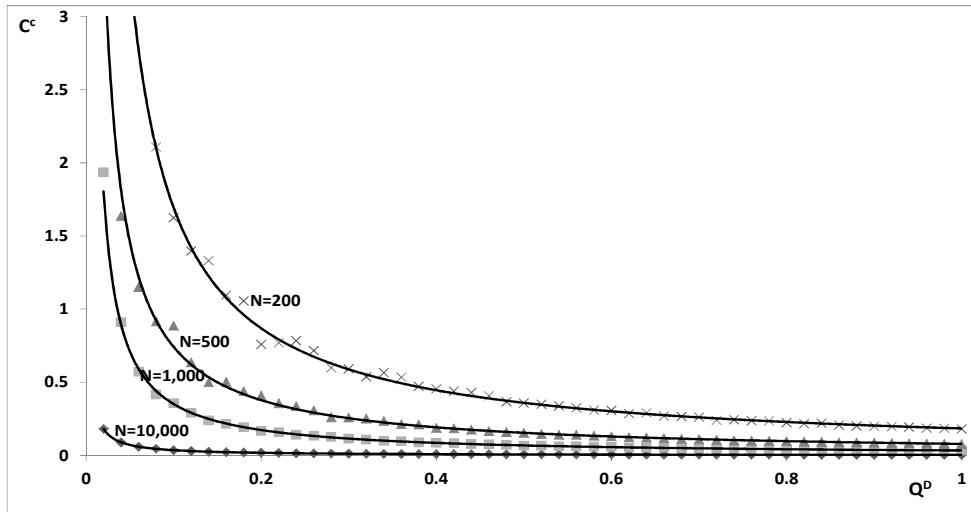
$$C^c = E[C^c(\hat{A})] = U * E[\Phi\left(\frac{\hat{A} - \mu_1}{\sigma}\right) + \Phi\left(\frac{\mu_0 - \hat{A}}{\sigma}\right)] \quad (17)$$

The calculation of the mean depends on a certain confidence interval CI – given an actual sample size of  $Q^D * N$ , with a confidence rate of  $\rho$  (i.e., a likelihood of  $1 - \rho$ ), the estimated threshold  $\hat{A}$  will reside within a  $\Delta$  range around A, where  $\Delta$  depends on N,  $Q^D$ , and  $\rho$ .

$$C^c(\rho, N, Q^D) = E \left[ C^c(\hat{A}) \middle| \hat{A} \subset CI \right] = \\ U * \frac{\int_{\hat{A} \subset CI} \left( \Phi\left(\frac{\hat{A} - \mu_1}{\sigma}\right) + \Phi\left(\frac{\mu_0 - \hat{A}}{\sigma}\right) \right) d\hat{A}}{(1-\rho) * \Delta(\rho, N, Q^D)} = U * \delta(\rho, N, Q^D) \quad (18)$$

$$\text{where } CI = \left[ \hat{A} - 0.5 * \Delta(\rho, N, Q^D), \hat{A} + 0.5 * \Delta(\rho, N, Q^D) \right]$$

The expression  $\delta(\rho, N, QD)$  in Equation 18 reflects the average likelihood that a certain item will be misclassified, given certain values of confidence level  $\rho$ , training-set size  $N$ , and DQ level  $QD$ . It is likely to decrease with a smaller  $\rho$ , larger  $N$ , and/or larger  $QD$ . Figure 6 shows the expected classification cost (CC) versus the data quality (QD) for different sample sizes, with  $U=1$  and  $\rho=0.05$  (the same training sets that were used in Figure 4 -  $\mu_0=2$ ,  $\mu_1=4$ ,  $\sigma=3$ ).



**Figure 6: Model Quality (QM) versus Data Quality, with  $U=1$  and  $\rho = 0.05$**

The similarity in behavior between Figure 4 and Figure 6 is noticeable – the expected cost is higher with lower sample size, and decreases further as the rate of missing values increases (lower  $Q^D$ ). With a very large  $N$  (here, the maximum take is  $N=10,000$ ), and with no missing values ( $Q^D=1$ ), the expected  $C^c$  nearly reaches the optimum ( $C^{c*} \approx 0.036$ ). At this large sample size the impact of missing values is relatively minor – there a significant change in  $C^c$  only when  $Q^D$  goes below 0.1.

### **Data Quality, Decision Quality and Cost-Benefit Tradeoffs**

Assuming that we now have the ability to complete missing values in our training set, at a cost of  $S$  units per missing items – would the benefits gained from completing those values justify the associated cost? The answer would be yes – if the reduction in misclassification cost will be higher than the cost of missing-values completion.

Assume that the current quality level is  $Q^{D/S}$ , and the target quality level is  $Q^{D/T}$ . If we have  $N^T$  items that need to be classified, the classification costs that will be saved by filling in missing values will be

$$\begin{aligned} \Delta C^C(Q^{D/T}) &= N^T * (C^C(\rho, N, Q^{D/T}) - C^C(\rho, N, Q^{D/S})) = \\ &N^T * U * (\delta(\rho, N, Q^{D/T}) - \delta(\rho, N, Q^{D/S})) \end{aligned} \quad (19)$$

The correction cost  $\Delta C^S$  of increasing the quality level from  $Q^{D/S}$  to a target quality level of  $Q^{D/T}$  is:

$$\Delta C^S(Q^{D/T}) = S * N * (Q^{D/T} - Q^{D/S}) \quad (20)$$

The net-benefit associating with missing-value competition is given by  $B(Q^{D/T}) = \Delta C^S(Q^{D/T}) - \Delta C^S(Q^{D/S})$ . We can now frame the question of what quality-level to target as an optimization problem:

Choose  $Q^{D/T}$  that maximizes:

$$B(Q^{D/T}) = N^T * U * (\delta(\rho, N, Q^{D/T}) - \delta(\rho, N, Q^{D/S})) - S * N * (Q^{D/T} - Q^{D/S}) \quad (21)$$

S.t.,  $Q^{D/T} \leq Q^{D/S} \leq 1$ ,  $B \geq 0$

Where,

$B$	The net-benefit associated with data quality improvement
$Q^{D/T}$ -	The target data quality level
$Q^{D/S}$ -	The given data quality level
$\rho$ -	The target confidence level
$N$ -	The number of samples in the complete training dataset
$N^T$ -	The number of samples to be classified
$\delta(\rho, N, Q^D)$	The average likelihood of misclassification
$U$	The expected cost of misclassifying a single item
$S$	The cost of fixing a single missing value

The objective function formulation in Eq. 21 is not linear and, obviously, does not have a close-form solution; however, the optimal solution can be approximated using a software-based optimization tool. As highlighted by a few studies (e.g., Ballou et al., 1998; Heinrich et al., 2009, Even et al., 2010) – DQ management decisions often involve substantial cost-benefit tradeoffs. The need for cost-benefit assessment is also reflected in the analysis done in this study – but with some separation between the datasets on which we act. The data correction cost is associated with the training set, used for building the model. On the other hand, the reduction in misclassification cost is associated with data items that are not part of the training set, but have to be classified according to the model developed.

### Discussion - Limitations and Future Extensions

The general methodology described earlier suggests that DQ may affect MQ, and hence CQ behavior. This section developed this argument further by demonstrating an analytical methodology that shows the explicit link between the three levels. This section introduced a more detailed development of that concept for binary LDA classifiers – a relatively simple, yet useful classifier. The development showed explicit and quantifiable links between the missing-value rate (as reflected by the DQ measure  $Q^D$ ), the model quality (in terms of minimizing the confidence-interval length), and the decision quality (in terms of minimizing misclassification costs). As shown in Equation 21, the mapping between the data quality level and the expected misclassification cost can be used for developing analytical tools that permit cost-benefit assessments. Based on the results of such assessments – the target quality level can be set, such that the margin between the classification-cost saved and the correction cost will be maximized.

To highlight the key concepts and arguments – the analytical development in this section was done under some simplifying and restrictive assumptions. Those assumptions should be relaxed in future extensions to this study, as summarized in Table 4.

<b>Issue</b>	<b>Assumption Made</b>	<b>Future Extensions</b>
<b>Dimensions</b>	• Scalar (“1-dimensional”) input	• Multidimensional input vector
<b>Classes</b>	• Two	• Any $K \geq 2$
<b>Symmetry</b>	<ul style="list-style-type: none"> <li>• Class distributions with different means, but identical STDEV</li> <li>• Same a-priory probability</li> <li>• Same number of samples per class</li> <li>• Same misclassification cost for “false positive” and ‘false negative’</li> </ul>	<ul style="list-style-type: none"> <li>• Asymmetry between classes in terms of standard deviations, a-priory probabilities, sample size, and misclassification costs</li> </ul>
<b>Distribution</b>	• Normal	• Other distributions, not necessarily symmetric
<b>Classifier Type</b>	• Linear, based on a separating hyper-plane	• Non-linear, based on more complex separation rules - e.g., Quadratic Discriminant Analysis (Duda and Hart, 2001)
<b>Missing-Values Pattern</b>	• Missing completely at random (MCAR)	• Patterns with certain non-random associations between missing values (e.g., MAR – Missing at Random; NMAR – Not missing at Random (Little, 1987))
<b>DQ Criterion</b>	• Missing-value defects	• Other DQ defect types – e.g., inaccurate, invalid, and/or outdated data items
<b>MQ Criterion</b>	• Confidence interval, calculated per parameter	• Other criteria that consider the entire model
<b>CQ Criterion</b>	• Minimizing classification cost	<ul style="list-style-type: none"> <li>• Maximizing accuracy, precision, sensitivity, and/or specificity</li> <li>• Maximizing classification value</li> </ul>
<b>Decision Scenario</b>	• Classification, based on a discrete set of classes	• Optimization – setting the optimal value within a continuous value range

**Table 4: Assumptions and Future Extensions**

## CONCLUSIONS

The negative impact of DQ defects on decision making has been broadly acknowledged in research and in practice. This study suggests that a possible way to understanding and quantifying this impact is by looking into the mediating role played by decision-support models. Such models are often estimated from training datasets – and when such a training dataset suffers from DQ defects, the model and the decisions that it supports are likely to be biased. This claim makes intuitive sense – however, not much was done to support it analytically. This study takes a step in that direction by offering an analytical framework that links the three levels of quality assessment - data quality, model quality, and decision quality. The analytical development demonstrated in this study is relatively simple – and its aim was to highlight and demonstrate the key concepts. As this study is still progressing – our goal is to examine comprehensive and complex decision scenarios, in which some of the assumptions made will be relaxed.

## REFERENCES

- [1] Askira-Gelman, I. GIGO or not GIGO: The Accuracy of Multi-Criteria Satisficing Decisions, *The ACM Journal of Data and Information Quality*, 3(2), Article 9, 2011, pp. 1-27
- [2] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.
- [3] Blake, R., and Mangiameli, P., The Effects and Interactions of Data Quality and Problem Complexity on Classification. *ACM Journal of Data and Information Quality*, 2 (2), Article 8, 2011, pp. 1-28
- [4] Davenport, T.H. Competing on Analytics. *Harvard Business Review*, 84(11), 2006, pp. 99-107
- [5] Duda, P. E. and Hart, D.G. S., *Pattern Classification*, Wiley & Sons, 2001.
- [6] Even, A., and Shankaranarayanan, G. Utility-Driven Assessment of Data Quality, *SIGMIS Database*, 38(2), 2007, pp. 76-93
- [7] Even, A., Shankaranarayanan, G., and Berger, P.D. Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting, *Decision Support Systems*, 50(1), 2010, pp. 152-163
- [8] Han, J. and Kamber, M. *Data Mining – Concepts and Techniques*, Elsevier, 2006
- [9] Heinrich, B., Kaiser, M. and Klier, M. A Procedure to Develop Metrics For Currency And Its Application In CRM, *The ACM Journal of Data and Information Quality*, 1(1), 2009 pp. 5-28
- [10] Little R.J.A. *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987
- [11] Liu, S., Duffy, A., Whitfield, R., and Boyle, I. Integration of Decision Support Systems to Improve Decision Support Performance. *Knowledge and Information Systems*. 22( 3), 2010, pp. 261-286
- [12] McLachlan J. G. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 1992
- [13] Parssian, A., Sarkar, S., and Varghese, S.J. Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product. *Management Science*, 50(7), 2004, pp. 967-982
- [14] Pipino, L.L., Lee, Y.W., and Wang, R.Y. Data Quality Assessment, *Communications of the ACM*, 45(4), 2002, pp. 211-218.
- [15] Redman, T.C. *Data Quality for the Information Age*, Artech House, 1996
- [16] Sessions, V., and Valtorta, M. Towards a Method for Data Accuracy Assessment Utilizing a Bayesian Network Learning Algorithm, *Journal of Data and Information Quality*, 1(3), Article 14, 2009, pp. 1-34
- [17] Shankaranarayanan G, and Kay Cai Y. Supporting Data Quality Management in Decision-Making. *Decision Support Systems*, 42(1), 2006, pp. 302-317
- [18] Shim, J.P., Warkentin, W., Courtney, J.F., Power, D. J., Sharda, E., Carlsson, C. Past, Present, and Future of Decision Support Technology, *Decision Support Systems*, 33, 2002, pp. 111-126

# CALYDAT: A METHODOLOGY FOR EVALUATING DATA QUALITY DIMENSIONS BASED ON DATA PROFILING TECHNIQUES

(Research-in-progress)

**Yonelbys Iznaga**

Universidad de las Ciencias Informáticas, Cuba

yiznaga@uci.cu

**César Guerra**

Universidad Politécnica San Luis Potosí, México

cguerra74@gmail.com

**Ismael Caballero**

Instituto de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha, España

Ismael.Caballero@uclm.es

**Abstract:** Any organization that needs to satisfy their business objectives and uses data to implement organizational processes, must have knowledge of how these data satisfy the preset quality requirements. These requirements are expressed by means of certain data quality dimensions. In some contexts, models and methodologies of data quality assessment require of mechanisms to control and monitor the level of quality of data. Proposing a methodology with a qualitative diagnosis of the data quality dimensions and using data profiling techniques to measure some of these dimensions, will have a significant impact on the processes of appropriate use of the data. The main contribution of this paper is a methodology that assesses the data quality, by diagnosing its dimensions through surveys and data profiling techniques. The paper also presents the results obtained in a real case study, which served to validate the methodology.

**Key words:** Data quality, data quality dimensions, data profiling, methodology.

## INTRODUCTION

During decades, data management has acquired a growing significance in companies, because data constitute the blood of the organization, and without them, corporations cannot align with their organizational strategy [7]. In 2002, only in the United States of America, the annual expenditure of poor data quality for enterprises was six billion dollars, according to estimations of TDWI (The Data Warehouse Institute) [20]. Because electronic data are so pervasive, data quality (hereafter DQ) plays a critical role in all business and governmental applications [1] and it is recognized as a relevant performance issue of operating processes [3]. Companies that decide to implement complex information systems such as Decision Support System (DSS), Executive Support Systems (ESS) or Enterprise Resource Planning (ERP), among others, should understand that the success of these systems also depends largely on their data. According to ISO / IEC 25012, DQ is defined as "*the degree to which the characteristics of the data are suggested conditions and needs when used under specific conditions.*" [9].

Therefore, before any operation, it is important to assess the suitability degree of the use of data involved in the task, according to the context in which they are. Data profiling is one of the techniques that helps diagnose the DQ in specific contexts, which is the "*data analysis systems to understand its content, structure, quality and dependencies*" [4]. Indeed, doing data profiling and monitoring the defects of data, are useful activities for assessing DQ in specific contexts.

Although, nowadays there exists some models, methodologies and tools to carry out the data profiling processes, in our context, it is possible to find some needs such as: assessment of some characteristics of DQ using techniques and tools of data profiling, definition of roles and responsibilities for the DQ control, organization of the process through the use of artifacts and documents and the frequent reporting to the organization of the DQ diagnosis, depending on user types and level securities, in order to involve and engage members and roles that interacts with these data.

This paper is organized in five main sections. Section II describes existing methodologies for DQ control and assessment, and data profiling models that currently exists. Section III presents CALYDAT, the proposed methodology and a description of their characteristics, principles, scope, processes, activities and people in charge. Section IV presents the obtained results from applying the proposed methodology in a real context. Finally, Section V presents the conclusions and the main intentions for future work.

## BACKGROUND

### ***Methodologies for DQ assessment***

Many authors have made contributions for DQ. Several of these have offered the most relevant categorizations of DQ dimensions, such as in [10, 14, 17, 21, 22, 25]. This research was based on the data quality characteristics introduced in ISO/IEC 25012, which are: accuracy, completeness, consistency, credibility, timeliness, accessibility, compliance, confidentiality, efficiency, traceability, portability, understandability, availability and recoverability [9].

For making a comparative study of existing methodologies for DQ assessment, we consider several aspects, including: dimensions used, cost and types of data and information systems involved. The methodologies for DQ assessment and improvement have been classified in four categories [1]:

- complete methodologies*, which provide support to both the assessment and improvement phases, and address both technical and economic issues;
- audit methodologies*, which focus on the assessment phase and provide limited support to the improvement phase;
- operational methodologies*, which focus on the technical issues of both the assessment and improvement phases, but do not address economic issues;
- economic related methodologies*: which focus on the evaluation of costs.

This research is based on audit methodologies. Some of these methodologies are AIMQ [13], CIHI [26], AMEQ [20] and IQM [5]:

**AIMQ** Methodology: (*A Methodology for Information Quality Assessment* [13]): It is the only methodology of information quality based on benchmarking. It draws heavily on the PSP/IQ model (Table 1), which classifies the DQ dimensions according to the interest and priority of users and administrators. **AIMQ** has four classifications for DQ: comprehensive, reliable, useful and usable, into which DQ dimensions fall. It uses questionnaires for the identification and diagnosis of both DQ dimensions and measures of information quality.

	<i>Conforms to specifications</i>	<i>Meets or exceeds the customer expectations</i>
<b><i>Product Quality</i></b>	<i>Sound information</i>	<i>Useful information</i>
<b><i>Service Quality</i></b>	<i>Dependable information</i>	<i>Usable information</i>

**Table 1. The PSP/IQ model.**

**CIHI** methodology (*Canadian Institute for Health Information* [26]): **CIHI** focus on the control of DQ of data stored in the Canadian Institute of Health Information, specifically in the monitoring of the size, heterogeneity and quality of the stored data. Data quality evaluation is based on a four-level hierarchical model. At the first level, 86 basic quality criteria are defined. These criteria are aggregated by means of

algorithms of composition into 24 quality characteristics at the second hierarchical level, and finally, these are aggregated into five DQ dimensions at the third level. Finally, the five dimensions are aggregated into one overall database evaluation at the fourth level.

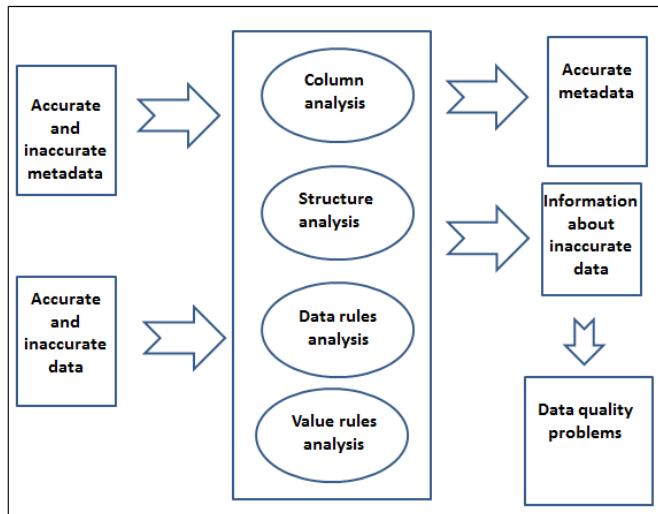
**IQM** methodology (*Information Quality Measurement* [5]): **IQM** conceived the provision of a quality framework adapted to the Web data. Among its entries, besides of the quality criteria, it has the tools and techniques used to measure the DQ. The result of evaluation is the most important outputs, which is a valuable guide for selecting and customization of the tools used by web administrators for creating, managing websites. IQM describes the following main phases: assessment planning, assessment configuration, Measurement and follow-up activities, where the most important processes are: the diagnosis of the data, the requirements analysis and evaluation of the DQ.

**AMEQ** methodology (*Activity-based Measuring and Evaluating of Product information Quality* [19]): **AMEQ** provide a rigorous basis for Product Information Quality assessment and improvement in compliance with organizational goals. The methodology is specific for the evaluation of DQ in manufacturing companies, where product information represents the main component of operational databases. In manufacturing companies, the association between product information and production processes is straightforward and relatively standard across companies [1]. AMEQ has five phases. The first one assesses the cultural preparation of the organization. The second one focuses on all information related to the product by process modeling and identification of critical areas. One of the outputs of this phase is a model of measurement techniques. The third phase focuses on the implementation of all activities and techniques for the measurement and evaluation. During the fourth phase the causes of DQ problems that have been detected after diagnosis of the dimensions will be investigated. The last one is responsible for monitoring and improving the quality of product information, through mechanisms of accountability of the processes and data.

After studying the characteristics of these audit methodologies, we consider that they are very useful, depending on its features and goals. However, according with some aspects like: the focus on the business processes of the organization, the definition of roles and responsibilities, the use of artifacts for document the process and the inclusion of data profiling techniques for the DQ evaluation; we conclude that, except AMEQ that utilizes the organizational processes in its process modeling, the rest of the methodologies are not based on business processes. They do not use roles and responsibilities in its phases and activities, they do not include data profiling techniques for the DQ evaluation and only CIHI has a well-defined documentation process.

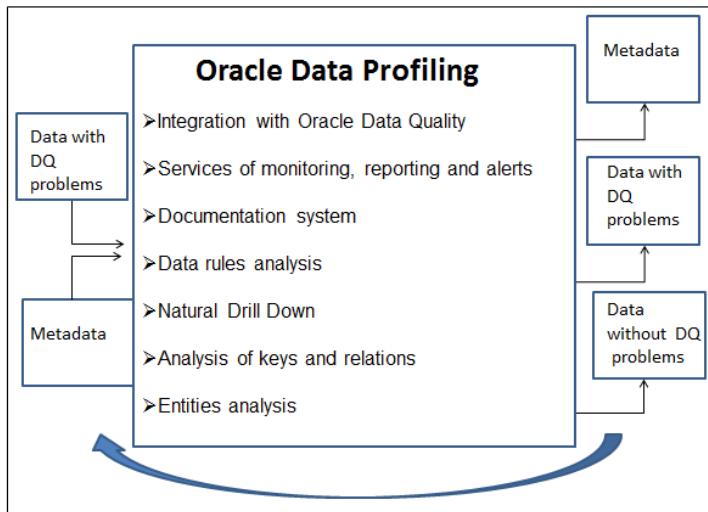
### ***Data profiling models, techniques and tools***

Several data profiling methods and techniques also contribute to the necessary assessment for the DQ control, where the fundamental approach is performed on the data collections. The DQ dimensions more widely used to assess DQ are: correctness, completeness and accuracy. One of the models available today is [4], which consists of one or more inputs of data and metadata, the application of research techniques, and as outputs, corrected metadata and information related with data, as shown in Figure 1.



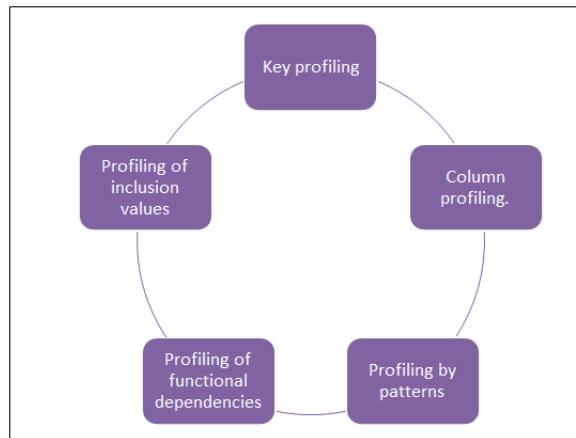
**Fig. 1. Data profiling model of [4]**

Oracle Corporation, a company that has developed a system for profile data, is oriented to the thorough investigation and close monitoring of its quality [15]. With a tool named *Oracle Data Profiling*, the user has the possibility to discover and infer rules based on data, and monitor their quality over time. As shown in Figure 2, the inputs and outputs are well defined, where data and metadata that were profiled, can be profiled again.



**Fig. 2. Data profiling techniques and process of [15]**

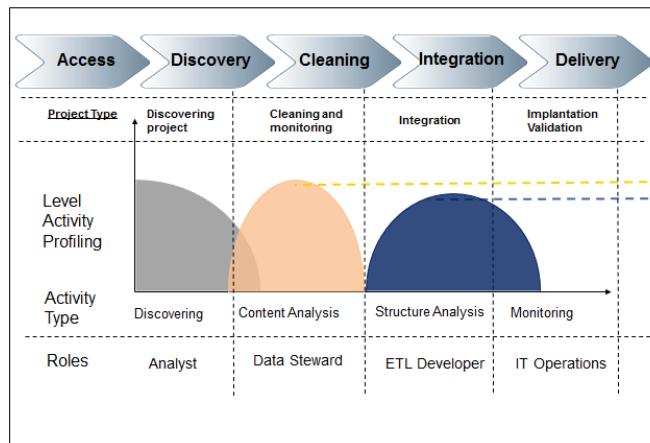
Microsoft offers a tool named Data Quality Services 2008 [11], with techniques and mechanism of data profiling, such as: candidates keys profiling, column profiling, data profiling using patterns, functional dependences profiling, and inclusion values profiling [11].



**Fig. 3. Data profiling model of [11]**

The Embarcadero Company is noted for its software design: *ER/Studio*. Through its *CA ERwin Data Profiler* tool, the user can combine the analysis and the data modeling in a practical way. In its own model highlights four key activities: analysis column, integration with data models, the discovery of keys and Extended Analysis of attributes [6].

Informatica Corporation [27], with its tool named *Informatica PowerCenter*, an enterprise platform that offers access, research, data profiling and data integration from any data source, and any format. It is a very important tool for data profiling and diagnoses the DQ. As shown in Figure 4, *Informatica PowerCenter* has five subsystems: Access, Discovery, Cleaning, Integration and Delivery [23].



**Fig. 4. Subsystems, levels, roles, activities, and techniques of data profiling of *Informatica Power Center***

## CALYDAT: A METHODOLOGY FOR DATA QUALITY CONTROL, ANALYSIS AND EVALUATION BASED ON DATA PROFILING TECHNIQUES

The contribution of this paper is a methodology to control, analyze and evaluating of DQ through the use of data profiling techniques and diagnosis of the DQ attributes. It consists of three phases, each of which contains processes, activities, artifacts, people in charge and tools. It is named as: *Methodology for the Control, AnaLYsis anD EvaluATion of Data Quality based on Data Profiling Techniques* (CALYDAT).

## **Scope of CALYDAT**

CALYDAT guides to the establishment of the control of DQ, based on the diagnosis of DQ dimensions and processes related with data profiling of relational databases, where activities, techniques and mechanisms are involved, as a guide for its implementation.

Unlike the others audit methodologies; CALYDAT is based on business processes of the organization, besides it defines roles and responsibilities for a better organization of the execution of its phases and activities. CALYDAT also proposes well-defined artifacts that help for documenting the implementation of the methodology, and it includes an added value: the using of the results of data profiling techniques for the DQ evaluation.

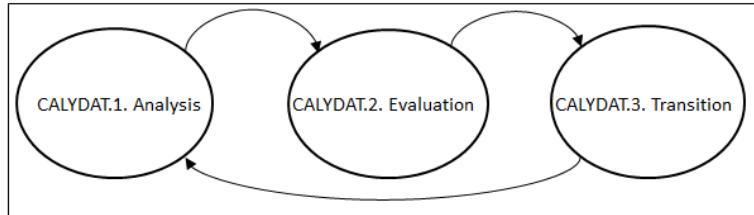
## **Fundamentals of CALYDAT**

CALYDAT is based on the following pillars:

- *It is focused on the DQ control of each organizational business process:* Its main objective is to implement a DQ monitoring system of the organizational process analyzed, and if a DQ problem raises, enable the possibility of detecting when occurred, the area, database or information system where the problems happened, and who are the people in charge.
- *Implication of roles that manage data:* It is based on committing all the roles involved in data management, which are in charge for monitoring or controlling the data quality.
- *Iterative and incremental:* Once a CALYDAT development cycle is completed, it should be executed again so that each iteration will cover each of the organizational processes involved in access, control and management of data in the organization.

## **Representation of CALYDAT**

CALYDAT is based on three phases: Analysis, Evaluation y Transition, as shown in the figure 5:



**Fig. 5: Phases representation of CALYDAT.**

The following subsections provide details of each one of the phases of the methodology:

### **CALYDAT.1. Analysis**

At this phase, the current status of a particular organizational process is studied, this implies to take into account the types of existing users, the data types, database administrators, etc., for preparing the infrastructure for the application of data profiling techniques and for the survey of diagnostic of DQ dimensions. In new iterations new organizational processes will be diagnosed. Table 2 shows their characteristics:

<b>Input Products</b>	Description of the organizational process
<b>Output Products</b>	Identified information, selected dimensions for the DQ evaluation.
<b>Activities</b>	1.1. Diagnosis, 1.2. Election of Requirements .
<b>Methods, techniques and tools</b>	Expert judgment, brainstorming, artifact for the Diagnosis of organizational process (please see artifacts presented in Appendix A)
<b>Roles</b>	Business analyst, DQ analyst

**Tab. 2: Phase of analysis.**

In this phase, the activities are:

#### **CALYDAT.1.1. Diagnosis**

It must be executed to get a first assessment of the current status of the selected organizational processes. For doing so, it is necessary to take into account the databases used by the selected organizational processes, the user types and the existing roles, the types and formats of data handled by the organization and database administrators. As input of this activity, aspects related with the diagnostic process should be provided, and as output, the identified information related with the organization are to be generated.

#### **CALYDAT.1.2. Election of the Requirements**

The DQ dimensions selected will be involved in the entire cycle of execution of the methodology for each one of the selected organizational processes. As input of this activity, some aspects of the diagnostic process must be provided, and as output, the list with the selected DQ dimensions should be generated.

#### **CALYDAT.2. Evaluation**

In this phase, an evaluation of the level of DQ of a relational database should be performed. This implies the use of some techniques like structure profiling, relational profiling, data rules profiling and the implementation of surveys for the diagnosis of DQ dimensions Table 3 shows their characteristics:

<b>Input Products</b>	Result of the diagnosis of organizational process, data source, metadata source
<b>Output Products</b>	Data profiled, metadata profiled, result of the survey for the diagnostic of DQ dimensions
<b>Activities</b>	2.1. Structure profiling, 2.2. Relational profiling, 2.3. Data rule profiling, 2.4. Conductions of a Survey for the diagnostic of DQ dimensions
<b>Methods, techniques and tools</b>	Profiling of table structures, and its functional dependences, data rules profiling, questionnaire of the survey for the diagnostic of DQ dimensions (see Appendix B), data profiling tools
<b>Roles</b>	Business analyst, DQ analyst, database administrator, database designer

**Table 3: Phase of evaluation.**

To achieve the goals of this phase, the team should execute the following activities:

#### **CALYDAT 2.1. Structure profiling**

It consists of thoroughly investigate each one of the columns and rows of tables in the source systems, applying a set of techniques to calculate statistical information and metadata. The most significant DQ dimensions are completeness, accuracy and precision. As input of this activity, services of data access, profiled and not-profiled data and metadata should be provided, and as output, artifacts, data profiled and metadata profiled are to be generated.

**Property profiling:** It refers to applying profiling techniques to determine table properties, such as number and percent of null values, unique, duplicates, blanks, data types, minimum and maximum size of characters, maximum and minimum values and domains, among others.

- **Regular expressions profiling:** It refers to applying pre-defined regular expressions to identify matches with the values of the attributes. You can define new expressions or use existing ones.
- **Language profiling:** Getting profiles of natural language terms and language elements stored as data, is very complex during the data profiling process. In this case, the domain plays an impor-

tant role in defining the dominant values in a column, and defining which values are written or spoken like a specific values (Matching Writing and Matching Sound respectively). For this, regular expressions or SQL statements and stored procedures or functions can be used. A repository of terms can also be used to store the letters or vowels of the alphabet and verify matches with the values of the columns.

### CALYDAT 2.2. Relational profiling

The main aim of this activity is to determine possible relationships and functional dependencies between tables or business objects, and discovering primary and foreign keys. With this activity it is possible to evaluate the degree of consistency. According to [1], the DQ dimension consistency refers to the violation of semantic rules defined on a data or a particular data set. In this case it will be profiled the violations of integrity constraints, specifically inter-relational constraints. As input, the services of data access, data and metadata profiled and unprofiled are to be provided; and as output, artifacts, primary keys, foreign keys, relationships between entities and the relational matrix should be generated.

For this case, several rules, SQL statements and data mining techniques can be applied [24], for example association rules [24], to find dependency percentages of some attributes related to others, and thus find possible foreign keys.

Business analyst, DQ analyst, database designer can use the following techniques and tools to achieve their objectives:

- *Analysis of primary key:* It is used to determine those values in the attributes that are unique and are candidates for primary keys.
- *Analysis of foreign key:* It is used to determine those attributes that have been detected from the rules of inclusion and the relational matrix. The associations identified can be used to predict behavior, and to reveal correlations and occurrences of events [24]. To evaluate the rules, the support is used. As shown below, Equation 1.1 indicates the number of cases covered by the rule, and confidence; Equation 1.2 indicates the number of values of one item that belongs to another item; and Equation 1.3, referred to the confidence, it indicates the number of cases correctly predicted by the rule. Confidence is expressed as the ratio between the number of cases in which the rule is met and the number of cases in which it applies, because the premises are satisfied.

If we consider the following item  $I = \{A, B, C, D, E\}$  where A, B, C, D, E are attributes of a particular database:

$$\text{Support}(A) = P(A) \quad \text{Equation. 1.1}$$

$$\text{Support}(A \subseteq B) = P(A \subseteq B) \quad \text{Equation. 1.2}$$

$$\text{Confidence}(A \subseteq B) = P(B | A) = \frac{P(A \subseteq B)}{P(A)} \quad \text{Equation. 1.3}$$

Where  $P(A)$  is the total value of the attribute A and  $P(A \subseteq B)$ , the number of values of attribute A that belongs to attribute B, where B can be repeated. Confidence is the ratio between both. This will determine the confidence of each of the attributes related with the rest, identifying the higher value, which constitute potential foreign keys.

- *Relational Matrix specification:* Technique that uses a two dimensional array to detect high levels of confidence from the result of applying the rules of inclusion, and thus identify possible relationships between attributes. The intersection of two attributes corresponds to a functional dependency between them, being represented by a box with a percentage value. This value is the confidence that exists between the two attributes. An example of a relational matrix is shown in Table 4, where the highlighted values represent the largest confidences (86.3, 90.5, 91.2, 76.1, 100, 79.5, 90.2, 94.1, 100, 100 and 78.4), and therefore, possible relationships between the attributes (A, B), (A, C), (B, H), (C, A), (E, C), (E, F), (E, G), (F, E), (F, G), (G, E) and (H, D), respectively.

As tools, any of the data profiling tools presented in section 2.2 could be proposed.

	A	B	C	D	E	F	G	H
A	34,2	76,1	48,6	10,4	54,7	15	27,9	
B	86,3		45,8	16,6	21,5	6,9	5,7	72
C	90,5	56,9		0,3	100	7,9	1,8	9,2
D	37,6	19,4	1,2		0	0	0	78,4
E	21,4	11,8	60,5	0		94,1	100	0
F	39,8	23,5	4,6	0	79,5		68,3	0
G	0,3	88	3,7	0	90,2	100		1,1
H	16,7	91,2	28	71,3	0	0	2,6	

Table 4: Example of relational matrix.

#### CALYDAT 2.3. Data rules profiling.

Activity aimed to the researching, discovering, verification and validation of data rules. It helps to specify the degree of conformity, which determines whether the data has attributes that adhere to standards, conventions or regulations and similar rules relating to DQ in a specific context of use [11]. As input of this activity, services of data access, data and metadata profiled and unprofiled, and as output, artifacts and data rules.

Business analyst and DQ analyst could use the following techniques and tools to get the specified results:

- *Analysis of default data rules:* Based on existing data rules in information systems or in databases of the organization, are checked to see if the results match with what is expected of them.
- *Discovery of data rules:* These rules are conditions that may involve one or more columns. They generally use conditionals like (if, then, <,>, =).

As tools we propose the data profiling tools that implement data rules profiling.

#### CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ evaluation.

The survey is a system for collecting information to describe, compare and explain knowledge, attitudes and behavior [12]. In this process a qualitative diagnostic of DQ is performed from the application of the survey. As input of this activity, the result of organizational analysis is to be provided, and as output, the result of the survey of DQ dimensions should be produced. With the aim of improving the diagnosis and evaluation of DQ, there are five types of users to whom the survey is proposed. Table 5 shows some examples of types of users:

User types	Examples of user types
Data user	Database administrators, database developers, ETL specialists, etc.
Requirement user	Requirement analysts, requirement specialists, etc.
Technology user	Network administrators, server administrators, IT specialists , etc.
Business user	Business analysts, executives, leaders, managers, customers, area and department directors, final user, etc.
Interface user	Web programmers, designers, ads and marketing specialists, etc.

Table 5: User types and examples of user types.

Business analyst, DQ analyst and database administrator can use the following techniques and tools to achieve their objectives:

- *Conducting of survey for the diagnostic of DQ dimensions:* This method is based on the DQ characteristics provided by the ISO/IEC 25012. It should be conducted periodically to the members of the organization that interact and manage data involved in the organizational process, with questions related to each of these DQ dimensions, so as to provide a qualitative and quantitative value of the level of quality of the data used within the organization. The Details on the survey can be seen in Appendix B. As proposed tools, the data profiling ones and the questionnaire can be suggested.

### 3.1.1 CALYDAT.3. Transition

In this phase the organizational process analyzed is monitored, continuing to the analysis. It should be reported the status of the DQ, to all roles and members involved in the organizational business process. It implements activities related to the process of monitoring and alerting DQ. Table 6 shows their characteristics:

<b>Input Products</b>	Result of the survey for the diagnostic of DQ dimensions.
<b>Output Products</b>	Artifacts, notifications and alerts.
<b>Activities</b>	3.1. Monitoring and control
<b>Methods, techniques and tools</b>	Notification, and alerts of DQ
<b>Roles</b>	DQ analyst

Table 6: Phase of transition.

#### CALYDAT 3.1. Monitoring and control

The goal of this activity is to notify and alert events related with the detection of poor DQ in any of the selected business processes of the organization. The people in charge should ensure the beginning for repeating the phase of analysis in a new organizational process. As input, the result of the survey for diagnostic the DQ dimensions should be entered, and as output, the specification of the artifacts, notifications and alerts should be generated.

DQ analyst can uses the following technique and tools to achieve their objectives:

- *Execution of the monitoring and alert:* CALYDAT proposes the implementation of a reporting solution, for the notification of the DQ dimensions assessment to members and roles related with the organizational process, about the current diagnostic of DQ in that process.

## RESULTS

In order to test the applicability of CALYDAT in a real environment, we used the methodology in an organization with well-defined business processes. Concretely, CALYDAT was applied to one business process named Control of mobile devices, where its main objective is to manage each mobile device in the agricultural fields where these mobiles work, its exact location, if they are stopped or moving, the fuel consumed, the kilometers traveled, etc. Obtained results are to be presented in this section.

We decided to apply CALYDAT to this scenario because of its own characteristics, for example, the organization has well-defined business processes, its data are stored in relational databases, it is possible to apply data profiling techniques for evaluating the DQ, also because it is a well-defined and complex organizational process where it is recommendable the use of artifacts that guide and document the application of CALYDAT. This experience involved the execution of techniques and activities of CALYDAT and the application of the survey (see the section CALYDAT 2.4. Survey) for the diagnostics of DQ dimensions.

Firstly, the members that will play the role of DQ analysts were identified; they would be the people in charge for the application of CALYDAT. DQ analysts and managers planned jointly the execution of the phases of the methodology, ensuring the availability of resources for the corresponding iterations in a periodical application of CALYDAT to other business processes of the organization.

Let's explain the application of each phase of CALYDAT to the business process Control of mobile devices. For carrying out the phase of **CALYDAT.1. Analysis**, the activity of Diagnostic was performed, where the concepts and features of the business process were identified, using the artifact Diagnosis of organizational process (see Appendix A).

<b>Activities</b>	<b>Selected DQ dimensions</b>
CALYDAT 2.1. Structure profiling CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Accuracy
CALYDAT 2.1. Structure profiling CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Completeness
CALYDAT 2.2. Relational profiling CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Consistency
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Credibility
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Currentness
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Accessibility
CALYDAT 2.3. Data rules profiling CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Compliance
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Confidentiality
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Efficiency
CALYDAT 2.1. Structure profiling CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Precision
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Traceability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Understandability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Availability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Portability
CALYDAT 2.4. Conduction of a Survey for the diagnostic of DQ dimensions	Recoverability

**Table 7: Selected dimensions for the application of CALYDAT.**

In addition, the dimensions that will be involved in the DQ evaluation were selected, as shown in the Table 7, where the *Activities* column refer to the activities of the phase of CALYDAT.2. Evaluation, where the DQ dimensions selected will be evaluated, and the *Selected DQ dimensions* column refers to the DQ dimensions that will be evaluated in each activities:

During the phase of **CALYDAT.2. Evaluation**, taking into account that the business process of the control of mobile devices and its data sources are stored in a relational database, we performed the activity of profiling structure, where highlighted the attributes **operation\_date**, **mobile\_state**, **year**, and **crop\_cycle** (as shown in Table 8), which presented DQ problems, particularly with the completeness dimension. For example the attribute **date\_operation** presented 17.5% of null values, the attribute **mobile\_state**, 6.5% of null values and the attribute **year**, a minimum value of 1278.

Attribute	Entity	Description of the attribute in the data model
operation_date	operation	Attribute that stores the date of the operation performed by the mobile device
mobile_state	mobile	Attribute that stores the state of the mobile device in the mobile entity
Year	operation	Attribute that stores the year that the operation was performed
crop_cycle	crop	Attribute that stores the number of cycles of an agricultural crop determined
state_code	mobile	Attribute that stores the state code in the mobile entity.
state_id	state	Attribute that stores the identifier of the status of the mobile in the state entity
device_id	device	Attribute that stores the identifier of the tracking device

Table 8: Attributes susceptible to receive a data profiling analysis of the control of mobile devices.

After checking the degree of completeness, checking if all values for each row are complete or not, the result obtained is shown in Table 9, by each of the entities in the database (see figure 6 and Table 10):

Business entities	Percentage of the evaluation result of completeness
mobile	23 rows with incomplete values, 312 rows in total: 92,62 %
crop	7 rows with incomplete values, 215 rows in total: 96,74 %
device	49 rows with incomplete values, 378 rows in total: 87,04%
state	4 rows with incomplete values, 37 rows in total: 89,19 %
operation	18 rows with incomplete values, 193 rows in total: 90,67 %

Table 9: Verification of the degree of completeness

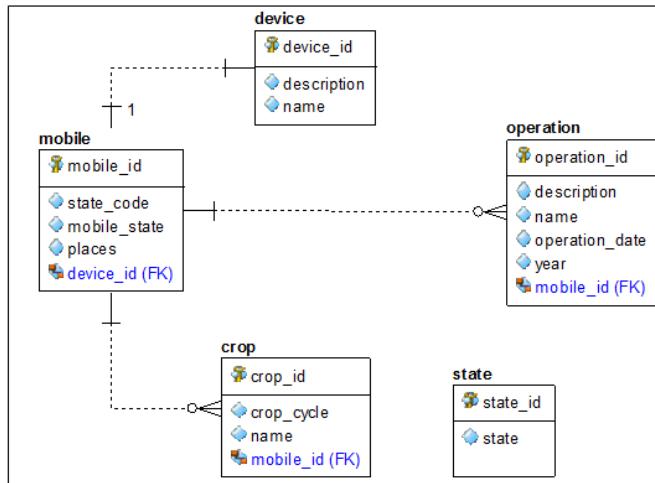


Fig. 6. Entity-Relation model of the control of mobile devices.

Entity	Description of the entity in the data model
Mobile	Entity that stores mobile devices: such as tractors, trucks, jeeps, etc.).
Crop	Entity that stores agricultural crops where worked the mobile devices
Device	Entity that stores the tracking devices carried by mobiles
State	Entity that stores the state of mobile devices
Operation	Entity that stores the operations of mobile devices

Table 10: Profiled entities of the control of mobile devices.

After the phase of CALYDAT.1. Analysis, specifically in the activity of diagnosis, it was noted that entities should not have rows with incomplete values, because in other processes it performed percentage calculations using these values, so that at least one row with incomplete values in some of these entities, represents a negative impact to the DQ dimension of completeness.

In the activity of CALYDAT 2.2. Relational profiling, after the processing, two attributes were detected as potentially relatable (state\_code and state\_id, see Table 8) between two unrelated entities (mobile and state, see Figure 6 and Table 10) with a 98.4% of confidence (see the section CALYDAT 2.2. Relational profiling, specifically the technique Analysis of foreign key). This helped to discover a violation of referential integrity, specifically in the dimension consistency.

Based on predefined business rules and in order to diagnose the DQ dimension of compliance of the organizational process analyzed, it was found that some rules were not complied with, such as the attribute values crop\_cycle which must be in the range between 1 and 100, and were found values such as 134, 121, 106 and 189. Also that the attribute device\_id must be unique, and was found the value 008 repeated twice, and the value 014 repeated three times. In crop\_cycle, attribute with data type varchar, were found 7 strings, and according to business rules, should store only numeric values.

During the activity of CALYDAT 2.4. Survey, it was applied the questionnaire (see Appendix B). Candidates to participate in the survey were chosen from members who work directly with the analyzed business process. The members are the database administrators and workers of the technology department, related to the business process of control of mobile devices. In total there were 12 members: three (3) database administrators, four (4) network administrators, two (2) server administrators, one (1) security specialist, the manager and the vice-manager of technology. The result of the survey is shown in Table 11. The average column corresponds to the average values for each dimension of all applied surveys, and it is a value ranged between 0 and 5.

According with the context where data are used, in this case data are used for storing and managing information related with the exactly location of mobile devices, so the values of Table 11 become relevant. The highest percentage values, corresponds to DQ dimensions which quality is adequate. Conversely, the lower percentages are the DQ dimensions with data quality problems. As result, the critical dimensions that need an urgent attention are: compliance, precision and recoverability.

Dimensions	Average	%
Accuracy	4,12	82,4
Completeness	4,37	87,4
Consistency	4,10	82
Credibility	3,79	75,8
Currentness	4,05	81
Accessibility	4,17	83,4
Compliance	2,21	44,2
Confidentiality	3,78	75,6
Efficiency	4,19	83,8
Precision	2,92	58,4
Traceability	3,75	75
Understandability	4,56	91,2
Availability	3,98	79,6
Portability	4,46	89,2
Recoverability	2,73	54,6

**Table 11: Results of the survey for the diagnostic of DQ dimensions of the process analyzed.**

During the execution of the phase of **CALYDAT.3. Transition**, we proposed the creation of a web site in the organization, with the corresponding levels of access, and based on the types of users. The notification of the diagnosis of DQ dimensions should be weekly. It was advised to the managers that they should select the data profiling tool, according to their needs and possibilities, and repeat the survey frequently, including others business process of the organization.

## CONCLUSIONS

This main contribution of this paper is CALYDAT, a methodology for the analysis, control and evaluation of DQ, through data profiling techniques and the application of surveys for the diagnostic of DQ dimensions, to various types of users. Its application in a real environment was satisfactory and provided the expected results, giving to the managers and members involved in the organizational process of control of mobile devices, a quantitative and qualitative evaluation of DQ. The type of user plays a fundamental role, which offers the possibility to detect more effectively the DQ problems, based on the role to which is directed the survey. As research methods during the process of developing the methodology, we used theoretical and empirical methods [28], including the method of survey. We empirically obtained the DQ dimensions used in CALYDAT, the types of users to which the questionnaire should be applied, the roles and responsibilities defined and the output products of the analysis phase. For the success of CALYDAT, it was necessary to consider the systemic method as a combined and integrated system of all phases and activities, with an iterative and incremental approach. Finally, the survey plays a key role for the evaluation of the DQ in CALYDAT.

In the future, we intend to develop a tool that supports the application of CALYDAT. This tool will have functionalities that allow execute data profiling techniques, and mechanisms for diagnosis the DQ dimensions: Accuracy, Completeness and Precision.

## BIBLIOGRAPHY

- [1] Batini C., Cappiello C., Francalanci C. and Maurino A., *Methodologies for Data Quality Assessment and Improvement*, ACM Computing Surveys, Vol. 41, No. 3, Article 16, 2009.
- [2] Catarci, T., and Scannapieco, M. 2002. *Data quality under the computer science perspective*. Archivi Computer 2.
- [3] Chengalur-Smith, I. N., Ballou, D. P., and Pazer, H. L. 1999. *The impact of data quality information on decision making: An exploratory analysis*. IEEE Trans. Knowl. Data Eng. 11, 6, 853–864.
- [4] E. Olson, Jack. *Data Quality—The Accuracy Dimension*. San Francisco, Elsevier Science, 2003.
- [5] Eppler, M. and M'Unzenmaier, P. 2002. *Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology*. In Proceedings of the 7th International Conference on Information Systems (ICIQ).
- [6] Erwin Studio, CA ERwin Data Profiler. 2009 [Accessed in 2009 October]; Available from: <http://www.ca.com/us/products/Product.aspx?ID=8235>
- [7] Ferdinandi, P.L. *Data warehouse advice for managers*, New York: AMACOM American Management Association., 1999.
- [8] Gomes, J.F., Maria José Trigueiros, *A Data Quality Metamodel Extension to CWM*, in *4th Asia-Pacific Conference on Conceptual Modelling (APCCM 2007)*. 2007: Australian.
- [9] ISO-25012, *ISO/IEC 25012: Software engineering - Software Product Quality Requirements and Evaluation (SQuaRE) - Data quality model*. 2008.
- [10] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P., Eds. 1995. *Fundamentals of Data Warehouses*. Springer Verlag.
- [11] Wong, A., Sutcliffe D., (2009). *Data Quality Services 2008*. 2008 [Accessed in 2009 Septem-

- [12] ber]. Available from: <http://msdn.microsoft.com/en-us/library/dd129900%28v=sql.100%29.aspx>
- [13] Lawrence, S., Kitchenham B., "Principles of Survey Research. Part 1: Turning Lemons into Lemonade" *Software Engineering Notes*, 2001. vol 26 no 6 pp. 16
- [14] Lee, Y.W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. *AIMQ: A methodology for information quality assessment*. *Inform. Manage.* 40, 2, 133–460.
- [15] Naumann, F. 2002. *Quality-driven query answering for integrated information systems*. Lecture Notes in Computer Science, vol. 2261.
- [16] Oracle Corporation, Oracle Data Profiling. 2007 [Accessed in 2012 June]; Available from: [http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledp\\_datasheet.pdf](http://www.oracle.com/technology/products/oracle-data-quality/pdf/oracledp_datasheet.pdf)
- [17] Microsoft Corporation, Microsoft Developer Network, MSDN [Accessed in 2012 July]; Available from: <http://msdn.microsoft.com/en-us/library/ff877917.aspx>
- [18] Redman, T. 1996. *Data Quality for the Information Age*. Artech House.
- [19] Rhind, Graham. *Poor quality Data. The pandemic problem that needs addressing*. 2007.
- [20] Su, Y. and Jin, Z. 2004. *A methodology for information quality assessment in the designing and manufacturing processes of mechanical products*. In Proceedings of the 9th International Conference on Information Quality (ICIQ). 447–465.
- [21] TDWI- The Data Warehouse Institute. 2009 [Accessed in 2009 October]; Available from: [http://tdwi.org/research/2009/09/mr-who-ensures-clean-consistent-data.aspx?sc\\_lang=en](http://tdwi.org/research/2009/09/mr-who-ensures-clean-consistent-data.aspx?sc_lang=en)
- [22] Wand, Y. and Wang, R. 1996. *Anchoring data quality dimensions in ontological foundations*. Comm. ACM 39, 11.
- [23] Wang, R. and Strong, D. 1996. *Beyond accuracy: What data quality means to data consumers*. J. Manage. Inform. Syst. 12, 4.
- [24] Minguez, A. *Fundamentos de Calidad de datos*. [Accessed in 2012, March]. Available from: <http://seminarisempresa.fib.upc.edu/aulasempresa/2009/programes/POWERDATA.html>
- [25] R. Agrawal and R. Srikant. (June 1994). *Fast algorithms for mining association rules in large databases*. In Research Report RJ 9839, IBM Almaden Research Center, San Jose, CA.
- [26] Bovee, M., Srivastava, R., and Mak, B. September 2001. *A conceptual framework and belief-function approach to assessing overall information quality*. In *Proceedings of the 6th International Conference on Information Quality*.
- [27] Long, J. and Seko, C. April 2005. *A cyclic-hierarchical method for database data-quality evaluation and improvement*. In *Advances in Management Information Systems-Information Quality Monograph (AMISIQ) Monograph*, R. Wang, E. Pierce, S. Madnick, and Fisher C.W.
- [28] Free informatica tutorials; [Accessed in May, 2012]. Available from: <http://free-informatica-tutorials.blogspot.com/>
- Hernández, R. and Coello, S. 2002. *El paradigma cuantitativo de la investigación científica*. EDUNIV. La Habana, Cuba. 82-96.

## APPENDICES

### Appendix A

#### Diagnosis of organizational process

##### Deliverable

<Organization name>

<Organizational process name>

<Version>

##### Version control

Date	Version	Description	Author
dd/mm/yy	<x.x>	<Details>	<Name>

##### Introduction

##### Purpose

[Define the main objective for the evaluation and diagnosis of organizational process.]

##### Scope

[It specifies which business processes and DQ dimensions shall apply. In this case the artifact for the diagnosis of organizational process will integrate with the survey for the Diagnosis of the DQ dimensions.]

##### References

[List of referenced documents]

Code	Title
[1]	Document 1
[2]	Document 2

##### Glossary

[In the glossary specifies a group of basic terms that are managed for the diagnosis of organizational process.]

##### Description of the diagnosis application

[It describes the implementation strategy for the diagnosis of the organizational process.]

##### Summary of the diagnosis in the business process:

[Summary of the results of the diagnosis of organizational process.]

##### Analysis of significant results:

[Analysis of the most relevant results obtained in the diagnosis and a summary of the main factors to consider.]

##### Conclusions

[Conclusions of the diagnosis of the organizational process.]

## **Appendix B**

### **Survey for the diagnostic of DQ dimensions**

#### **Deliverable**

<Organization name>

<Organizational process name>

#### **Introduction**

This survey is defined for the investigation and the correct diagnosis of the DQ dimensions, where the participation of roles and members who interact and use the data is very important. Below are a number of aspects which should be marked with an X the value in the scale that is considered appropriate to characterize the current state of data quality dimensions. In case of indecision or ignorance in any aspect, please do not make any X in the corresponding aspect. The collection is a term used in the survey for referring to data or data set that will be analyzed by each of the dimensions.

#### **General aspects:**

Line/Area/Group where it belongs: \_\_\_\_\_

Role played: \_\_\_\_\_

Alternatives to respond to an aspect are listed below:

Nomencla-ture	A	B	C	D	E
Qualitative equivalence	Yes, quite	Yes, but not enough	Little	Very little, almost none	No, none

**Survey: Diagnostic of DQ dimensions**

Taking into account the following initial requirements:

The business area to diagnose: \_\_\_\_\_

The business concept to diagnose: \_\_\_\_\_

The source, data source or agent in charge (person or system) to enter data: \_\_\_\_\_

The data or the data set that must be evaluated:

In the range of time: From: \_\_\_\_\_ (day/month/year) To: \_\_\_\_\_ (day/month/year)

		A	B	C	D	E
Accuracy	Does the collection have the value and the actual characteristics expected?					
Completeness	Is the collection completed and has all the expected values?					
Consistency	Is the collection free of inconsistencies, contradictions in relation to other data?					
Credibility	Does the collection have adequate credibility and reliability?					
Currentness	Do you think the collection is updated with respect to the specified time range or with respect to the current time?					
Accessibility	Can be the collection properly managed through its access?					
Compliance	Does the collection comply with business rules or restrictions?					
Confidentiality	Does the collection have the appropriate confidentiality and security?					
Efficiency	Does the collection have the expected levels of efficiency and performance?					
Precision	Does the collection have the adequate accuracy and precision?					
Traceability	Is the access to the collection being audited by traces or tracks?					
Understandability	Is the collection understandable and interpretable by users?					
Availability	Can the collection be properly retrieved by authorized users or applications?					
Portability	Will maintain the collection its quality if is moved from one system to another?					
Recoverability	Will maintain the collection its quality despite occurrences of failures?					

# KEY-BASED BLOCKING OF DUPLICATES IN ENTITY-INDEPENDENT PROBABILISTIC DATA

(Research-in-Progress)

**Fabian Panse**

University of Hamburg, Germany  
[panse@informatik.uni-hamburg.de](mailto:panse@informatik.uni-hamburg.de)

**Wolfram Wingerath**

University of Hamburg, Germany  
[wingerath@informatik.uni-hamburg.de](mailto:wingerath@informatik.uni-hamburg.de)

**Steffen Friedrich**

University of Hamburg, Germany  
[friedrich@informatik.uni-hamburg.de](mailto:friedrich@informatik.uni-hamburg.de)

**Norbert Ritter**

University of Hamburg, Germany  
[ritter@informatik.uni-hamburg.de](mailto:ritter@informatik.uni-hamburg.de)

**Abstract:** Currently, in many application areas the demand on probabilistic data grows. Duplicate entity representations are an essential problem of data quality, for certain databases as well as for probabilistic databases. Traditional duplicate detection approaches are based on pairwise comparisons. For dealing with large data sets, however, a comparison of all entity representation pairs is impractical and the search space is usually reduced by blocking techniques. The majority of blocking techniques is based on the usage of keys created from the original representations. These techniques, however, are only designed to deal with certain keys and hence cannot be used for probabilistic data without any adaptation. In this paper, we propose an adaptation of existing blocking techniques to data uncertainty based on the creation of certain keys from the probabilistic data. Moreover, we discuss some approaches for adapting the techniques' core functionalities to handle probabilistic keys. A final set of experiments evaluates the quality of our certain key based approaches in terms of pairs completeness and pairs quality.

**Key Words:** Probabilistic Data, Duplicate Detection, Blocking, Sorted Neighborhood Method

## 1. INTRODUCTION

Today, a large amount of real-life applications [1] [2] naturally produce uncertain, imprecise or vague information. For accurately storing such imperfect information probabilistic databases [3] [4] [5] have been developed. For meaningfully integrating probabilistic data originating from different sources or for cleaning a single probabilistic database, duplicate entity representations<sup>1</sup> need to be identified. Techniques for duplicate detection are usually based on pairwise comparisons of entity representations [6] [7]. However, for detecting duplicates in large data sets, a pairwise comparison of all representations is by far too expensive in storage as well as in time. Instead the search space has to be initially reduced to a manageable size by the usage of blocking techniques [8] [9] (also known as indexing [10]) as for example the Sorted Neighborhood Method [11]. The most of these blocking techniques are based on the usage of key values which are generated by the entity representations' data (in the following, we use the words 'key value' and 'key' synonymously). In probabilistic entity representations, however, the data used for key value creation can be uncertain. Thus, from applying a traditional key definition function probabilistic keys, i.e. keys with multiple possible instances, can result. Existing blocking variants are not designed to deal with probabilistic keys and hence cannot be used for probabilistic data without any adaptation.

---

<sup>1</sup> In certain relational data, an entity representation corresponds to an ordinary tuple, but in probabilistic data an entity is usually represented by more complex constructs as x-tuples [5] or tuple-blocks [3].

In this paper, we consider an adaptation of existing blocking techniques to the uncertainty and impreciseness modeled in probabilistic data in two ways. First, by resolving uncertainty during key value creation, i.e. by applying methods for creating certain keys from probabilistic entity representations, and second by adapting the core functionality of the blocking technique to probabilistic keys. The advantage of creating certain keys is that the core functionality of the blocking technique remains substantially unchanged and blocking can be applied as usual. In contrast, an adaptation to probabilistic keys implies a reimplemention of the whole blocking technique and hence adaptations to one blocking technique cannot be simply adopted to other ones. Due to a variety of probabilistic data applications restrict themselves to the usage of entity-independent data models, i.e. representation systems in which the uncertainties of different entity representations are not correlated, as BID-tables [3] or ULDBs [5] without lineage, we restrict ourselves to this class of probabilistic data models as well. A consideration of probabilistic data with entity dependencies is planned for future research.

The main contributions of this paper are:

- Strategies to adapt existing blocking techniques to entity-independent probabilistic data by creating certain keys from the uncertain data,
- Discussion on adapting the core functionality of the Sorted Neighborhood Method to probabilistic keys,
- An exhaustive experimental evaluation on the effectiveness and the accuracy of the proposed adaptations.

### 1.1. Motivating Example

As a motivating example, we consider the probabilistic entity representations of the three Movies 1-3 presented in Figure 1. Assume that the key of each movie is generated by concatenating the first three characters of its title and the last two digits of its production year. The title and the production year of Movie 1 are certain values and hence creating a certain key does not pose a problem. Although the title of Movie 2 is uncertain, for each of its possible instances the same key result, i.e. 'Bat01'. In contrast, the title's three first characters of Movie 3 are either 'Bat' or 'Ret' and hence are uncertain. A simple idea to solve this problem is to create a single certain key for each movie, but it is not clear which certain key represents Movie 3 at best. One intuitive solution is to take the key of the movie's most probable instance, which is 'Ret95'. Nevertheless, to take the key which is most probable at all (in this case 'Bat95') is maybe more appropriate. Another option is to represent Movie 3 by multiple certain keys, i.e. both 'Ret95' and 'Bat95'. We also could initially create a probabilistic key, but then the retained uncertainty has to be resolved during the remaining steps of the considered blocking technique. In summary, there are a lot of potentialities for handling this problem, but it is unclear which of them solves the problem at best.

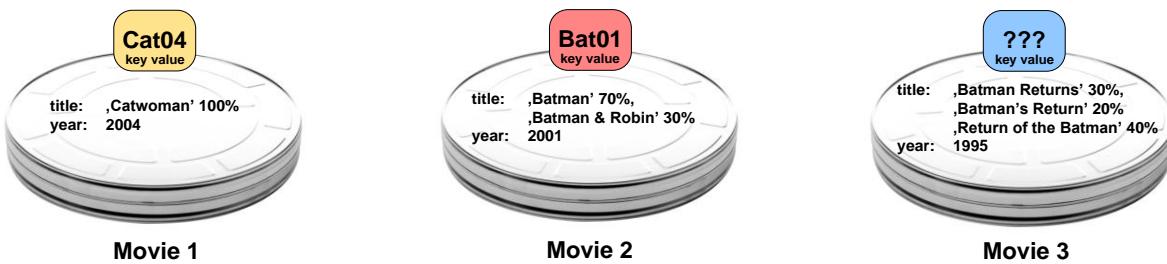


Figure 1: Probabilistic entity representations of three sample movies containing uncertain information

### 1.2. Outline

The paper is structured as follows: We start with some basics on probabilistic data, duplicate detection, and existing blocking techniques in Section 2. Then we present our strategies to adapt blocking to probabilistic data in Section 3. First we discuss some approaches based on certain keys (Section 3.1). In Section 3.2 we then propose some adaptations of the Sorted Neighborhood Method to probabilistic keys. We evaluate our newly defined strategies experimentally in Section 4. Finally, we examine related work in Section 5. Section 6 concludes the paper.

## 2. BASICS

In this section we give a short overview on probabilistic data and introduce some basics on duplicate detection and search space reduction (blocking). Moreover, we will go into detail with the Sorted Neighborhood Method which we will use as a blocking technique representative throughout this paper.

### 2.1. Probabilistic Data

A probabilistic relational database is defined on an ordinary relational database schema. According to the possible world semantics [12] the instantiation of a probabilistic database is theoretically defined as  $PDB = (\mathbf{W}, \mathbf{P})$  where  $\mathbf{W} = \{W_1, \dots, W_n\}$  is a finite set of possible instances of this database (also called as possible worlds) and  $\mathbf{P}: \mathbf{W} \rightarrow (0,1]$ ,  $\sum_{W \in \mathbf{W}} P(W) = 1$  is the probability distribution over these instances.

				<b>Possible World</b>	<b>Probability</b>
$t_1$				$W_1 = \{t_{2,1}, t_{3,1}\}$	$P(W_1) = 0.1 \times 0.35 = 0.035$
				$W_2 = \{t_{1,1}, t_{2,1}, t_{3,1}\}$	$P(W_2) = 0.56 \times 0.35 = 0.196$
				$W_3 = \{t_{1,2}, t_{2,1}, t_{3,1}\}$	$P(W_3) = 0.34 \times 0.35 = 0.119$
$t_2$				$W_4 = \{t_{2,1}, t_{3,2}\}$	$P(W_4) = 0.1 \times 0.25 = 0.025$
				$W_5 = \{t_{1,1}, t_{2,1}, t_{3,2}\}$	$P(W_5) = 0.56 \times 0.25 = 0.14$
				$W_6 = \{t_{1,2}, t_{2,1}, t_{3,2}\}$	$P(W_6) = 0.34 \times 0.25 = 0.085$
$t_3$				$W_7 = \{t_{2,1}, t_{3,3}\}$	$P(W_7) = 0.1 \times 0.2 = 0.02$
				$W_8 = \{t_{1,1}, t_{2,1}, t_{3,3}\}$	$P(W_8) = 0.56 \times 0.2 = 0.112$
				$W_9 = \{t_{1,2}, t_{2,1}, t_{3,3}\}$	$P(W_9) = 0.34 \times 0.2 = 0.068$
				$W_{10} = \{t_{2,1}, t_{3,4}\}$	$P(W_{10}) = 0.1 \times 0.2 = 0.02$
				$W_{11} = \{t_{1,1}, t_{2,1}, t_{3,4}\}$	$P(W_{11}) = 0.56 \times 0.2 = 0.112$
				$W_{12} = \{t_{1,2}, t_{2,1}, t_{3,4}\}$	$P(W_{12}) = 0.34 \times 0.2 = 0.068$

Figure 2: Sample x-relation (left) and its corresponding set of possible worlds (right)

Entity-independent probabilistic data models are specific probabilistic representation systems that restrict the possible world space to databases in which the instance and existence of one entity representation is independent from the instance and existence of any other entity representation. Although entity-independent probabilistic data models are no complete representation systems, i.e. there are sets of possible worlds which cannot be represented by such a data model; they are commonly used, because they are easier to manage than a complete one.

The simplest entity-independent probabilistic data model is a tuple-independent probabilistic data model [25] in which each entity is represented by a single tuple that is assigned with a probability score. In this representation system only the uncertainty on an entity's existence can be modeled in the probabilistic data. In this paper, we focus on entity-independent probabilistic data models that also allow a representation of the uncertainty on the entities' instantiations as ULDBs [5] and BID-tables [3] in which an entity is represented by an x-tuple or a block of disjoint tuples respectively.

Without any loss of generality, we use the ULDB model as a representative throughout this paper. The ULDB model [5] based on the x-tuple concept. Each x-tuple consists of a set of mutually exclusive alternatives each defined as a certain tuple which is assigned with a confidence score (attribute p). In the following, the set of alternatives (possible instances) of an x-tuple  $t$  is denoted as  $pI(t)$ . Moreover, the  $j^{th}$  alternative of an x-tuple  $t_i$  can be expressed by the form  $t_{i,j}$ . Maybe x-tuples (tuples for which non-existence is possible, i.e., for which the sum of its alternatives' probabilities is smaller than 1) are indicated by '?'. In the ULDB model different interpretations of the confidence values exist [5]. In our work we focus on probabilistic data, therefore, we always interpret confidence as probability. Relations containing one or more x-tuples are called x-relations. A sample movie x-relation with three x-tuples along with its possible world space is shown in Figure 2. Since x-tuple  $t_1$  is a maybe tuple with two alternatives and x-tuple  $t_3$  is a non-maybe tuple with four alternatives, the movie x-relation represents a set of twelve possible worlds.

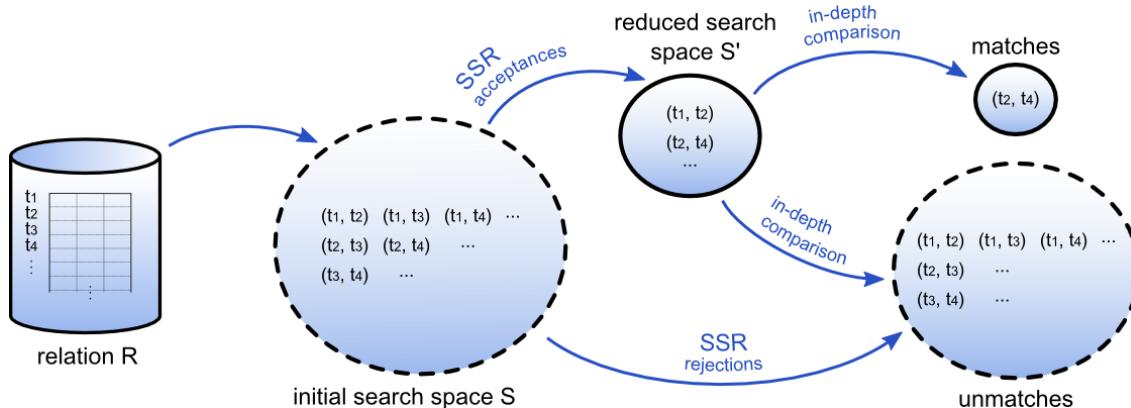
When clear from the context, we sometimes simply use 'tuple' to refer to x-tuples (and hence entity representations in general) and 'relation' to refer to x-relations.

## 2.2. Duplicate Detection

Duplicate detection means the problem of identifying multiple entity representations that refer to the same real-world entities. The most approaches for duplicate detection are based on pairwise entity representation comparisons [7] [6] [13] [14]. Such approaches can be conceptually decomposed into four phases [6]:

1. **Search Space Reduction:** Since a comparison of all pairs of tuples is mostly too inefficient, the search space is usually reduced using heuristic blocking techniques (see Section 2.2.1).
2. **Attribute Value Matching:** Similarity of tuples is usually based on the similarity of their corresponding attribute values. Despite data preparation, syntactic as well as semantic irregularities remain. Thus, attribute value similarity is quantified by syntactic and semantic means [6]. From comparing two tuples, we obtain a comparison vector  $\vec{c} = \langle c_1, \dots, c_n \rangle$ , where  $c_i$  represents the value similarity of the  $i$ th attribute.
3. **Decision Model:** The comparison vector is input to a decision model [13] which determines which set a tuple pair  $(t_1, t_2)$  is assigned to: matching tuples (M) or unmatched tuples (U).
4. **Duplicate Clustering:** Decision models only make decisions for single tuple pairs. To get a globally consistent result a clustering technique [6] needs to be applied.

For delimiting from the cheap comparison methods done by search space reduction techniques described later, we call the combined execution of the attribute value matching and the decision model as an in-depth comparison. We propose methods for in-depth comparisons of  $x$ -tuples in [15].



**Figure 3: The principal functionality of a search space reduction for duplicate detection. The dashed boundaries of the initial search space and of the set of unmatches indicate that these sets are never materialized.**

### 2.2.1. Search Space Reduction

Without reduction, the search space of a duplicate detection on an input relation  $R = \{t_1, t_2, \dots, t_n\}$  based on pairwise comparisons is principally the set of all possible pairs of tuples belonging to  $R$  (see Figure 3):

$$S = \{(t_i, t_j) \mid t_i, t_j \in R \wedge i < j\}$$

Since two tuples only need to be compared once and a tuple does not need to be compared with itself, the initial search space consists of  $\frac{|R| \times (|R|-1)}{2} = \frac{n^2-n}{2}$  tuple pairs (complexity  $O(n^2)$ ). In large data sets with millions or more tuples, the number of tuple pairs to be compared explodes and hence the duplicate detection process becomes infeasible. For that reason, the search space has to be initially reduced before comparing tuples in-depth. Reduction is realized by rejecting pairs of tuples being no duplicates for sure and adding them to the set of unmatches.

### 2.2.2. Evaluation Measures

Blocking is effective, if the number of rejected tuple pairs is high. Nevertheless, it is only accurate, if no true duplicate pair is rejected. In general, blocking is based on cheap comparisons and hence is known to cause two kinds of errors: false acceptance (short *FA*), i.e. leaving an actual unmatch in the search space, and – even worse – false rejection (short *FR*), i.e. removing an actual match from the search space by assigning it to the set of unmatches.

False rejection is worse than false acceptance, because an actual match that is rejected is not considered again and therefore changes the duplicate detection result for the worse, whereas a false acceptance is eventually corrected during the in-depth comparison.

To score accuracy and effectiveness we use the two measures pairs completeness (PC) and pairs quality (PQ) as proposed by Christen [10]. Pairs completeness represents the share of true acceptance (short *TA*) in all duplicate pairs (*TA*  $\cup$  *FR*), and pairs quality represents the share of true acceptances in the accepted tuples pairs (*TA*  $\cup$  *FA*):

$$PC = \frac{|TA|}{|TA| + |FR|}$$

$$PQ = \frac{|TA|}{|TA| + |FA|}$$

Note, compared to the pairs quality (also known as precision) achieved by in-depth comparison, a pairs quality of around 0.02 usually resulting from blocking is rather low, but compared to the pairs quality of the initial search space ( $\approx 2 \times 10^{-6}$ ), the percentage of increase is really high.

### 2.3. Existing Blocking Techniques for Certain Data

Currently several blocking techniques have been proposed (see [10] for a survey). The most of these techniques based on the usage of key values. The goal of this paper is not to present adaptations to probabilistic data for all of the key-based blocking techniques, but rather to point out different approaches for adaptation and to compare them with each other. In this paper, we consider three blocking techniques. We use the Sorted-Neighborhood Method (short SNM), which is a state-of-the-art blocking technique, to illustrate our adaptation strategies based on certain keys and discuss ways to adapt the SNM to probabilistic keys in Section 3. In our experiments in Section 4, we additionally use Standard Blocking [8] [14] (short SB) and Robust Suffix-Array Blocking [16] (short SAB).

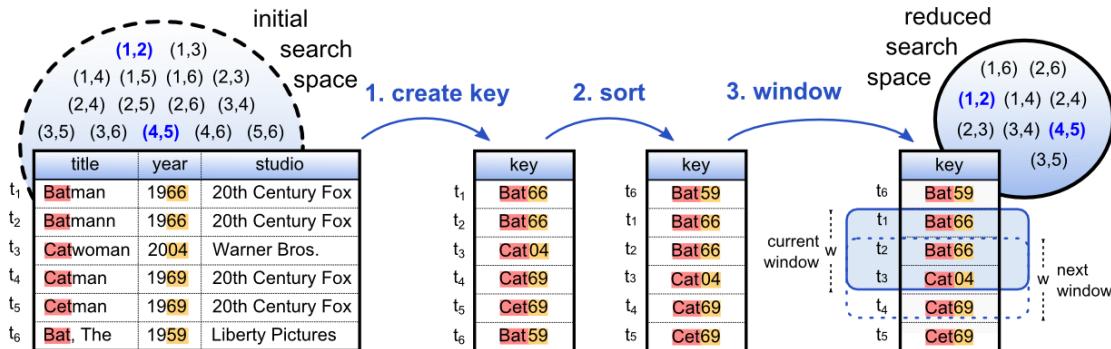


Figure 4: The three steps of the Sorted Neighborhood Method

#### 2.3.1. The Sorted Neighborhood Method

The standard variant of the SNM [11] reduces the search space in three steps (for illustration see Figure 4):

1. **Key Creation:** First, for each tuple  $t$  a key  $\kappa(t)$  is computed by concatenating characters of some identifying attributes as for example identification numbers, names, addresses, etc.. In our example, we concatenate the first three non-space characters of the title and the last two digits of the production year.
2. **Sorting:** Second, the tuples are sorted - usually lexicographically - by their respective keys.
3. **Windowing:** Finally, a window of fixed size  $w$  (in our example  $w=3$ ) slides sequentially over the sorted tuples. All tuples being within the window at the same time are paired with each other and added to the resultant search space. Due to the fixed window size, each tuple is compared with at most  $2w-2$  tuples from its immediate neighborhood.

The underlying assumption of the SNM is that duplicate tuples have similar keys and hence are sorted close together. According to [11], large window sizes do not lead to a high pairs completeness, but the rate of false acceptances grows very fast with the window size. For that reason, pairs completeness is often increased by using a multi-pass approach [11]. In this approach instead one, multiple key definition functions are used, each function in one pass. The final search space results in all candidate pairs detected for at least one pass (or more than  $k$  passes respectively). It is obvious that the resultant pairs quality is lower than in a single pass approach. However, the risk of not choosing the best key definition function is lowered and the result is usually more accurate. Assuming a data set with  $n$  tuples and a window of a fixed size  $w$ , the total number of tuple pair comparisons resulting from using the SNM with a single pass is  $O(wn)$  [11].

### 3. BLOCKING APPROACHES FOR PROBABILISTIC DATA

In the previous sections we introduced the ULDB model, described the process of duplicate detection in certain data and went into detail with the SNM. This section is devoted to the adaptation of blocking to the ULDB model. The one big issue here is that probabilistic entity representations may result in probabilistic keys. So in order to make blocking applicable to probabilistic data, the uncertainty of the keys has to be resolved. There are basically two approaches to those adaptations: generating only certain keys and thus resolving uncertainty during the key value creation, or generating probabilistic keys and so resolving the uncertainty during the remaining steps of the respective blocking technique (e.g. the sorting step or the windowing step of the SNM). For each of both approaches, we identified several strategies. With respect to the ULDB model, certain keys as well as probabilistic keys can be considered as non-maybe x-tuples defined on a single attribute. Whereas, a certain key has exactly one alternative, a probabilistic key can have multiple alternatives.

An important fact is that if the source data is certain each variant of our proposed adaptations (based on certain keys as well as probabilistic keys) lead to the same results as the original variants of the corresponding blocking techniques, i.e. our strategies are generalizations of the already existing techniques.

#### 3.1. Adaptations based on Certain Keys

An adequate strategy for building certain keys from x-tuples is by far not so straight forward as already illustrated in our motivating example in Section 1.1, because all the uncertainty in the tuple's data needs to be resolved. For certain key creation, we discuss four strategies. In the multi-pass over possible worlds (Section 3.1.1) a separate pass is applied to some of the database's possible worlds (each a certain relation). In key-per-tuple (Section 3.1.2) for each x-tuple a certain key is built by applying a traditional key definition function on a certain tuple representative. In key-per-alternative (Section 3.1.3) we create a key per x-tuple alternative (each a certain tuple). In key-per-representative (Section 3.1.4) we first compute a set of certain representatives for each x-tuple and then create a key for each of them. Some variants of these strategies can be also applied to immediately created probabilistic keys instead of the original x-tuples (concept *Uncertain Keys First*, see Section 3.1.5).

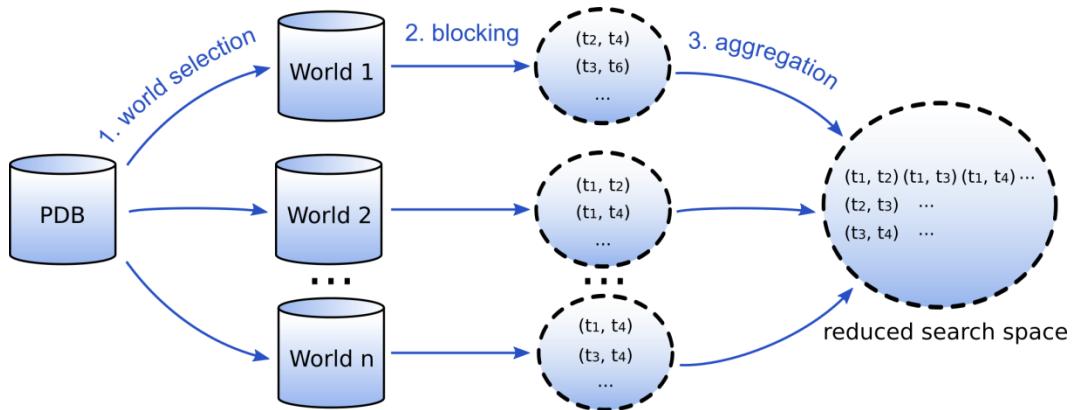
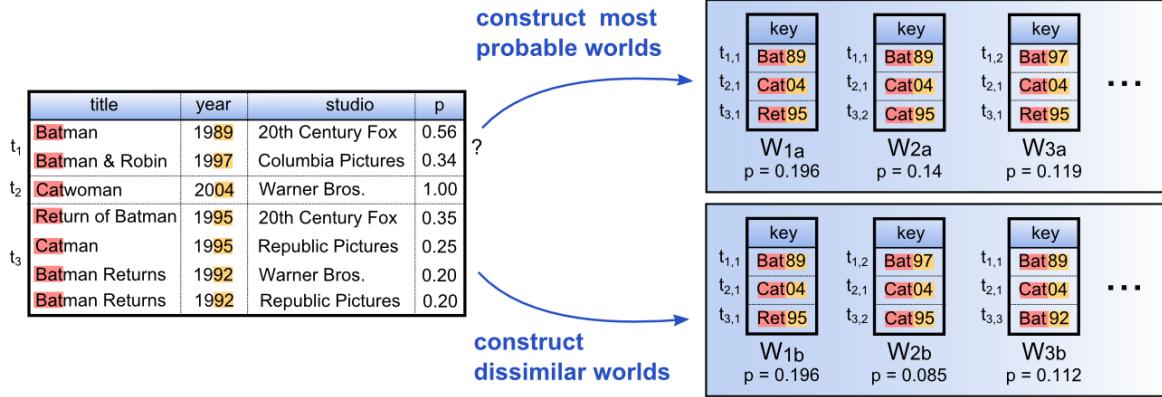


Figure 5: Basic concept of the multi-pass over possible worlds

##### 3.1.1. Multi-Pass over Possible Worlds

The idea for the first strategy is based on the fact that each possible world of an x-relation is an ordinary relation on which blocking can be applied as usual. Thus, a conceptually simple way to perform blocking with certain keys on

an x-relation is to construct its corresponding set of possible worlds (see Section 2.1), to apply the conventional blocking technique to each world individually, and to aggregate the resulting search spaces to a single one by the set union operator or by a voting strategy. The basic concept of the *multi-pass over possible worlds* strategy is illustrated in Figure 5.



**Figure 6: The variants of constructing the most probable worlds or constructing some dissimilar worlds respectively**

The problem is that the number of possible worlds of large x-relations is usually tremendous and running passes on all possible worlds is infeasible in practice. Moreover, some tuples are not present in some worlds and thus cannot be paired with other tuples for later in-depth comparison (for instance, in the sample of Figure 2 tuple  $t_1$  is missing in world  $W_1$ ). Therefore, instead to all blocking is only applied to a set of selected worlds.

The decision which possible worlds should be used is not easy to make; once the first run has been performed on the most probable world, additional passes over the next few most probable worlds will not improve the result very much, because the most probable worlds are usually very similar. For a better result, worlds should be considered that have not only a rather high probability, but are also as dissimilar from one another as possible. We implemented two variants of this strategy. One is to construct the  $k$ -most probable worlds and the other is to construct a set of  $k$  highly dissimilar possible worlds (see Figure 6). Both variants consider only worlds with all x-tuples present.

---

**Input:** x-relation  $R$

---

1. Let  $W_{MP} = \{argmax_{I \in pI(t)} p(I) | t \in R\}$
2. Compute for remaining alternatives  $t_{i,j}: w(t_{i,j}) = p(t_{i,j}) / max_{I \in pI(t)} p(I)$
3. Rank remaining alternatives  $t_{i,j}$  into list  $L_{alt}$  by  $w(t_{i,j})$
4. Let  $MostProbableWorlds = \{W_{MP}\}$
5. While  $|MostProbableWorlds| < k$ 
  - (a) Remove top element  $t_{i,j}$  from  $L_{alt}$
  - (b)  $NewWorlds = \emptyset$
  - (c) For each world  $W \in MostProbableWorlds$ :
    - $W_{New} = (W - pI(t_i)) \cup t_{i,j}$  with  $P(W_{New}) = P(W) \times w(t_{i,j})$
    - Add  $W_{New}$  to  $NewWorlds$
  - (d) Add all  $NewWorlds$  to  $MostProbableWorlds$
6. Rank  $MostProbableWorlds$  by probability into list  $L_{worlds}$

---

**Output:** First  $k$  elements of  $L_{worlds}$

---

**Algorithm 1: Compute the  $k$  most probable worlds**

Since all x-tuples are independent to each other, the  $k$ -most probable worlds can be built as described in Algorithm 1: First the most probable world is created by taking the most probable alternative from each x-tuple. Second the remaining alternatives are sorted into the list  $L_{alt}$  by a weight which is computed from the alternatives probabilities in descending order. Then as long as we have less than  $k$  worlds, we make copies from all already created worlds,

remove the top element  $t_{i,j}$  of  $L_{alt}$  and replace the current alternative of x-tuple  $t_i$  in each copied world by  $t_{i,j}$ . Finally we rank the set of created worlds<sup>2</sup> by their probabilities and take the  $k$  most probable ones.

<b>Input:</b> x-relation $R$	
1. Let $DissimilarWorlds = \emptyset$	
2. For $i = 1, \dots, k$ :	
(a) Let $CurrentWorld = \emptyset$	
(b) For each x-tuple $t \in R$ :	
If ( $i \leq  pI(t) $ )	
Add the $i$ th most probable alternative of $t$ to $CurrentWorld$	
Else	
Add the most probable alternative of $t$ to $CurrentWorld$	
(c) Add $CurrentWorld$ to $DissimilarWorlds$	
<b>Output:</b> World Set $DissimilarWorlds$	

### Algorithm 2: Compute $k$ dissimilar worlds

The basic idea of the second variant is to perform blocking on several possible worlds that are very dissimilar from each other. The probabilities of the constructed worlds are only of secondary importance. Here (see Algorithm 2), the most probable world with all tuples present is constructed for the first pass. Afterwards, a possible world is built by using only the second most probable alternative of each tuple. Accordingly, a possible world is then built from the third most probable tuple alternatives, and so on. This procedure is repeated, until all alternatives have been used or the user-defined threshold  $k$  is reached. If for any constructed world a tuple has no more new alternative, the most probable one is used for the remaining worlds. The number of worlds constructed by this procedure is rather small, as it cannot be greater than the maximum number of alternatives per tuple. Furthermore, each additional pass is likely to add many new tuple pairs and thus to improve the result much. So, this variant of the possible world strategy seems by far more promising than constructing the most probable worlds.

The biggest handicap of the multi-pass over possible worlds is its execution time. However, because all passes are independent to each other, we plan to reduce execution time by a parallel implementation using the Map-Reduce framework. The idea is to push each world to another mapper so that all passes can be done at the same time. Finally we use one reducer per x-tuple pair to decide if this pair belongs to the reduced search space or not. A similar approach has been already implemented by Kolb et al. [17] who perform a parallel multi-pass Sorted Neighborhood Method on certain data.

Theoretically, each variant of the multi-pass over possible worlds strategy is identical to a multi-pass over some variants of the key-per-tuple strategy (see Section 3.1.2), by using a different function for computing an x-tuple representative in each pass. However, finding a set of functions leading to the same results as the variants presented above is not trivial. For that reason, we consider this concept as an own strategy.

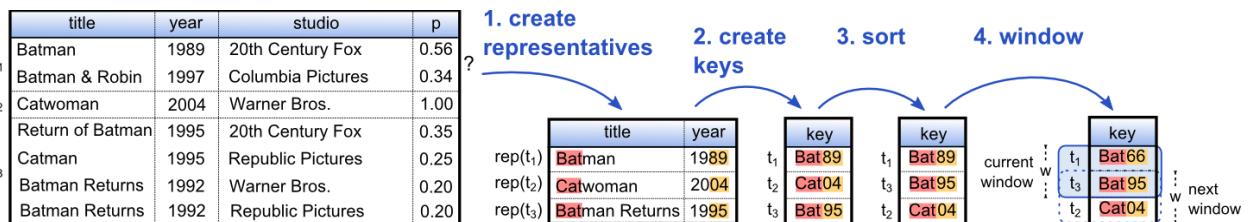


Figure 7: Key-per-tuple: In this example, a tuple's representative is computed from its most probable attribute values

<sup>2</sup> Note, by this algorithm a same world can be result from changing different worlds, but since we use a set of worlds we consider such duplicate worlds to be automatically removed. Moreover, the probability computation of the new worlds is only correct for the copy of the world with the most probable alternative of the considered x-tuple. Thus, we retain the highest probability when removing duplicate worlds.

### 3.1.2. Key-per-Tuple

This strategy resolves the uncertainty by computing exactly one certain key value for each x-tuple. As illustrated in Figure 7 and Algorithm 3, this strategy is composed of two steps. The simple idea is to compute a certain tuple for every x-tuple as a representative (Step 1) and then to create a key from this representative (Step 2).

For computing a certain x-tuple representative, metadata such as probabilities as well as the actual attribute values can be used. Of course, when computing a representative for key value creation, only key attributes have to be considered. Each x-tuple alternative corresponds to a certain tuple. Thus, computing a certain x-tuple representative from a set of alternatives is similar to computing a representative for multiple conflicting duplicate tuples in the fusion of certain data [18]. The only difference here is that x-tuple alternatives are per definition complete and no handling of null values is required. Moreover, x-tuple alternatives are assigned with probabilities and hence additional meta data for computing a representative is available. Following Bleiholder et al. [18], there are basically two strategies of computing a single representative of a whole tuple set: deciding strategies, in which simply one of the already existing tuples is chosen as a representative, or mediating strategies, in which from the given tuples a new representative is computed, i.e. the resultant representative does not necessarily belong to the input set.

<b>Input:</b> x-relation $R$ , key definition $\kappa$ , blocking technique $B$
1. Let $KeyTuplePairs = \emptyset$
2. For each x-tuple $t \in R$ :
(a) Create the tuple representative $rep(t)$
(b) Add $(\kappa(rep(t)), t)$ to $KeyTuplePairs$
3. Let $S$ be the search space that results from performing $B$ on $KeyTuplePairs$
<b>Output:</b> Search Space $S$

**Algorithm 3: key-per-tuple**

**Deciding Strategies:** A very simple deciding strategy is to pick the most probable alternative for each x-tuple. This is equivalent to perform blocking on just the most probable world without missing tuples (see Section 3.1.1). A more complex deciding strategy is based on the Distributional Cluster Feature (DCF). Andritsos et al. [19] use the DCFs to compute a tuple representative which in turn is used for computing a probability for each tuple of a duplicate cluster. Since by using this approach, the computed representative is not an element of the considered domain, the representative itself cannot be used for key value creation, but rather the x-tuple's alternative having the lowest distance to the x-tuple representative has to be used. Since this approach is most likely too time consuming for the blocking purpose and since our experiments showed that using a single key per x-tuple do not lead to best blocking qualities, we did not implement this variant so far.

<b>function</b>	<b>type</b>	<b>description</b>
cry with the wolves	dec.	take the most often occurring value
most probable value	dec.	take the most probable value
roll the dice	dec.	pick a value randomly
longest value	dec.	take the longest value
median/average	med.	compute the median/average of all values
expectation value	med.	compute the expected value

**Table 1: Conflict resolution functions which can be used for mediating strategies**

**Mediating Strategies:** In many situations an alternative computed with a mediating strategy represents an x-tuple better than one of the already existing ones. Mediating strategies are usually applied on an attribute-by-attribute basis. In other words, the tuple representative results from computing a single value representative for each of its attributes. Functions for merging single attributes are denoted as conflict resolution functions [18], because a representative is computed from multiple conflicting input values. To each attribute a different resolution function can be applied. Like the whole strategies, resolution functions can be of a deciding or a mediating style. By using a deciding function one of the existing values is chosen. Two typical deciding functions are *cry with the wolves* where the most often occurring value is taken or *roll the dice* where one of the given values is picked randomly. By mediating functions from a set of given values a new value is created. A typical mediating function is *meet in the middle*, by which the average value or the median is computed. Since x-tuple alternatives are assigned with probabilities, additional conflict resolution functions are possible and often more convenient, e.g., a deciding

function in which the most probable value is chosen or a mediating function in which the expected value is computed. A set of conflict resolution functions which can be used for computing a representative of an uncertain attribute value is listed in Table 1.

Naturally, different techniques may be used for different attributes, e.g. the median or the expectation value can be used for numbers, while string values can be processed with taking the most probable value. Moreover, we generally use the roll the dice function as a fallback strategy, when the primary used function delivers an ambiguous result.

To illustrate the difference between deciding strategies and mediating strategies only consisting of deciding functions, we consider the x-tuple  $t_3$  with its four alternatives presented in Figure 6. By choosing the most probable alternative,  $t_3$  is represented by  $rep(t_3) = t_{3,1}$ . In contrast, by choosing the most probable value for each attribute (mediating strategy with deciding functions) the representative  $rep(t_3) = ('Batman Returns', 1995)$  results, which is not equal to any alternative of the considered x-tuple.

### 3.1.3. Key-per-Alternative

In our third strategy, we do not create a single key for each tuple, but for tuple alternatives, so that tuples may have more than one key computed for them. As a consequence, a tuple can appear several times in the sorted list as shown in Figure 8. Since tuples may appear several times in one window, the number of different x-tuples per window can vary. In order to prevent this effect, we redefine the window size as the number of different x-tuples per window instead of the number of (key,tuple) pairs per window.

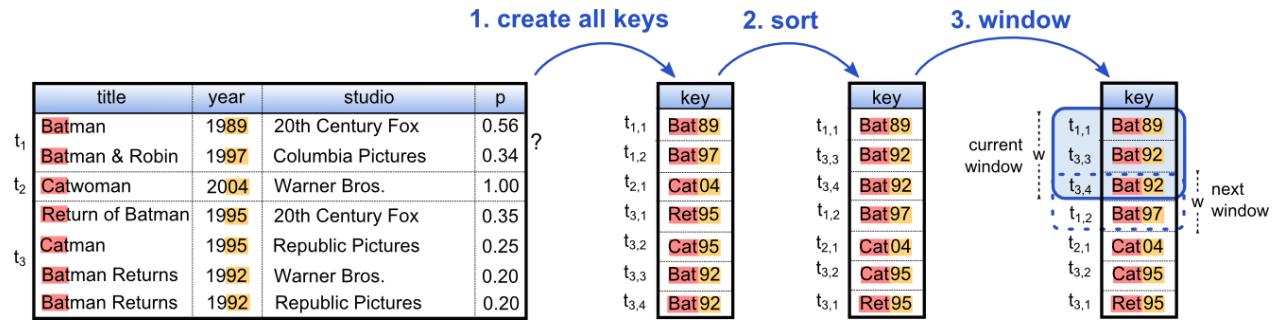


Figure 8: The standard variant of the key-per-alternative strategy

There are many approaches to decide which alternatives are used for key value creation. One of them is to simply use all alternatives. Another idea is to use only a predefined number of alternatives per tuple or to use the most probable alternative of every tuple and, in addition, a share of the remaining alternatives, e.g. the 100,000 most probable remaining alternatives in the database.

In this paper, we consider two variants: (a) the standard variant (KpA-All) which creates a key for all alternatives, and (b) the Top- $k$ -variant (see Algorithm 4), which creates a key for the  $k$  most probable alternatives of each x-tuple.

---

**Input:** x-relation  $R$ , key definition  $\kappa$ , blocking technique  $B$

1. Let  $KeyTuplePairs = \emptyset$
2. For each x-tuple  $t \in R$ :
  - (a) For each  $i \in 1, \dots, \min(k, |pI(t)|)$ :
    - i. Let  $I$  be the  $i$  most probable alternative of  $t$
    - ii. Add  $(\kappa(I), t)$  to  $KeyTuplePairs$
3. Let  $S$  be the search space that results from performing  $B$  on  $KeyTuplePairs$

---

**Output:** Search Space  $S$

---

Algorithm 4: Top- $k$ -variant of key-per-alternative

### 3.1.4. Key-per-Representative

Our fourth and newest strategy is basically a generalization of key-per-tuple and key-per-alternative and hence is a mixture of both concepts. The underlying idea is to create multiple key values per tuple (as in key-per-alternative)

each derived from a generated tuple representative (as in key-per-tuple). Thus, in key-per-representative we combine the concepts of key-per-tuple and key-per-alternative.

Key-per-representative can be specialized to key-per-tuple by generating only a single x-tuple representative, and it can be specialized to key-per-alternative by generating x-tuple representatives only with deciding strategies.

### 3.1.5. Concept of *Uncertain Keys First*

In the concept of *Uncertain Keys First*, instead of working on the original set of x-tuples, the proposed methods for certain key value creation are applied on intermediately generated probabilistic keys, each being an x-tuple with one attribute. Since multiple, maybe each less probable, alternatives of an x-tuple can have the same keys, the most probable alternative of an x-tuple's probabilistic key can differ from the key of the most probable alternative of this x-tuple. As a result, the keys created by using the *Uncertain Keys First* concept can be more representative for the considered x-tuples than the keys resulting from applying the key creation strategy commonly.

For illustrating the *Uncertain Keys First* concept and for demonstrating the differences to the standard approach, we consider the tuple  $t_3$  from Figure 7. Assume that we apply the Top-2-variant of the key-per-alternative strategy. Instead of creating certain keys for the two most probable alternatives of each x-tuple, we choose the two most probable alternatives of each x-tuple's probabilistic key. For that purpose, in a first step, for each x-tuple a probabilistic key is created. Since the alternatives  $t_{3,3}$  and  $t_{3,4}$  of tuple  $t_3$  have the same key 'Bat92', the probability of the corresponding alternative of the probabilistic key is equal to the sum  $p(t_{3,3}) + p(t_{3,4}) = 0.4$ . In the second step, the intended Top-2-selection is applied to the probabilistic keys. Thus, in our example, the third x-tuple is represented by the keys 'Bat92' and 'Ret95' instead by the keys 'Ret95' and 'Cat95'.

Before evaluating the quality of this concept by our experiments in Section 4, we first discuss the feasibility of the *Uncertain Keys First* concept for the different variants of our certain key based strategies:

- **Key-per-Tuple:** In deciding strategies the key of one (e.g. the most probable) alternative is taken. Choosing the most representative alternative of an x-tuple's probabilistic key seems more qualified than choosing the key of the most representative x-tuple's alternative. In contrast, mediating whole instances (x-tuple alternatives) seems more qualified than mediating single attribute values (probabilistic key alternatives), because keys are composed by proportion of different attributes and hence have no inherent semantics. For that reason, we suggest to take the *Uncertain Keys First* concept for variants only based on deciding strategies and not to use this concept for mediating variants or mixed ones.
- **Multi-Pass of Possible Worlds:** The *Uncertain Keys First* concept should improve the accuracy of the Top- $k$  variant, because more representative worlds are selected. In contrast, in the variant of dissimilar worlds, worlds are arbitrarily selected. Thus, we cannot make any appropriate forecast for that variant.
- **Key-per-Alternative:** For the Top- $k$ -variant the *Uncertain Keys First* concept should improve accuracy. If keys for all alternatives are created; the results of both concepts are equivalent.

## 3.2. The Sorted Neighborhood Method with Probabilistic Keys

In this section, we shortly discuss in which ways the core functionality of the SNM can be adapted to probabilistic keys. Sorting tuples by their key values corresponds to a tuple rank scenario where the keys serve as ranking scores and the lexicographic order serves as ranking order. Thus, we consider existent techniques for ranking probabilistic tuples [20] to resolve the uncertainty in the sorting step by building a sorted list of x-tuples based on their probabilistic keys or to resolve the uncertainty in the windowing step by sliding the window over a set of possible sorting lists.

### 3.2.1. Single Ranking Approaches

In single ranking approaches from probabilistic keys a single certain ranking is computed in the sorting phase.

- **Most Probable Ranking (SNMMPR) :** The base idea of this adaptation is to rank (sort) the probabilistic tuples by the most probable ranking of their key values. By using a key definition function  $\kappa$ , this can be realized by sorting based on the two relations ' $<_P$ ' and ' $=_P$ ', which are defined as:

$$t_1 <_P t_2 \Leftrightarrow p(\kappa(t_1) < \kappa(t_2)) > p(\kappa(t_2) < \kappa(t_1)) \text{ and } t_1 =_P t_2 \Leftrightarrow \neg(t_1 <_P t_2) \wedge \neg(t_2 <_P t_1)$$

Since just another order relation is used, complexity is dominated by the sorting time ( $\Rightarrow O(n \log(n))$ ).

- **Expected Position Ranking (SNM<sub>ExpR</sub>)** : This approach based on the idea to compute the expected rank position per x-tuple and then to rank all tuples by this position. For a finite set of possible ranking scores per tuple this computation can be done in  $O(n \log(n))$  [20].
- **Expected Score Ranking (SNM<sub>ExpS</sub>)** : This approach based on the idea to transform the blocking key into a numerical value, to use this value as a ranking score and then to rank the tuples by their expected score. For simple transformations this approach is dominated by the ranking time ( $\Rightarrow O(n \log(n))$ ).
- **Uncertain Rank Aggregation (SNM<sub>URA</sub>)** : In this approach, a ranking is computed which has the minimal average (expected) distance to all possible rankings. For attribute uncertainty models such an aggregation based on the footrule distance can be done in polynomial time ( $O(n^{2.5})$ ), whereas a computation based on the Kendall tau distance is known to be NP-Hard [20].

### 3.2.2. Multiple Ranking Approaches

As multiple ranking approaches, we consider approaches which do not produce a single ranking result, but resolve uncertainty in the windowing phase.

- **Sorted U-Rank Neighborhood (SNM<sub>URN</sub>)** : This approach is based on the rank function  $l\text{-UTop-Rank}(i,j)$  which is defined by Ilyas et. al [20]. This rank function returns the  $l$  most probable x-tuples that appear at the rank position  $i \dots j$ . Let  $w$  be the used window size, in the Sorted U-Rank Neigborhood we pair all x-tuples that result from  $l_1\text{-UTop-Rank}(i,i)$  with all x-tuples that result from  $l_2\text{-UTop-Rank}(i-w,i+w)$ , where  $l_2 > l_1$  (for example  $l_2 = w \times l_1$ ).

### 3.2.3. Comparison

Since the most probable sorting should be more representable than the sorting resulting from the most probable world, the SNM<sub>MPR</sub> is expected to supply a better blocking quality than the Top-1 variant of key-per-alternative. However, by using a single ranking approach each x-tuple is represented only once in the sorted list. Thus, an x-tuple  $t_i$  is only close to a second x-tuple  $t_j$ , if  $t_j$  is similar to the other neighbors of  $t_i$ , too. Therefore, similar to key-per-tuple, x-tuple uncertainty can be only restrictedly considered, because there exist no single sort position for an x-tuple with dissimilar alternatives which is appropriate to find all of its duplicate candidates. As a consequence, from single ranking approaches we can expect a blocking quality which is similar to the quality of key-per-tuple. First experiments for SNM<sub>MPR</sub> and SNM<sub>ExpS</sub> confirmed that intuition.

## 4. EXPERIMENTAL EVALUATIONS

In our experimental evaluations, we analyzed the differences in blocking quality of our proposed adaptations based on creating certain keys. Hereby, we especially focused on the robustness against a varying data dirtiness and a varying data uncertainty. Moreover, we evaluated for which variants the *Uncertain Keys First* concept was actually valuable. Finally, we compared the quality results of our adaptations for different blocking techniques.

### 4.1. Probabilistic Test Data

Getting large sets of unclean probabilistic real-life data being labeled, i.e. each duplicate pair is exactly known, is nearly impossible. For that purpose, we produced some synthetic data sets for revalidating the quality of our proposed strategies. In order to make the data as realistic as possible, we decided to use real-life data from an existing certain database. So we extracted title, production year, studio and director of about 300,000 movies from the online movie database IMDb<sup>3</sup> with the Java application JMDB<sup>4</sup> and stored the data to an HSQLDB<sup>5</sup>.

For generating probabilistic data from the duplicate-free certain data, we programmed a Java application named ProbDataGen<sup>6</sup>. With ProbDataGen it is possible to choose among several HSQL databases holding certain movie data to generate a probabilistic movie database with duplicates, where the user can make several adjustments, e.g. the number of duplicates, the maximal number of alternatives per tuple, or the datas' degree of dirtiness.

To improve the reliability of our experimental results further on, we use a standard data setting for the movie tables in our experiments. The characteristic of this standard setting is adopted from the characteristic of a real-life CD-

<sup>3</sup> The Internet Movie Database (<http://www.imdb.com>)

<sup>4</sup> Java Movie Database (<http://www.jmdb.de>)

<sup>5</sup> HyperSQL DataBase (<http://hsqldb.org>)

<sup>6</sup> <http://vsis-www.informatik.uni-hamburg.de/projects/QloUD/ProbDataGen>

dataset<sup>7</sup> with duplicates. We adjust the percentage of duplicates, the average duplicate cluster size and the average similarity of the true duplicates to this real-life data set. In experiments where data characteristics are modified for experimental reasons, we used this setting as a fixed point and only changed the analyzed characteristic. We think that these adjustments make our experiments as realistic as possible, even though synthetic data sets are used.

All the data sets (along with descriptions of their characteristics) we used in our experiments are available at <http://vsis-www.informatik.uni-hamburg.de/projects/OloUD/ICIQ2012/TestData>.

## 4.2. Experimental Settings

We performed five experiments. For space limitations, for the first four experiments we show only the results for the SNM which in our mind were most illustrative. In the last experiment, we also used Standard Blocking (SB) and Robust Suffix-Array Blocking (SAB) to make an overall comparison between different blocking techniques.

1. In the first experiment, we made an overall comparison of the certain key based variants proposed in this paper. We evaluated and compared their quality in terms of pairs completeness, pairs quality and runtime. In this experiment, we used the SNM with a fixed window size  $w=10$  and a key built by the first 12 non-space characters of the movie title (parameter  $k_l$ ) and the last two digits of the production year. Moreover, we used movie tables generated with our standard data setting.
2. Duplicate detection is especially required to work on dirty data, i.e. source data with poor quality. Thus, in Experiment 2, we evaluated the robustness of our variants against a varying dirtiness of the source data. For that purpose we used six sets of movie tables each generated with different settings for dirtiness. Since we consider duplicate detection, we measure quality as the average similarity of the true duplicate pairs (the lower the average duplicate similarity, the dirtier the data). For measuring similarity, we took the Monge-Elkan distance [6], which is known to work well for most domains. In this experiment, we used  $k_l=12$ . Moreover we used the SNM with a specifically chosen  $w$  for each strategy so that all strategies produced a search space of similar size (this should enable a fair comparison of pairs completeness).
3. In the third experiment, we evaluated the robustness against a varying data uncertainty. For that purpose, we changed the average number of alternatives per x-tuple. We used the SNM with  $w=10$  and  $k_l=12$ .
4. In the fourth experiment, we evaluated the impact of the *Uncertain Keys First* concept on the resultant blocking quality. In this experiment, we used the SNM with  $w=10$  and  $k_l=12$ .
5. In our final experiment, we compared the results from the SNM with the results from Standard Blocking (SB) and Robust Suffix-Array Blocking (SAB). For comparison, we conducted runs with two different experimental objectives. First, we executed the KpA-All variant on several databases with varying quality to test the robustness of the blocking techniques against poor data quality (Objective 1). Then, we compared the results for a selected set of adaptation approaches w.r.t. these three techniques on our standard data set (Objective 2). For the first objective, we took the KpA-All variant, because it was the adaptation approach performing best for all three techniques. For the second objective, we took our standard data set and performed for each blocking technique KpT, KpA-All, Diss(10) and Top-1.

For our experiments we consider the adaptation variants listed in Table 2.

<b>shorthand</b>	<b>variant description</b>
Top-1	a single pass over the most probable world (identical with MPW-1 and KpA-Top-1)
MPW-10	a multi-pass over the 10 most probable worlds
Diss( $k$ )	a multi-pass over $k$ dissimilar worlds
KpT	a key-per-tuple variant which build a representative by using the most probable value of each attribute
KpA-All	the standard variant of key-per-alternative using all alternatives for key value creation
KpA-Top- $k$	a key-per-alternative variant which uses the $k$ most probable alternatives for key value creation
KpR	a key-per-representative variant which takes all alternatives plus a tuple built by the most probable attribute values as representatives

**Table 2: The variants (along with their shorthand symbols) of our certain key based approaches used in the experiments**

<sup>7</sup> [http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/datasets/cd\\_datasets.html](http://www.hpi.uni-potsdam.de/naumann/projekte/repeatability/datasets/cd_datasets.html)

We applied each experiment on generated data sets of 102,692 x-tuples with 4,380 duplicate pairs. If not stated otherwise, each x-tuple has at most 10 alternatives (5.46 alternatives in average). All experiments were performed on a machine with an Intel(R) 3.1GHz quad-core processor, 8GB main memory, and a 64-bit operating system.

### 4.3. Experimental Results

#### 4.3.1. Experiment 1: Overall Comparison of Adaptation Strategies using the SNM

The absolute values of pairs completeness and pairs quality are shown in Figure 9. Table 3 shows the blocking quality of different variants in relation to the blocking quality produced by KpA-All. Figure 10 shows the runtime of the different variants.

As expected and shown by the experimental results, a multi-pass over the  $k$  most probable worlds with  $k > 1$  did not bring any advantage, because no new candidate pairs result from the subsequent passes, but runtime increased linear with growing  $k$ . In contrast, a multi-pass over dissimilar worlds was extremely beneficial. Already for small window sizes and short keys a good pairs completeness ( $PC > 0.9$ ) was achieved. The goodness lacked with fewer worlds to be constructed, but was still of good quality by using 5 dissimilar worlds (see Diss(5) in Figure 9 and in Table 3).

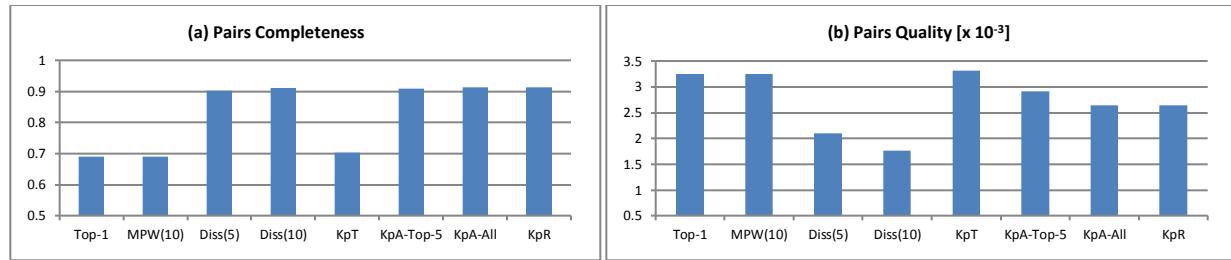


Figure 9: Pairs completeness and pairs quality of different adaptation variants, each performed with the SNM

Interestingly, the used KpT variant which creates an x-tuple representative by using the most probable value of each attribute performs a little bit better than using the most probable alternative as the representative (Top-1). That shows that mediating strategies can be useful to create an x-tuple representative. Combining mediating strategies and deciding strategies for creating a set of x-tuple representatives, as we did it with the KpR variant, was not successful, i.e. it did not improve the KpA-All variant in any of the performed experimental runs.

The conclusion of this experiment is that for the SNM producing multiple keys per tuple turned out to be more accurate than creating a single one. Of course, the resultant search space grows with the number of alternatives used for key value creation, but the resultant values of pairs quality are all of an acceptable size. The trade-off between accuracy and effectiveness is perfectly illustrated by the results shown in Table 3. The strategies using a single key per x-tuple (KpT, Top-1) are most effective (smallest search space and lowest runtime), but less accurate than the strategies using multiple keys per x-tuple (KpA-All, KpA-Top-5, Diss(5), Diss(10)).

In summary, due to the higher priority of pairs completeness, the variants which produce multiple key per tuple (KpA, Diss( $k$ )) turned out to be best suitable to adapt the SNM to probabilistic data.

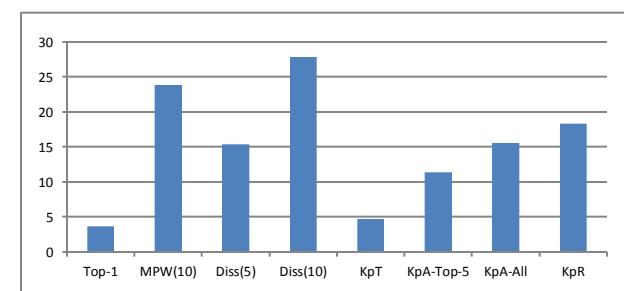


Figure 10: Runtime [in sec] of different adaptation strategies based on certain key values performed with the SNM

Table 3: Comparison of different variants to KpA-All (best results are underlined)

### 4.3.2. Experiment 2: Robustness against a varying Dirtiness of the Source Data

Since we set all strategies so that they produced search spaces of similar sizes, we present only results on pairs completeness in Figure 11. As you can see, all the variants produced a result of good quality if the source data were of good quality (similarity of 0.93), too. Nevertheless, the blocking quality shrank rapidly when the source data became dirtier. In general, it is easy to see that the five considered variants can be grouped into two classes. The first class contains KpA-All, KpA-Top-5 and Diss(10). These variants worked acceptable for the three cleanest data sets and became only bad for the data sets with the poorest quality. The second class contains KpT and Top-1. The blocking quality of these variants was bad in the most cases. This experiment shows that using multiple keys for x-tuples makes the blocking process more robust against a varying dirtiness of the source data.

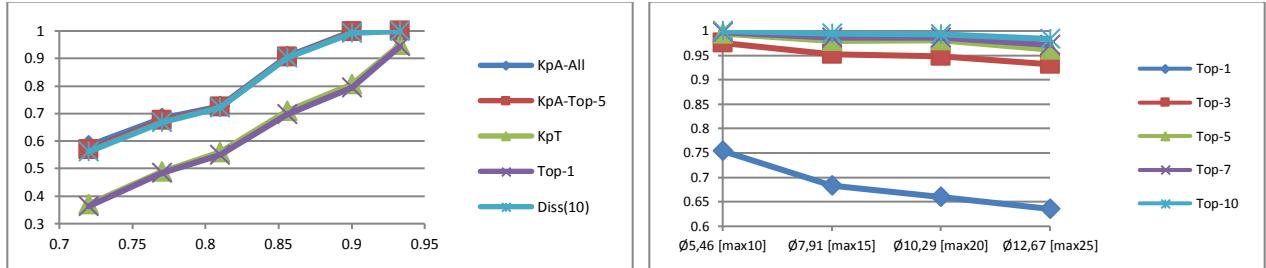


Figure 11: Pairs completeness for different variants of the adapted SNM w.r.t. a varying quality of the source data

Figure 12: Pairs completeness for KpA Top- $k$  w.r.t. a growing number of x-tuple alternatives

### 4.3.3. Experiment 3: Robustness against a varying Uncertainty of the Source Data

In the results of the previous experiments, the KpA-All variant shows the best performance on pairs completeness. However, creating a key for each alternative can be very ineffective for databases with a high degree of uncertainty, i.e. the average number of alternatives per x-tuple is very high. For that reason, we were interested in the loss of quality we will suffer, if we use only the  $k$  most probable x-tuple alternatives instead all of them. To evaluate that fact, we conducted a set of experiments with different settings for  $k$  on four different sets of movie tables, each with another degree of uncertainty. The experimental results on pairs completeness are depicted in Figure 12. The notation  $\emptyset a$  [max  $b$ ] on the x-axis denotes that in the corresponding movie table the average number of alternatives per x-tuple was  $a$  and the maximal number of alternatives an x-tuple can have was  $b$ . The values of the individual variants are computed in relation to the result of the variant KpA-All, i.e. a result of 1.0 for a setting  $k$  means that the Top- $k$  variant detected all the duplicates which have been detected by the KpA-All variant.

The Top-1 variant performed significantly worse than the KpA-All variant, but for  $k > 2$  the loss of true positives compared to KpA-All is less than 5%, even if the maximal number of alternatives per x-tuple is up to 25. Certainly, the relative number of correctly detected duplicate pairs shrank, if data uncertainty grew, but this loss of quality is of an acceptable size. To show the complexity which comes along with a high setting of  $k$ , we also compared the absolute size of the resultant search space and the execution time (see Figure 13). The higher  $k$ , the more the search space grew proportional with the uncertainty of the data. In contrast, for low values of  $k$ , e.g.  $k = 1$  or  $k = 3$ , the size of the search space was mostly independent from the degree of uncertainty. Moreover, the runtime of KpA-All grew extremely with a growing number of x-tuple alternatives, whereas the runtime for the other variants grew less significantly, the lower  $k$ .

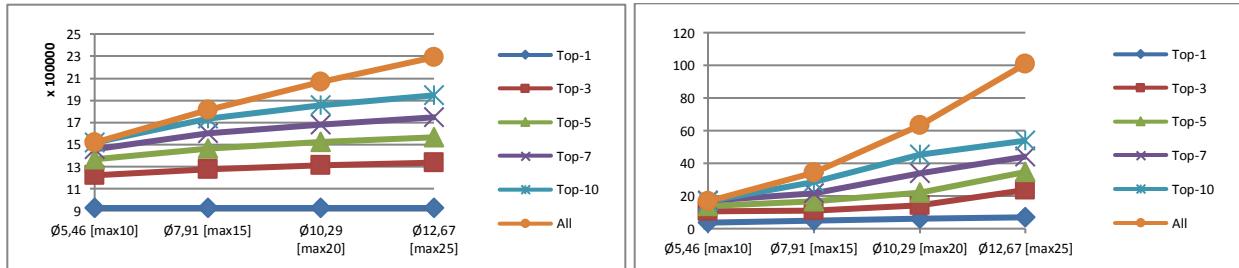


Figure 13: Absolute search space sizes and runtimes [sec] for different variants of KpA Top- $k$  w.r.t. a growing number of x-tuple alternatives

#### 4.3.4. Experiment 4: Uncertain Keys First

To test the idea of *Uncertain Keys First*, we conducted a set of experiments and tested different variants of the key-per-alternative strategy and the multi-pass over possible worlds strategy. Recall, these are the two strategies for which we expected that *Uncertain Keys First* could have a positive impact (see Section 3.1.5).

As expected, *Uncertain Keys First* improved the pairs completeness of the KpA-Top- $k$  variants as well as the pairs completeness of the multi-pass over the  $k$  most probable worlds, but surprisingly decreases pairs completeness of the multi-pass over  $k$  dissimilar worlds. We detect that this impact is substantially independent from the window size and the quality of the data. The most interesting effect of *Uncertain Keys First* was observed for the KpA-Top- $k$  variant. The degree of improvement decreased with growing  $k$ , i.e. is maximal for  $k = 1$ , and increased with the number of alternatives per x-tuple. The average amount of improvement (scored in percentage of pairs completeness) w.r.t. different settings of  $k$  as well as the average amount of improvement w.r.t. a growing number of alternatives per x-tuple are shown in Figure 14. In both cases, we aggregated over the remaining dimension.

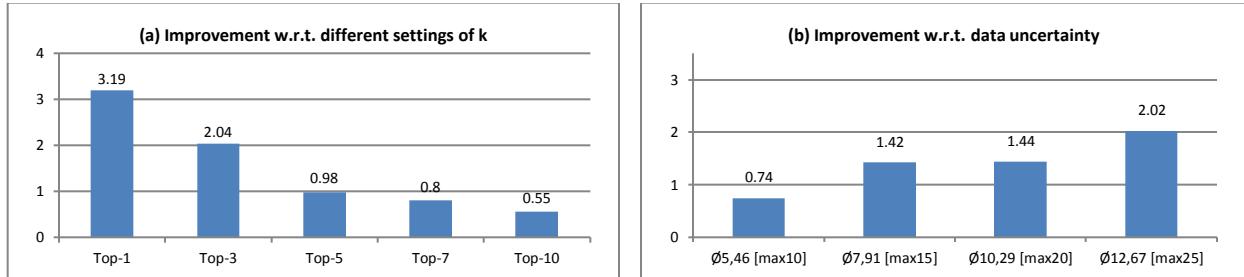


Figure 14: The improvement achieved by using *Uncertain Keys First* with respect to (a) different settings of the KpA Top- $k$  variant and (b) a growing number of alternatives per x-tuple

#### 4.3.5. Experiment 5: Overall Comparison of Different Blocking Techniques

The results of the robustness test are shown in Figure 15. In databases of good quality (similarity > 0.9) all three techniques achieved an outstanding pairs completeness close to 1. In contrast pairs completeness shrank significantly for databases with poor quality. SAB was by far the most robustness technique. Even for an average duplicate similarity of 0.72 SAB achieved a pairs completeness of nearly 0.9. In contrast the pairs completeness of SB and SNM decreased down to 0.74 (SB) or 0.61 (SNM) respectively. Surprisingly, SB performs better than SNM. Moreover, SAB achieved by far the highest pairs quality and produced the smallest search space. The pairs quality of SB and SNM were nearly identical. In general, pairs quality shrank, if the duplicate pairs became more dissimilar. The results of our second objective are depicted in Figure 16. They show that SAB performed best for all of the adaptation variants. Second in quality was SB. SNM achieved the poorest results. You can see that the differences in blocking quality of the certain key variants are the same for all three techniques: KpT performed better than Top-1 what shows that using the most probable alternative is generally not the best variant to create a tuple representative. Moreover, the resultant qualities of the different blocking techniques vary at most in the variants producing a single key. In contrast, for KpA-All and Diss(10) all three techniques produced similar results. The single key strategies as KpT produce a smaller search space and hence had a better pairs quality than the strategies producing multiple keys.

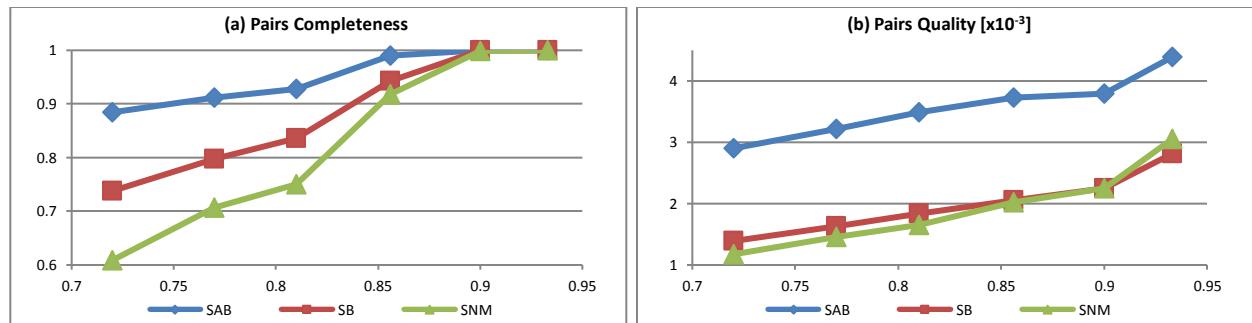


Figure 15: Pairs completeness and pairs quality of KpA-All performed with SAB, SB and the SNM w.r.t. databases of different qualities (measured by the average similarity of all true duplicates)

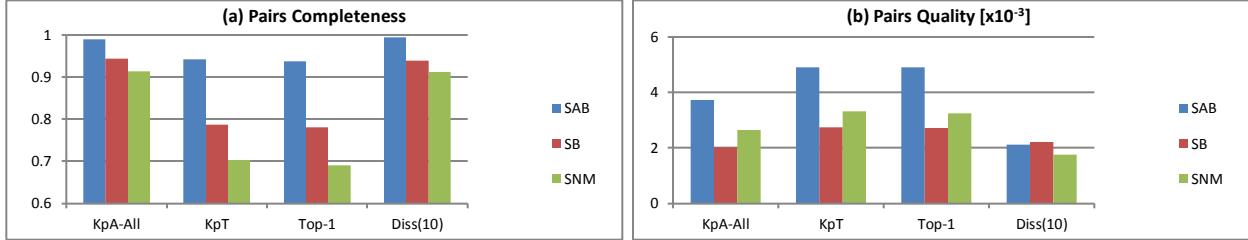


Figure 16: Pairs completeness and pairs quality of some adaptation strategies performed with SAB, SB and the SNM

#### 4.4. Experimental Conclusions

The experiments presented above show the feasibility of our approaches. Moreover, with key-per-alternative and the multi-pass over dissimilar worlds, they lift out two adaptation strategies which were best fitting for all three considered blocking techniques. Moreover, they were most robust against a poor quality of the source data. Only in scenarios where the search space must be as small as possible, a single key approach as key-per-tuple is maybe a better choice. The critical point of KpA is the one discussed in Experiment 4. Using all the alternatives for key value creation can affect the efficiency of this approach negatively. For that reason a Top- $k$  variant with  $k > 2$  is sometimes better suitable. In that case the concept of *Uncertain Key First* can improve the effectiveness further on, but slightly increases the search space. The drawback of Diss( $k$ ) is its long runtime for high settings of  $k$ . However, this weak point should be erased by a parallel implementation as we plan it in future research.

### 5. RELATED WORK

Duplicate detection in general [6] [7] [14] [21] [13] and blocking in particular [10] are handled in several works. Existing blocking techniques that are based on the use of key values are Standard Blocking [8] [22], the Sorted Neighborhood Method [11] [23] [24], Q-gram Indexing [9], Suffix-Array Blocking [24] [16], K-way Sorting [25], Similarity-Aware Inverted Indexing [26], Sorted Blocks [27], String Map based Indexing [28], Priority Queue [29], TI-similarity [30], and Adaptive Filtering [31]. Further blocking techniques are Locality-Sensitive Hashing [32] and Fuzzy Blocking [33], Canopy Clustering [34] [35], Spectral Neighborhood Blocking [36], and blocking with MFIBlocks [37]. Kolb et al. [17] consider a parallelization of duplicate blocking using the Map-Reduce programming model. Approaches for blocking based on semantic relationships between data items are proposed in [38] (tuple relationships given by foreign keys) and [39] (hierarchical relationships in XML documents). In [40] blocking data items with heterogeneous data structures is considered. Further interesting and useful work on blocking can be found in [41] [42] [43] and [44].

Some duplicate detection approaches produce probabilistic data as result data for modeling ambiguous duplicate decisions [45] [46] or for modeling uncertain merging results [19]. None of these studies, however, handle probabilistic data as source data. In contrast, in current research on the integration of uncertain data [47], deduplication is not considered. To the best of our knowledge, we are the first who consider the problem of blocking in the context of duplicate detection in probabilistic data. Nevertheless, to adapt blocking to probabilistic data we make recourse to techniques already used in the fusion of certain data tuples as proposed in [18] [19]. Moreover, we made some first proposals about the in-depth comparison of  $x$ -tuples in [15].

### 6. CONCLUSION

Duplicate tuples are pervasive problems of data quality. To efficiently apply duplicate detection on large data sets, the search space has to be initially reduced by a blocking technique. Until now, duplicate detection, and especially blocking, has only been considered for certain data. Nevertheless, duplicates are a quality problem in probabilistic databases, too. In this paper we propose different strategies to adapt the Sorted Neighborhood Method, which is a state-of-the-art blocking technique, to probabilistic source data. We present strategies based on certain keys created from probabilistic entity representations and shortly discuss possible strategies based on probabilistic keys. The benefit of using certain keys is that these strategies can also be applied to other key-based blocking techniques without any specific adaptation. In contrast, strategies based on probabilistic keys need to be tailor-made for each blocking technique. Our experimental evaluations of the certain key approaches show that creating multiple certain keys per entity representation is more effective than creating a single certain key per entity representation. Moreover, using multiple keys turned out to be more robust against a varying dirtiness or uncertainty of the source

data than using a single key. Finally, we observe that intermediately created probabilistic keys can improve the efficiency of the approaches based on multiple certain keys further on.

In future research, we aim to accelerate our blocking approaches, especially the multi-pass over possible world approaches, by using the Map-Reduce framework. Moreover, we plan to focus on strategies for probabilistic key based blocking adaptations in more detail.

## REFERENCES

- [1] D. Suciu, A. Connolly and B. Howe, "Embracing Uncertainty in Large-Scale Computational Astrophysics," *MUD Workshop*, pp. 63-77, 2009.
- [2] D. Z. Wang, E. Michelakis, M. J. Franklin, M. Garofalakis and J. M. Hellerstein, "Probabilistic declarative information extraction," *ICDE*, pp. 173-176, 2010.
- [3] D. Suciu, D. Olteanu, C. Re and C. Koch, *Probabilistic Databases*, Morgan & Claypool Publishers, 2011.
- [4] T. J. Green and V. Tannen, "Models for incomplete and probabilistic information," *EDBT Workshops*, pp. 278-296, 2006.
- [5] O. Benjelloun, A. D. Sarma, A. Y. Halevy and J. Widom, "Uldbs: Databases with uncertainty and lineage," *PVLDB*, pp. 953-964, 2006.
- [6] F. Naumann and M. Herschel, *An Introduction to Duplicate Detection*, Morgan & Claypool Publishers, 2010.
- [7] A. K. Elmagarmid, P. G. Ipeirotis and V. S. Verykios, "Duplicate Record Detection: A Survey," *TKDE*, pp. 1-16, 2007.
- [8] M. Jaro, "Advances in Record Linkage Methodologies as Applied to Matching the 1985 Census of Tampa Bay, Florida," *Journal of American Statistical Society* 84, pp. 414-420, 1985.
- [9] R. Baxter, P. Christen and T. Churches, "A comparison of fast blocking methods for record linkage," *ACM SIGKDD Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pp. 25-27, 2003.
- [10] P. Christen, "A survey of indexing techniques for scalable record linkage and," *TKDE*, 2011.
- [11] M. A. Hernandez and S. J. Stolfo, "The Merge/Purge Problem for Large Databases," *SIGMOD Conference*, pp. 127-138, 1995.
- [12] S. Abiteboul, P. C. Kanellakis and G. Grahne, "On the representation and querying of sets of possible worlds," *Theor. Comput. Sci.*, pp. 158-187, 1991.
- [13] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems, Berlin: Springer, 2006.
- [14] J. R. Talburt, *Entity Resolution and Information Quality*, Morgan Kaufmann Publishers, 2011.
- [15] F. Panse, M. van Keulen, A. de Keijzer and N. Ritter, "Duplicate Detection in Probabilistic Data," *ICDE Workshops*, pp. 179-182, 2010.
- [16] T. de Vries, H. Ke, S. Chawla and P. Christen, "Robust record linkage blocking using suffix arrays," *CIKM*, pp. 305-314, 2009.
- [17] L. Kolb, A. Thor and E. Rahm, "Multi-pass sorted neighborhood blocking with mapreduce," *Computer Science - R&D*, p. 45–63 , 2012.
- [18] J. Bleiholder and F. Naumann, "Data fusion," *ACM Comput. Surv.*, p. 41, 2008.
- [19] P. Andritsos, A. Fuxman and R. J. Miller, "Clean Answers over Dirty Databases: A Probabilistic Approach," *ICDE*, p. 30, 2006.
- [20] I. Ilyas and M. A. Soliman, *Probabilistic Ranking Techniques in Relational Databases*, Morgan & Claypool Publishers, 2011.
- [21] P. Christen, *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, 2012.
- [22] P. Lehti and P. Fankhauser, "A precise blocking method for record linkage.,," *DaWaK*, p. 210–220, 2005.
- [23] S. Yan, D. Lee, M.-Y. Kan and C. L. Giles, "Adaptive sorted neighborhood methods for efficient record linkage.,," *JCDL*, pp. 185-194, 2007.
- [24] A. N. Aizawa and K. Oyama, "A fast linkage detection scheme for multi-source information integration.,,"

- WIRI*, p. 30–39, 2005.
- [25] A. Feekin and Z. Chen, "Duplicate detection using k-way sorting method.,," *SAC*, p. 323–327, 2000.
  - [26] P. Christen and R. Gayler, "Towards scalable real-time entity resolution using a similarity-aware inverted index approach.,," *AusDM*, pp. 51-60, 2008.
  - [27] F. Naumann and U. Draisbach, "A Generalization of Blocking and Windowing Algorithms for Duplicate Detection," *ICDKE*, 2011.
  - [28] L. Jin, C. Li and S. Mehrotra, "Efficient record linkage in large data sets.,," *DASFAA*, pp. 137-152, 2003.
  - [29] A. E. Monge and C. Elkan, "An efficient domain-independent algorithm for detecting approximately duplicate database records.,," *DMKD*, 1997.
  - [30] S. Y. Sung, Z. Li and S. Peng, "A fast filtering scheme for large database cleansing," *CIKM*, p. 76–83, 2002.
  - [31] L. Gu and R. A. Baxter, "Adaptive filtering for efficient record linkage.,," *SDM*, 2004.
  - [32] H. sik Kim and D. Lee, "Harra: fast iterative hashed record linkage for large-scale data collections.,," *EDBT*, p. 525–536, 2010.
  - [33] J. Nin and V. Torra, "Blocking anonymized data.,," *AGOP*, p. 83–87, 2007.
  - [34] W. W. Cohen and J. Richman, "Learning to match and cluster large high-dimensional data sets for data integration.,," *KDD*, p. 475–480, 2002.
  - [35] A. McCallum, K. Nigam and L. H. Ungar, "Efficient clustering of high-dimensional data sets with application to reference matching.,," *KDD*, p. 169–178, 2000.
  - [36] L. Shu, A. Chen, M. Xiong and W. Meng, "Efficient spectral neighborhood blocking for entity resolution.,," *ICDE*, p. 1067–1078, 2011.
  - [37] B. Kenig and A. Gal, "Efficient entity resolution with mfiblocks," Technion, 2011.
  - [38] J. Nin, V. Munt'és-Mulero, N. Martínez-Bazan and J.-L. Larriba-Pey, "On the use of semantic blocking techniques for data cleansing," *IDEAS*, pp. 190-198, 2007.
  - [39] S. Puhlmann, M. Weis and F. Naumann, "Xml duplicate detection using sorted neighborhoods.,," *EDBT*, p. 773–791, 2006.
  - [40] G. Papadakis, E. Ioannou, C. Niederee and P. Fankhauser, "Efficient entity resolution for large heterogeneous information spaces.,," *WSDM*, pp. 535-544, 2011.
  - [41] M. Bilenko, B. Kamath and R. J. Mooney, "Adaptive blocking: Learning to scale up record linkage.,," *ICDM*, pp. 87-96, 2006.
  - [42] P. Christen, "Towards parameter-free blocking for scalable record linkage.,," 2007.
  - [43] M. Michelson and C. A. Knoblock, "Learning blocking schemes for record linkage.,," *AAAI*, 2006.
  - [44] S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald and H. Garcia-Molina, "Entity resolution with iterative blocking.,," *SIGMOD*, p. 219–232, 2009.
  - [45] G. Beskales, M. A. Soliman, I. F. Ilyas and S. Ben-David, "Modeling and Querying Possible Repairs in Duplicate Detection," *PVLDB*, p. 598–609, 2009.
  - [46] M. van Keulen, A. de Keijzer and W. Alink, "A Probabilistic XML Approach to Data Integration," *ICDE*, pp. 459-470, 2005.
  - [47] P. Agrawal, A. D. Sarma, J. Ullman and J. Widom, "Foundations of uncertain-data integration," *PVLDB*, pp. 1080-1090, 2010.

# RESEARCH ON THE ROLE OF SOCIAL MEDIA AND MOTIVATION TO USE IN THE LOCAL COMMUNITY

## INDEX OF INFORMATION QUALITY AND PRIVATE SPACE FUNCTION

(Research-in-Progress)

**Yasuhiro Tanaka<sup>+++</sup>**  
Senshu University, Japan  
S120062@senshu-u.jp

**Akihisa Kodate<sup>\*\*</sup>**  
Tsuda College, Japan  
kodate@tsuda.ac.jp

**Abstract:** Use of social media in an attempt to aggressively provide information to local communities is increasing, but methods for specifically evaluating the effects of their use merit are for further research. This research aims to evaluate the effects of social media in the local community and identify their role by quantifying characteristics of motivation to use social media. To that end, Web-based questionnaire surveys were conducted with general users in four cities (250 persons per city) and across Japan (2,000 persons). In the process of quantification, we conducted an evaluation based on information quality and the index of private space function by applying the TAM (Technology Acceptance Model) rating scale, which have been used in analyzing Internet use. The analysis suggests that users of social media evaluate their value as communication tools. It is also found that recognition of the private space function will not influence the use of social media but the evaluation of information quality.

**Key Words:** Social media, information quality, contextual IQ, index of private space function, Technology Acceptance Model

## 1. INTRODUCTION

Since the arrival of the age of Web 2.0, social media, including blogs and social network services (SNSs), have become remarkably widespread. The recent surge in the use of Twitter is drawing attention. The *2010 White Paper Information and Communications in Japan* [10] reports the results of a survey in which respondents were asked to choose all the social media that they had ever used from ten types of social media, i.e., blogs, video-sharing websites, bulletin board services, social network services (SNSs), information-sharing websites, microblogging, social gaming, community broadcasting, the Metaverse, and augmented reality, showing that 77.3% had used blogs, followed by video-sharing sites and bulletin board services at 62.8%, and SNSs at 53.6%. With regard to the frequency of using SNSs, blogs, or microblogging, about 30% of respondents answered that they use them almost daily. Growing numbers of local communities have been launching SNSs as a platform where these social media directly function as tools for regional revitalization. As of February 2011, 469 local SNSs existed, but it has been pointed out that not many of them have brought about any effects of regional revitalization with proactive

<sup>++</sup>Senshu University, Institute for development of Social Intelligence, Center for Social Capital Studies  
Waseda University, Institute of Asia-Pacific Studies

<sup>\*</sup> Institute for Information and Communications Policy of the Ministry of Internal Affairs and Communications

<sup>\*</sup> Department of Computer Science, Faculty of Liberal Arts, Tsuda College

participation by local residents [11]. Meanwhile, with the spread of Twitter and smartphone use, the number of local governments that have started to use Twitter to provide information is increasing. The number of local governments registered on govtter<sup>26</sup> totals 250 as of June 30, 2012. There is also a report on an initiative that uses Twitter to promote the city of Yokote in Akita Prefecture.

Kaplan and Haenlein [6] state that various social media can be classified by using indices of social presence/media richness and self-presentation/self-disclosure. For example, blogs are high in self-presentation/self-disclosure but low in social presence/media richness, while SNSs are as high as blogs in terms of self-presentation/self-disclosure but higher than blogs in social presence/media richness. Users can choose which media to use depending on their purpose. Effective use of these media is likely to contribute to new development of local media. To that end, as the study by Goto, et al. [5] shows, quantitative analysis of the possible effects of social media will become increasingly important.

A well-known analysis model of information system acceptance is the Technology Acceptance Model (TAM) by Davis [1]. Lee, et al. [8] point out, based on their analysis of 101 titles of literature on TAM published from 1986 to 2003, that TAM has continually been extended according to the subject of analysis, and that TAM2, which was introduced by Venkatesh and Davis in 2000, has further improved the accuracy. Recently in 2009, it was used in research conducted by Kondo's team on determinants and actual use of the Internet. On the other hand, the simplicity of the model has been pointed out [18], and studies on analytical models suitable for social media such as SNSs have been conducted [Theotokis 2009] [4], the effects of which have not been fully verified. Models for evaluating system acceptance, like TAM, focus on evaluation of systems, including convenience, ease of use, and usefulness; they cannot be considered evaluation models that fully take into account service value, or cumulative information quality, of social media that has been rapidly developing in recent years. In evaluating the acceptance of social media, we hypothesize that an evaluation based on the information quality in the context of communication enables us to identify motivation to use social media.

In addition, the impact of media communication space or the function of that space must be considered as factors that influence social media use behavior. Recently, Japan has had frequent occurrences of cases where people, primarily the younger generations, post their own criminal acts or information (e.g., private information on celebrities) that come to their knowledge in the course of their duties, such as part-time jobs, and these are drawing social criticism. Space created by social media functions as public space while being oriented toward private communication space; social media seem to create a paradoxical space, which is unique to virtual space. We assume that recognition of the space function specific to social media greatly influences motivation to use and use behavior. Accordingly, this research applies the index of private space function developed by Tomari's team [14] to analyze the possible impacts that users' recognition of media space, which is developed by social media, may have on motivation to use.

In light of the results of the preceding study by Kondo and Umino [7], this research intends to evaluate motivation to use social media based on objective information quality in the context of communication, rather than from the perspective of system evaluation, by incorporating the indicator of information quality presented by Wang, et al. [17] into a TAM-based social media acceptance evaluation. We also assume that addition of the index of private space function to the model allows evaluation of motivation to use social media from two perspectives: information quality, and recognition of space function.

This research also includes a survey of the characteristics of social media acceptance by Internet users in regions that are advanced in informatization, where social media have been adopted on local government's independent initiative.

We conducted questionnaire surveys with general Internet users in four cities (250 persons per city) and across Japan (2,000 persons) on motivation to use social media.<sup>27</sup> We use the survey results and develop

<sup>26</sup> Open Government. "Govtter: Collection of Central and Local Governments' Initiatives Using Twitter." 2010. Web. June 29, 2012. <<http://govtter.openlabs.go.jp/>>

<sup>27</sup> Of the ten types of social media, this paper covers blogs, bulletin board services, SNSs, information-sharing websites, and microblogging, while excluding video-sharing websites, social gaming, community broadcasting, the

an evaluation model that combines TAM-based information quality with the index of private space function to discuss conditions for social media to be used. Social media used in this research are listed in Table 1.

Social Networking Service (SNS)	An online service that facilitates the building of social networks on the Internet by linking users. It offers a wide variety of functions, including diary, review, and video-sharing. SNSs covered in this study are services that provide these functions in a comprehensive manner, like Facebook or mixi.
Blog	A website that typically displays diary-like entries in reverse chronological order and allows readers to leave comments, which are updated regularly. This research defines blog users as persons who have their own blog site and publish their diaries.
Microblogging	A simplified blog that limits the number of characters to about 140, and opens its content to the general public and to a specific group (followers). This research focuses on "Twitter," which is more familiar to general users, instead of microblogs.
Information sharing website	A website service provided for the exchange and sharing of information, including product reviews, word of mouth, or cooking recipes (e.g., COOKPAD). This research defines information sharing websites as part of SNS functions, and does not survey the websites alone.
Bulletin Board Service (BBS)	A function implemented to provide a platform for discussions, exchanging information or chatting on the Web. This research defines BBS as part of SNS or blog functions, and does not survey BBSs alone.

**Table 1 Social Media Covered by This Research**

This paper outlines the survey in Section 2, analyzes motivation to use social media by applying TAM in Section 3, and summarizes the entire research in Section 4..

## 2. OUTLINE OF THE SURVEY

This study analyzes and discusses triggers for social media use based on the results of an online questionnaire survey and interview survey. The two surveys are outlined as follows.

### 2.1 Outline of Web-based questionnaire surveys

To identify motivation to use social media, Web-based questionnaires, as well as interviews, were conducted with the residents of four cities (cities of Mitaka, Okayama, Yamaguchi, and Matsumoto), which were selected from those cities launching local social media. In conjunction with this, a nationwide web-based questionnaire was also conducted. In choosing the four cities, we referred to the *Local Government Informatization Yearbook 2009–10* [12]. We selected the cities of Mitaka and Okayama from cities with the highest scores for information/service, accessibility, and informatization policy, and the cities of Yamaguchi and Matsumoto as median cities.

An outline of the Web-based questionnaire is shown in Table 2.

---

Metaverse, and augmented reality as hardly any have been launched or operated by local governments.

Survey period	From February 18 to 22, 2011
Survey method	Internet questionnaire survey using the survey panel
No. of valid response	3,000 persons
Respondents	250 persons in Mitaka City, Tokyo
(Targeting over 18 years old)	250 persons in Okayama City, Okayama 250 persons in Yamaguchi City, Yamaguchi 250 persons in Matsumoto City, Nagano
	2,000 persons across Japan
No. of questions	134 questions
Questionnaire item breakdown	Personal attributes (age, gender, area of residence, disposable income, etc.) Use of social media Self-efficacy (general and electronic devices) Technology acceptance model evaluation items Index of private space function Use of the Internet and communication services

**Table 2 Survey Outline**

The age structure of the respondents is shown in Figure 1. The average age of all respondents was 41.9. Average age in the survey cities was 44.5 in Mitaka, 42.5 in Okayama, 41.0 in Yamaguchi, and 40.6 in Matsumoto. The average age across Japan, excluding the four cities, was 41.8.

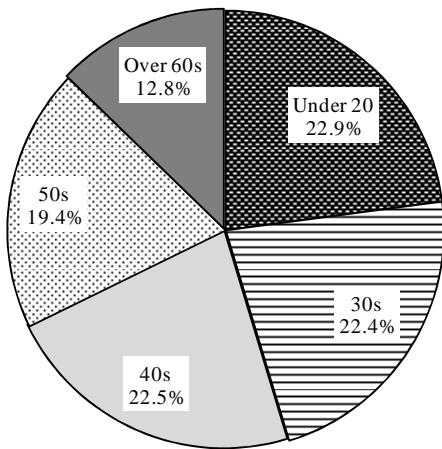
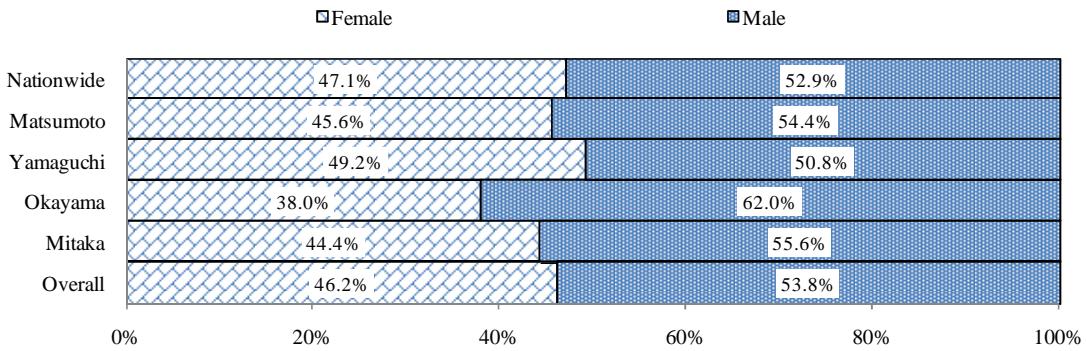
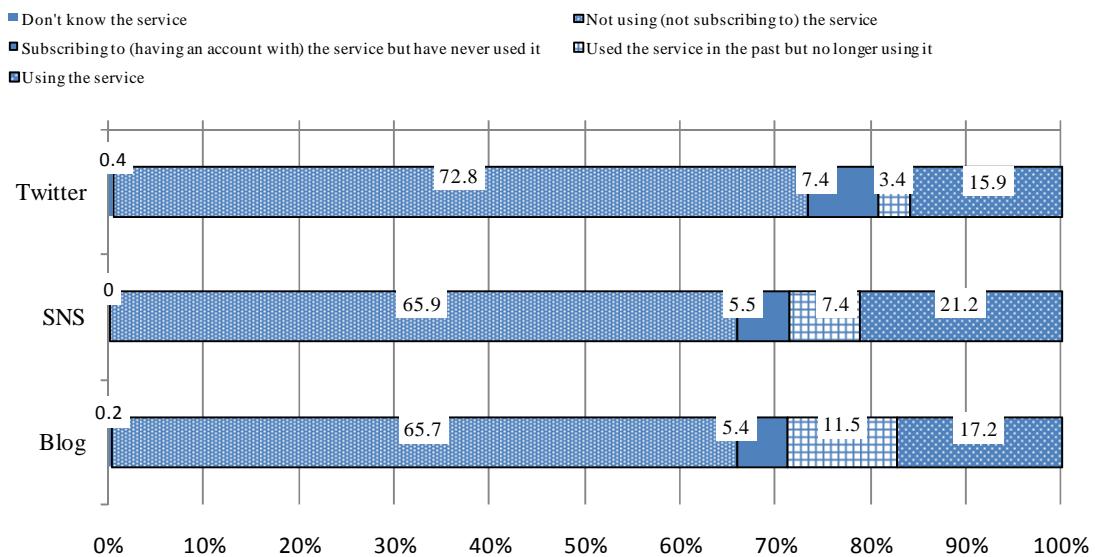
**Figure 1 Age Structure of Respondents**

Figure 2 shows the male-female ratio of the respondents. The proportion of males was the highest in Okayama at 62.0%, while it was the lowest in Yamaguchi at 50.8%.

**Figure 2 Male to Female Ratio of Respondents**

Among the all respondents, there were 515 blog users, 635 SNS users, and 478 Twitter users. Figure 3 gives details of the use of these social media.

**Figure 3 Subjects' Use of Blogs, SNSs and Twitter**

### 3. ANALYSIS OF TRIGGERS FOR SOCIAL MEDIA USE

This section aims to identify motivation to use the so-called social media, including SNSs, blogs, and Twitter, in light of the results of the preceding study by Kondo and Umino [Kondo and Umino 2009]. We developed a model by applying TAM based on exploratory factor analysis, and then derived three estimate results: motivation to use for all subjects, motivation to use for social media users, and motivation to use for the subjects in the four selected cities.

#### 3.1 Outline of TAM

To clarify motivation to use the Internet, Kondo and Umino used the Technology Acceptance Model (TAM) to verify whether motivation to use the Internet in Japan is explainable. They explain that TAM is a human behavior and attitude model introduced by Davis [2] [3] to explain computer users' be-

havior. Studies by Kondo explain the details of TAM. TAM and TAM2, which is an extended model, are for modeling and analyzing the processes in which perceived usefulness and perceived ease of use form attitudes toward use. TAM was originally advocated by Davis, but Taylor and Todd [13] point out the need to modify or extend the model. In response to these comments on TAM, Venkatesh and Davis [16] proposed TAM2 by extending and modifying TAM with social norms, user experience, and so on.

### **3.2 Analysis of motivation to use**

In analyzing motivation to use, this research applied the TAM structure model used by Kondo's team to design and analyze the survey sheets.

#### **Setting of evaluation items**

In light of the preceding study by Kondo and Umino, we prepared questions to clarify motivation to use social media based on the rating scale method used by Venkatesh and Davis [16] (see Table 3), to which we applied categories and dimensions of information quality invented by Wang's team. Furthermore, we added items on communication that seem important as social media rating scales (see Table 4), and items on the index of private space function. The questionnaire items were presented to be responded to on the five-level Likert scale: "Strongly agree," "Agree," "Neither agree nor disagree," "Disagree," and "Strongly disagree."

With regard to the dimensions of information quality suggested by Wang's team, we excluded key indices for evaluating information qualities of accuracy and objectivity. Social media enable users to create communities as they communicate freely. In other words, the important values of social media may lie in objective information that cannot be measured with the dimensions of accuracy and objectivity. According to Leo [9], the intrinsic IQ category defined by Wang is an index to quantitatively and objectively measure values of information, while dimensions included in the contextual IQ category are defined as application-dependent metrics, or subjective standards of value that can be evaluated only in a specific context. And in social media, various users send and share information based on their own information sources or subjective feelings. Therefore it is difficult to evaluate their overall accuracy and objectivity of social media, however respondents can evaluate accuracy and objectivity of individual information.

To evaluate motivation to use in the context of communication between users, this research weighs the evaluation of the contextual IQ of social media, and accordingly, excludes some information quality dimensions, including accuracy and objectivity in the intrinsic IQ category, with the aim of developing an evaluation model centering on the contextual IQ dimension.

Private space is defined as "domain (space and time) of individuals where they can act freely, separate from their social roles, without worrying about what others think of them" [19]. Tomari's team [14] shows that private space can be structured into three types with seven functions. These three types of space are: space that can be used exclusively, space that can be shared, and space where individuals can liberate themselves. And the seven functions are: tension release (TR), self-contemplation (SCo), focus on an issue (FI), frank communication (FC), change of pace (CP), emotional release (ER), and self-change (SCh). Tomari's team conceptualizes private space function (PSF) as a function of living space (time) to fulfill seven desires toward private space. From one aspect, private space functions mean seven inner desires toward private space, while, from another aspect, they refer to functions of living space to satisfy the desires from the perspective of living space (time). Therefore, a measure of the space functions must be evaluated from both sides (degrees of necessity and securement) [19].

The index of private space function used in this research is a simple version consisting of seven items and seven indices (the original version has seven items and 31 indices) developed by Tomari and Yoshida [Tomari and Yoshida 1999]. Furthermore, with regard to change of pace and emotional release that were considered slightly low in Cronbach's alpha in the Tomari team's credibility validation, this research excluded change of pace, which slightly overlaps tension release, and consequently incorporated

six functions and six items into the model.

Q1	A blog or SNS is indispensable to me.
Q2	The services are useful for my life and work.
Q3	My parents or family, and many of my friends are using the services.
Q4	Using a blog or SNS is the king of status symbols.
Q5	I enjoy using a blog or SNS.
Q6	I feel anxious that if I don't use a blog or SNS I will be left behind.
Q7	The services are free of charge.
Q8	I want to use a blog or SNS if I can.
Q9	How to use the services is clear and easy to understand.
Q10	Using a blog or SNS is crucial to my work or study.
Q11	A blog or SNS is much more convenient than their alternatives.
Q12	I can use a blog or SNS without being taught by someone or referring to manuals or books/websites that explain how to use it.
Q13	It is difficult for me to understand how to operate a mobile phone.
Q14	A blog and SNS improves my ability.
Q15	I can use it without thinking or learning much.
Q16	I'm using a blog or SNS because I want to use it.
Q17	I'll face inconvenience in my work or study if I don't use a blog or SNS.
Q18	I can manage to use a blog or SNS even if I don't know how to operate it and there is no one around to teach me.
Q19	How to operate it on a personal computer is too difficult to understand.
Q20	Using a blog or SNS improves the efficiency of my life or work.
Q21	I can use it easily and do what I want to do with it.
Q22	I'm using a blog or SNS but it's not because someone asked me to do so.
Q23	Using a blog or SNS brings me economic benefit or income.
Q24	I feel anxious that a blog or SNS will lead to leakage or misuse of personal information.
Q25	I can't use a blog or SNS without someone teaching me.
Q26	I'll face inconvenience if I don't use it because many people around me are using it.
Q27	Many people in my workplace or school use it.
Q28	People with a high standard of living are using blogs or SNSs.
Q29	The services have great advantages when comparing their benefits against their fees or costs.
Q30	I feel worried that I will become a victim of fraud since I use a blog or SNS.
Q31	Personal computer or other devices required are too expensive.

**Table 3 Questionnaire Items based on TAM2 Rating Scale Method in Preceding Study**

Q32	I keep a closer relationship with my family and friends by using a blog or SNS.
Q33	I can become friends with new people by using a blog or SNS.
Q34	I can spread my ideas or opinions to the world by using a blog or SNS.
Q35	I can find answers to my worries and problems by using a blog or SNS.
Q36	I can speak my mind on a blog or SNS.
Q37	I can share my hobbies or interests with many people by using a blog or SNS.

**Table 4 Additional Questionnaire Items of This Research**

Regarding the scope of data on actual media use behavior, which serves as a final dependent variable, we created a composite variable for analysis by combining use and non-use of SNSs, blogs, and Twitter, with the aim of reflecting not only the actual SNS use but also the actual use of social media as a whole, including blogs and Twitter.

#### **Estimate results based on preceding study model**

To analyze motivation to use social media, this research conducted exploratory factor analysis on TAM2 evaluation items to redefine the factors constituting the model and build a new model, and then incorporated the index of private space function into the new model.

#### **Development of an estimation model**

To develop a model for analyzing motivation to use social media, we conducted an exploratory factor analysis with regard to set evaluation items. For factor extraction, we used the maximum likelihood method, and conducted an analysis with promax oblique rotation. For factor analysis, we used PASW

Statistics 18 (currently, SPSS).

In our first exploratory factor analysis, Q11, Q29, Q31 and Q32 were small in factor loading. We therefore excluded them, and conducted the analysis again.

As a result of the exploratory factor analysis, we extracted seven factors. By referring to TAM factors, etc., we determined the first factor as “perceived usefulness,” the second as “perceived ease of use,” the third as “communication,” the fourth as “evaluation of benefits of use,” the fifth as “anxiety over use,” the sixth as “potential risk,” and the seventh as “subservience to others.” The resultant correspondence between these factors and the information quality categories and dimensions defined by Wang is listed in Table 5 below.

IQ Category	TAM (thesis) category	IQ Dimension	Question No.	TAM Item
Contextual IQ	Perceived usefulness	Value added	Q17	Face inconvenience in work or study without it
		Value added	Q20	Improves efficiency
		Value added	Q10	Crucial to work or study
		Value added	Q23	Brings income
		Value added	Q14	Improves ability
		Reliability	Q26	Face inconvenience without using it
		Reputation	Q6	Anxious that I'll be left behind
		Reputation	Q28	High living standard of users
		Reputation	Q4	Status
		Understandability	Q15	Can use without thinking
Representational IQ	Perceived ease of use	Understandability	Q18	Can use without being taught
		Understandability	Q21	Can do what I want to do with it
		Understandability	Q12	Can use without manuals
		Understandability	Q9	How to use is clear and easy
		Understandability	Q22	Using it voluntarily
		Accessibility	Q7	Free of charge to use
		Reliability	Q34	Communicate my opinion
Intrinsic IQ	Communication	Reliability	Q35	Find answers to my worries
		Reliability	Q33	Can increase my friends
		Reliability	Q36	Can speak my mind
		Reliability	Q37	Can share hobbies
		Reputation	Q5	Can enjoy it
Intrinsic IQ	Benefit evaluation	Reputation	Q8	Want to use it if possible
		Reputation	Q1	Indispensable
		Reputation	Q2	Useful service
		Q16		Using it because I want to
		Q13		Mobile phone is difficult to use
Representational IQ	Anxiety over use	Understandability	Q19	Personal computer is difficult to use
		Understandability	Q25	Can't use it without someone to teach me
		Understandability		
Accessibility IQ	Potential risk	Security	Q30	Worried about fraud
		Security	Q24	Worried about personal information leakage
Intrinsic IQ	Subservience to others	Reputation	Q27	Everyone in my workplace or school are using it
		Reputation	Q3	People around me are using it

**Table 5 Correspondences between TAM Evaluation Items and Information Quality**

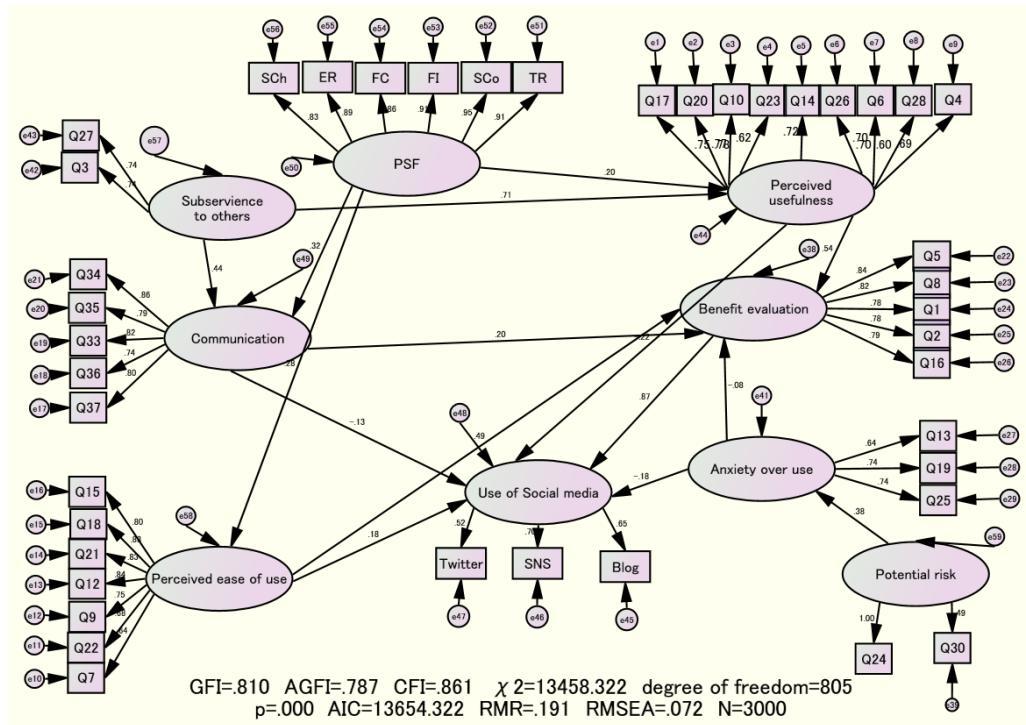
In order to develop an analytical model based on the seven factors obtained through the exploratory factor analysis, we conducted a correlation analysis of the factor score for the respective factors. Table 7 shows the results of the correlation analysis.

	1st factor	2nd factor	3rd factor	4th factor	5th factor	6th factor	7th factor
1st factor (Q17, Q20, Q10, Q23, Q14, Q26, Q6)		.298**	.500**	.611**	.236**	.247**	.001**
2nd factor (Q15, Q18, Q21, Q12, Q9, Q7)	.298**		.404**	.554**	.555**	.264**	.152**
3rd factor (Q34, Q35, Q33, Q36, Q37)	.500**	.404**		.549**	.343**	.164**	.236**
4th factor (Q5, Q8, Q1, Q2)	.611**	.554**	.549**		.482**	.084**	.056**
5th factor (Q13, Q19, Q25)	.236**	.555**	.343**	.482**		.168**	.149**
6th factor (Q30, Q24)	.247**	.264**	.164**	.084**	.168**		.334**
7th factor (Q27, Q3)	.001**	.152**	.236**	.056**	.149**	.334**	

**Table 6 Correlation between Extracted Factors (Pearson's Correlation Coefficient)**

## Estimate results 1 (all subjects)

We incorporated the index of private space function into the model, which we developed based on Tables 5 and 6, to conduct a covariance structure analysis. Consequently, we obtained the model shown in Figure 4 and the analysis results on motivation to use social media.



#### **Figure 4 Analysis Results for All Subjects (N=3,000)**

In terms of the goodness of fit of the model, GFI=0.810 and AGFI=0.787, which are smaller than the standard model fit of GFI=0.9 but exceeds the model fit when the TAM2 model is used for analysis (GFI=0.670, AGFI=0.599). All paths are confirmed as significant ( $p<0.01$ ).

As a result of model analysis, the path coefficient of “evaluation of benefits of use” was 0.75, which means that it is a major motivation to use social media.

This evaluation model proved that perceived usefulness and perceived ease of use influence evaluation of the benefits of social media use, and that the influenced evaluation of the benefits of use is the factor that eventually decides whether or not to use social media. Unlike the evaluation of the benefits of use factor in the preceding studies, the evaluation of the benefits of use factor presented in this model is assumed to comprehensively evaluate the media's value and benefits based on users' subjective recognition and sense.

The assumption of this research is that recognition of the private space function directly affects the use of social media. According to the analysis results, however, the path from the index of private space function to social media was not significant and was therefore rejected. It was confirmed that perceived ease of use, perceived usefulness, and communication influence the evaluation of information quality.

Perceived usefulness, perceived ease of use, and anxiety over use are assumed to be users' subjective recognition and value judgments of media function. The subjective value judgment seems to influence a comprehensive evaluation of the benefits of media use. In other words, the benefits of social media are evaluated based on users' subjective value judgments. That is, a value judgment of social media should not be evaluated based on objective indices, but based on the user's subjective value judgment, or contextual IQ.

### Estimate results 3: Comparative analysis of motivation to use social media in four cities

Next, we conducted model analyses for four cities (cities of Mitaka, Okayama, Yamaguchi, and Matsumoto; 250 subjects in each city) selected for an interregional comparison. We used the same analysis model as the model used in analyzing all subjects. For the scope of data on actual media use behavior, we set a composite variable of "social media use," which we created by combining use and non-use of blogs, SNSs and Twitter.

The analysis results are shown in Figures 5, 6, 7 and 8.

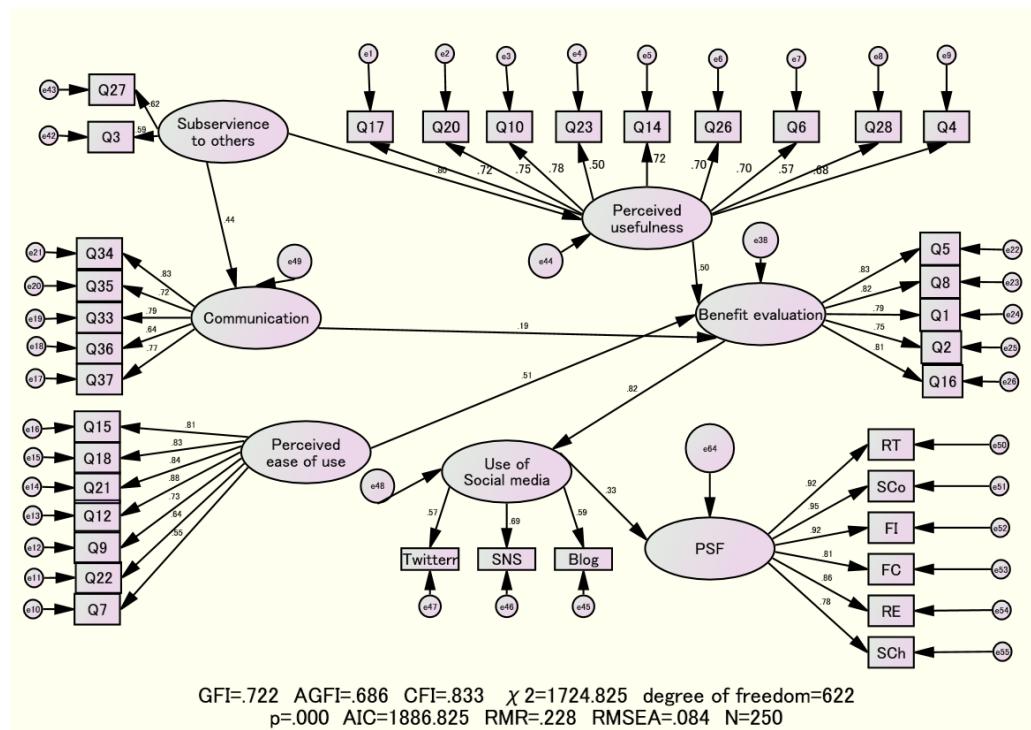


Figure 5 Analysis Results of Social Media Users in Mitaka (N=250)

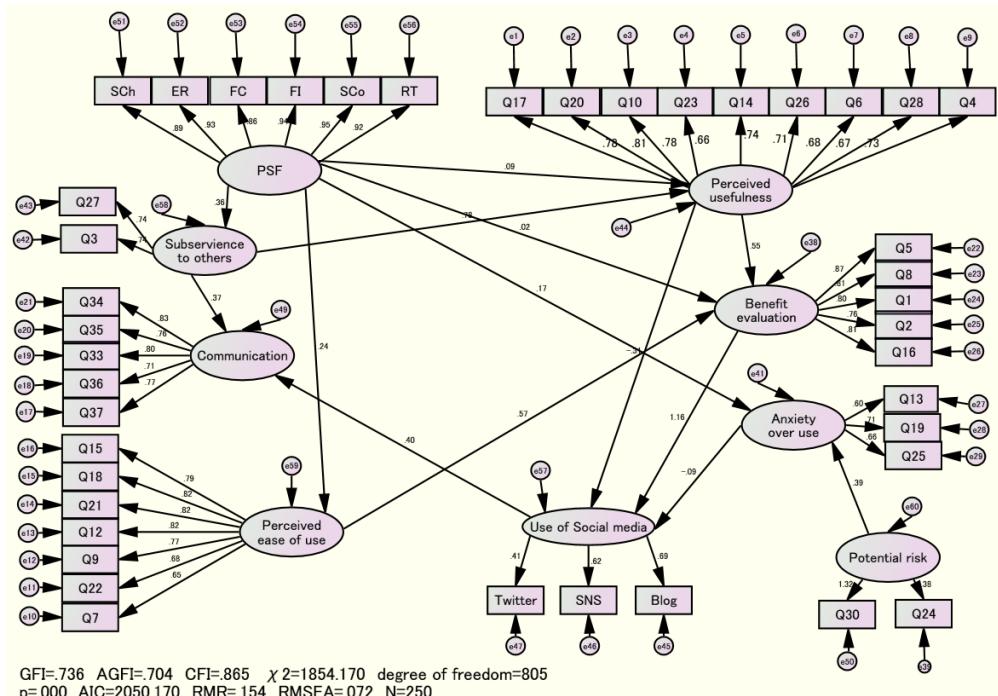


Figure 6 Analysis Results of Social Media Users in Okayama (N=250)

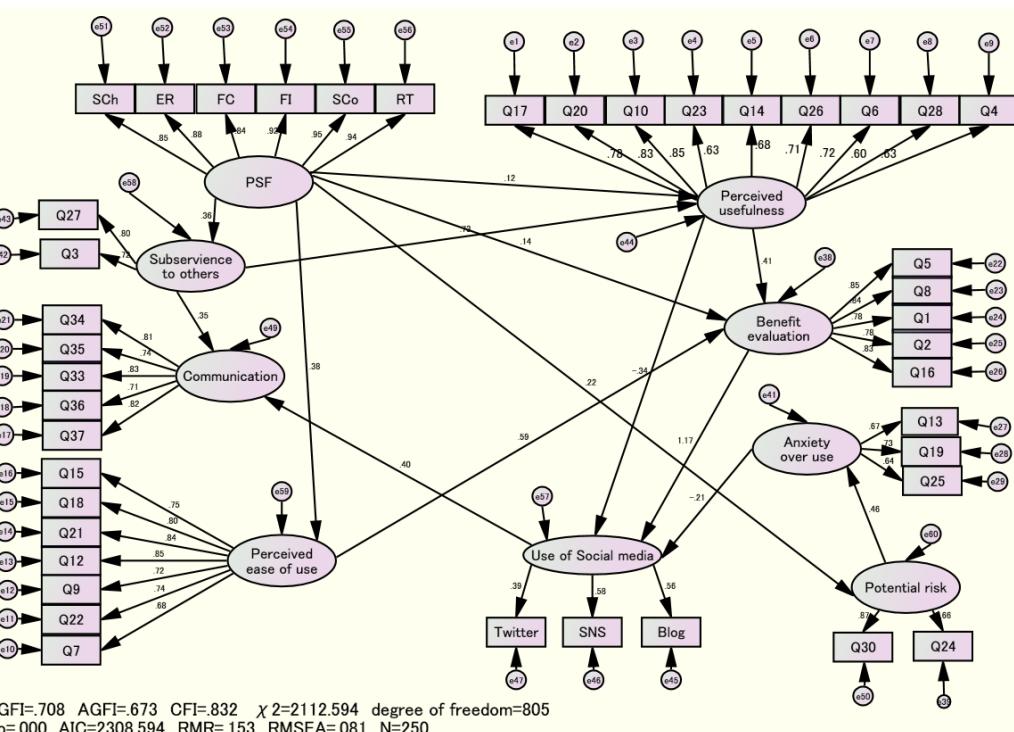


Figure 7 Analysis Results of Social Media Users in Yamaguchi (N=250)

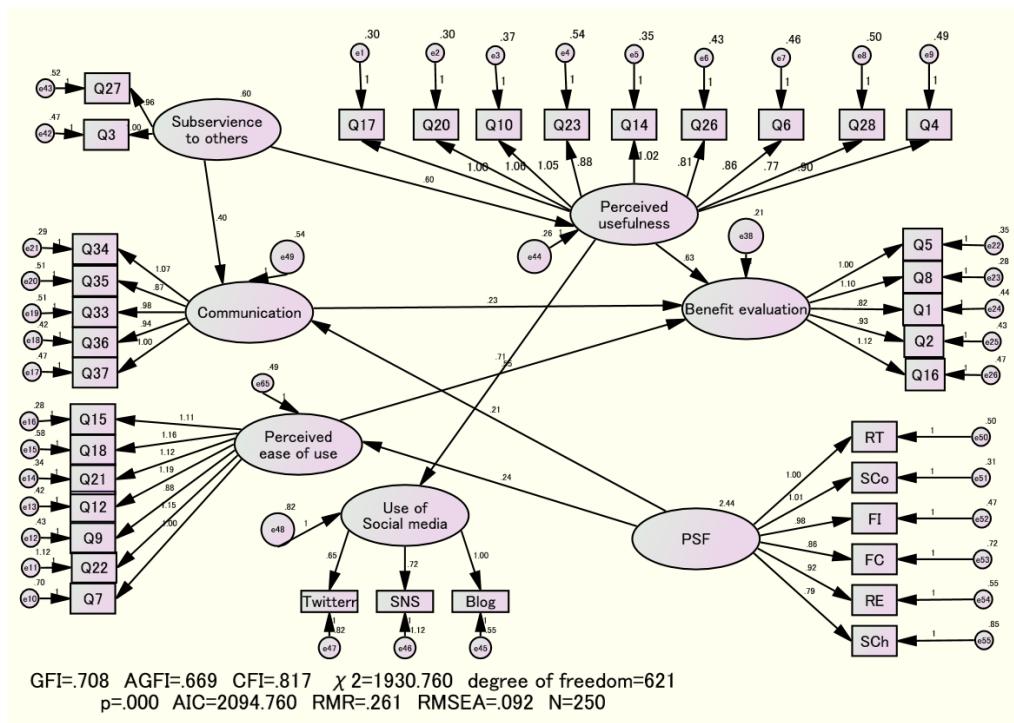


Figure 8 Analysis Results of Social Media Users in Matsumoto (N=250)

Comparison between the estimate results for each region with regard to the significance of the path coefficient confirmed two differences. In the cities of Mitaka and Okayama, the path coefficient from the perceived usefulness factor to the social media use factor was negative, while the path coefficient was not significant in the cities of Yamaguchi and Matsumoto. Only in Okayama was the path coefficient from perceived usefulness to social media use negative, while the path coefficients were not significant in the others.

As for the anxiety over use factor, the cities can be divided into two, Mitaka and Matsumoto, and Okayama and Yamaguchi. Between Okayama and Yamaguchi, there were differences in the significance of paths, but, in principle, all model factors developed in the analysis of all subjects were significant. On the other hand, in Mitaka and Matsumoto, the path from the anxiety over use factor was not significant ( $p>0.05$ ), so that the factor was excluded from the model.

With regard to the impact of the index of private space function, each city had different characteristics. In Mitaka, the path from recognition of the index of private space function to social media use and to various information quality items did not reach the level of significance, and therefore they were rejected. On the other hand, the influence of social media use on recognition of the private space function was confirmed. That is indicative of the possible impact of users' social media usage on recognition of media space.

In both Okayama and Yamaguchi, the impact of social media on information quality evaluation was confirmed, as in the case of the overall results. Particularly in Yamaguchi, the impacts of social media use on evaluation of the communication factor were confirmed.

In Matsumoto, it was confirmed that recognition of the private space function has an impact only on evaluation of the communication factor.

## **4. SUMMARY**

As a result of the analysis of motivation to use social media, this research revealed that perceived usefulness and perceived ease of use, which represent users' subjective value judgments, influence the evaluation of benefits of use, and that the evaluation of benefits of use is the major factor that motivates people to use social media. This means that introduction of an evaluation of information quality, including contextual IQ, allows evaluation of motivation to use social media and use behavior. Meanwhile, although we assumed that users' recognition of the private space function might influence social media use, the analysis results show that users' recognition of the media space function affects, not the use behavior, but evaluation of the information quality of social media. The impact of the recognition of space function on information quality evaluation requires further in-depth discussion.

Analysis of the characteristics of the four cities surveyed reveals that the impact of users' recognition of the private space function greatly differs between the four cities. Particularly in Mitaka, it is confirmed that social media use affects recognition of the space function. The differences due to regional characteristics in information quality evaluation and in the impacts of recognition of the media space function on motivation to use social media must be analyzed in detail, in association with the impacts of social media launched independently by local governments, as well as in light of the characteristics of local residents.

In developing an evaluation model, this research excluded accuracy and objectivity from the intrinsic IQ category defined by Wang, for value judgment and evaluation of motivation to use social media centering on communication between users based on the dimension of subjective contextual IQ. In Japan, however, local governments are taking initiatives to use social media for regional revitalization and information disclosure. Many of them use Facebook, Twitter, or other commercial social media, while some regions, such as the cities that this research picked up, have launched independent social media. In the meantime, as businesses have opened official Facebook pages and use Twitter, social media is becoming an indispensable tool for marketing and promotion. This social media use by governments and businesses is assumed to necessitate information quality of accuracy and objectivity, which this research excluded.

In light of these situations where public information from governments and businesses is used in social media, conducting an analysis by developing a model that incorporates accuracy and objectivity to comprehensively evaluate motivation to use social media is an issue for further research.

## **Acknowledgements**

We are especially grateful to Mayumi Yamauchi, Former Senior Researcher, Institute for Information and Communications Policy (IICP) of the Ministry of Internal Affairs and Communications for the great help extended to our survey and writing.

We would also like to thank individuals who provided cooperation with our research: Mr. Kasuya of the Planning Bureau, Okayama City; Mr. Ando of the Information Planning Division, Planning Bureau, Okayama City; Mr. Akita and Mr. Ikeda of Okayama Electronic Data Processing System Center Co., Ltd.; Mr. Takemoto and Mr. Minematsu of Okayama City Safe and Secure Network Office; Mr. Yonetomi, Information Management Division, General Affairs Department, Yamaguchi City; Mr. Harada, Yamaguchi Cable Television Co., Ltd.; Mr. Goto, Planning Department, Mitaka City; and Mr. Soyano and Mr. Kamikawa, Tourism and Hot Springs Division, Matsumoto City.

We conducted this research jointly with IICP with subsidies for Grants-in-Aid for Young Scientists (B) from the Japan Society for the Promotion of Science (Project No. 22700247) and a special research fund from Tsuda College. Our thanks go to all who have supported us.

## REFERENCES

- [1] Davis F.D, "Technology Acceptance Model for Empirically Testing New End-user Information Systems Theory and Results," *Unpublished Doctoral Dissertation*, MIT, 1986
- [2] Davis F. D., "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, September, pp.319-340,1989
- [3] Davis F. D., Bagozzi R. P. and Warshaw P. R., "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science*, Vol. 35, No.8, pp.982-1003, 1989
- [4] Donaldson O. and Duggan E., "Examining SNS Adoption through Motivational Lens," *Proceedings of the Seventeenth Americas Conference on Information Systems*, 2011
- [5] Goto S., Suwa H. and Ohta T., "Relation between the Purposes and the Effects of the Regional Social Networking Services: A Quantitative Analysis," *Journal of Socio-Informatics*, Vol.22-2, pp.17-26, 2011 (in Japanese)
- [6] Kaplan A. M. and Haenlein M., "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons Volume 53, Issue 1*, pp.59-68, January-February, 2010
- [7] Kondo K. and Umino A., *Research on motivation to use the Internet*, Ministry of Internal Affairs and Communications , 2009 (in Japanese)
- [8] Lee Y., K. Ozar A. K. and Larsen K. R. T., "Technology Acceptance Model: Past, Present and Future," *Communications of the Association for Information Systems(Vol. 12,Article 50)*, pp.752-780, 2003
- [9] Leo P., Wang R., David K., and William R., "Developing Measurement Scales for Data-Quality Dimensions," *Advances in Management Information Systems Vol.1*, 2005
- [10] Ministry of Internal Affairs and Communications, *2010 White Paper of Information and communication in Japan*,(<http://www.soumu.go.jp/johotsusintoeki/whitepaper/eng/WP2010/2010-index.html> ), 2010 (2012/6/30)
- [11] Ministry of Internal Affairs and Communications, *Survey of Leading-edge Examples of ICT Utilization in Japan*, 2010 (in Japanese)
- [12] Nikkei Business Publications, *Local Government Information Yearbook 2009–10*, Nikkei Business Publications, Inc., 2009(in Japanese)
- [13] Taylor S. and Todd P., "Understanding Information Technology Usage: A Test of Competing Models," *Information Systems Research Vol. 6 No. 2*, pp.144-176, 1995
- [14] Tomari S. and Yoshida F., "Psychological meanings of private space and its functions: A review of privacy research and proposal for new models", *Tsukuba Psychological Research Vol.20*, 1998(in Japanese)
- [15] Venkatesh V. and Davis F. D., "A Model of the Antecedents of Perceived Ease of Use: Development and Test," *Decision Sciences 27. 3.*, pp.451-481, 1996
- [16] Venkatesh V. and Davis F. D., "A theoretical extension of the technology acceptance model: Four longitudinal field studies", *Management Science 46:2*, pp.186-204, 2000
- [17] Wang R. Y. and Strong D. M., "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems, Vol. 12, No. 4*, pp.5-34, 1996
- [18] Wu P. F., "Opening the Black Boxes of TAM: Towards a Mixed Method Approach," *Thirtieth International Conference on Information Systems*, 2009.11.15-18, Phoenix, Arizona, U.S.A.
- [19] Yoshida F. and Hori H., *Index of Psychology II*, Saiensu Sha Co., Ltd. Publishers, 2001(in Japanese).

# **QUALITY OF SOCIAL MEDIA DATA AND IMPLICATIONS OF SOCIAL MEDIA FOR DATA QUALITY**

(Research-in-Progress)

**G. Shankaranarayanan**  
Babson College, U.S.A.  
[gshankar@babson.edu](mailto:gshankar@babson.edu)

**Bala Iyer**  
Babson College, U.S.A.  
[biyer@babson.edu](mailto:biyer@babson.edu)

**Donna Stoddard**  
Babson College  
[dstoddard@babson.edu](mailto:dstoddard@babson.edu)

**Abstract:** In the recent past, data generated by social media technologies have become part of organizational data. This data, together with the traditionally collected transactional data, is being used for marketing, product innovation and customer support. Understanding the data quality of the data generated by social media technologies is a critical first step towards managing the quality of organizational data today. In this paper, we present our findings from examining the quality of social media data and the impact of social media data on the quality of transactional data. Specifically, we look at the traditional dimensions of data quality and examine their applicability to social media data. We believe this is a first step towards gaining a better understanding of how to evaluate the quality of social media data. It also offers insights into the use of social media data for improving the quality of transactional data. With social media data, we posit that believability, a quality dimension that has received little attention in the context of traditional data, will gain significantly in stature. We present a model for evaluating believability and suggest methods for gauging believability of social media data.

**Key Words:** Social Media, Social Media Data, Social Media Tools, Data Quality, Quality Dimensions, Believability

## **INTRODUCTION**

Organizations manipulate the data and analyze it in multiple complex ways to satisfy the need to gather business intelligence and to monitor internal processes. Through partnerships as well as B2B web portals, organizations exchange data and use data from other organizations for mission-critical decisions. A key source of organizational data has been business transactions that are part of the business processes within the organization. In the recent past, data generated through the use of social media and social networks have also become part of the organizational data. It is evident that organizations use this new data for a variety of purposes such as generating and tracking leads, supporting customers, generating new product ideas and understanding market conditions. We refer to the data gathered through traditional means as transactional data. Included in transactional data are the data collected by clickstream and data on web analytics (e.g., Google analytics) as we treat these as data collected on web-transactions (such as browsing and shopping). We refer to the data gathered through social media as social media data and also as non-transactional data.

Data is an organizational asset. Organizations gain value from the use of data for managing day-to-day operations, understanding the effectiveness and efficiency of their internal and external business processes and by gaining business intelligence through data analysis. High quality data offers superior usability

and business benefits. If organizational data is of poor quality, then, organizational performance is adversely affected. Unfortunately, organizational data is susceptible to quality defects [23]. Today, managing data quality is more critical than ever before given the value that organizations gain from organizational data. Over the last two decades research in data quality management has proposed several different techniques for managing data quality. These techniques may be broadly classified into three categories. There are techniques and methods that help measure data quality (e.g., [2][27][30]). Then there are techniques that help improve data quality (e.g., [8][10][15][19][25]). Finally, there is the research that examines the impact of data quality on decision making in organizations to better manage data quality (e.g., [4][9][11][35]). All three categories described above are founded on the notion that data quality is a multi-dimensional construct ([6][26][34]). Data quality is measured along multiple dimensions such as accuracy, completeness, timeliness, and relevance, to name a few. The quality dimensions are discussed in greater detail in the next section. It is important to note that all of the research in data quality examines data quality in the context of traditional, transactional data. Very little has been done to examine the quality of data obtained through social media tools and technologies.

The high-level objectives of this research are to understand how to manage the quality of social media data and to understand whether social media tools and technologies can help improve data quality. As a step towards these objectives, we first examine the applicability of data quality management techniques (that have been successfully applied to the management of transactional data) to social media data. We focus on the data quality dimensions that have been used to measure and manage quality of transactional data and evaluate the applicability of these dimensions to social media data. *The first contribution that this paper makes is an analysis of existing quality dimensions and their applicability to manage the quality of social media data.* Second, we examine the implications of social media data for managing data quality. Specifically, we focus on social media tools and how organizations use these to manage the quality of the social media data. *This is the second contribution of this paper and a key first step towards understanding the utility of social media data for data quality.* We then examine believability, a data quality dimension we posit, will play an important role in the context of managing data quality of social media data. Believability has not received a lot of attention from data quality researchers. *The third contribution of this paper is a model for evaluating believability of social media data and an illustration of how believability can be evaluated using social media tools and data.*

The remainder of the paper is organized as follows. Section 2 presents an overview of the relevant literature to define the scope of this research paper. Section 3 describes the dimensions of data quality as applied to transactional data and discuss the applicability of these dimensions to social media data. In section 4, we describe the impact of social media data on the quality of transactional data using examples of real-life social media data and its application for practice. We also describe a model for evaluating believability and discuss the components of believability that can be measured using social media tools and data. We conclude the paper by reiterating the key contributions our plans for further research into managing the quality of social media data.

## BACKGROUND AND RESEARCH SCOPE

We begin by describing social media data to identify the differences between social media data and transactional data. Despite the existence of a large body of work on structured, unstructured and semi-structured data, we summarize the salient issues for the purpose of scoping our work. We also describe the research that addresses data quality of social media data. To position our research in the context of “big data”, we draw attention to the fact that “big data” includes transactional data, social media data, data from sensors, data from GPS and telecommunications, besides the images and pictures that are generated and posted. Big data is the term used to refer to data sets that grow so large that these cannot be managed by traditional databases and data management tools. Our research here only targets social media

data and distinguishes it from transactional data, both being sub-components of “big data”.

## **Social Media Data**

A popular definition of social media incorporates the concepts of Web 2.0 and user-generated content. Kaplan and Haenlein define social media as a group of internet-based applications that build on the ideological and technological foundations of Web 2.0 and that allow the creation and exchange of user-generated content [20]. Broadly speaking, social media includes collaborative projects such as wikis (e.g., Wikipedia) and social book-marking applications (e.g., Delicious and Digg). Wikis allow users to contribute, edit and delete content (mainly textual). Book-marking applications allow groups of users to aggregate and rate content (of any type). Social media also includes blogs, content communities (e.g., BookCrossing, Flickr, Slideshare and YouTube), social networking sites (e.g., Facebook and MySpace), virtual game worlds (e.g., World of Warcraft) and virtual social worlds (e.g., Second Life).

Social media data is data that is generated by users, using social media. Based on a recent report from IDC, nearly two-thirds of the 1.2 zettabytes of data that is digitally available today is user generated content [17]. Although companies use all of the social media types mentioned above, some such as virtual world (game and social) are used more for advertising and generate a lot less data than the others. For the purposes of this research we define social media data as data generated by users on social media sites including collaborative projects, blogs, content-communities and social networking sites.

A transaction is a business-related activity in an organization that executes a part or whole of a business process. The data generated by the transaction represents that specific transaction and provides a complete understanding of that transaction. Hence, transactional data is stand-alone. Its value is self-contained and it conveys a clear meaning. For example, a purchase transaction representing a customer (say, CX) purchasing a red Spirit bicycle for \$600.00 on a specific date informs the organization (the bicycle retailer) that a customer purchased a bicycle. From this transaction, the organization also knows who purchased it, what the selling price is, date of purchase etc. The data from the transaction is self-sufficient and does not need any additional context to explain its importance and meaning. On the other hand, the customer’s (CX) tweet that he/she acquired a red Spirit bicycle for \$600.00 offers little value to the retailer. In fact, the value of this data may be dependent on why the customer tweeted this message – whether it was a status update letting friends/followers know that the customer now owns a new red Spirit or whether it was a response to someone else tweeting the fact that they bought a similar product for \$700.00, or whether it is an announcement that the Spirit is available for \$600.00! Social media data is hence context-sensitive. Its purpose is known only to the user that generated the data. Its purpose, from the view point of the retailer is indeterminate and can only be inferred (with possible error) from the context and tone. Today, understanding the “sentiment” behind social media data is the biggest challenge facing organizations. This argument assumes that the retailer can connect the customer that purchased the bicycle with that customer’s tweet – a different problem that we do not address in this research!

Transactional data has a well-defined structure<sup>28</sup>. Its semantics is unambiguous and its meaning can be inferred without error from the data and its structure. There is a well-defined domain from which the values of each data element may be extracted and the values are constrained by the domain. Social media data does not have a well-defined structure (i.e., is ambiguous or irregular, has a structure that is not useful and/or the structure is not easily identifiable because it does not conform to any known/pre-defined data model). Its meaning is context sensitive and ambiguous. The value of a data element in social media data is not constrained by a pre-defined domain nor is it restricted by any pre-defined range.

---

<sup>28</sup> Some transactional data generated by CRM systems are not “structured” the way we have defined it here. These do have some minimal structure and are referred to as semi-structured or subtly-structured data. This data, however, do have a well-defined purpose and its meaning can be inferred without error.

Transactional data is typically captured in a database. The database is based on a data model that defines the structure using which the data is captured in the database. The data model is defined based on the different current and anticipated usage requirements specified by the organization (and users within). When a database is designed to capture transactional data, the organization knows the set of current and anticipated purposes as to why the data is captured this way and knows how to interpret the data explicitly. It is difficult or impossible to use the data that has been structured in a specific way for any other purposes other than those that were used to define its structure. On the other hand, social media data is typically created with a purpose known only to the creator (the blogger, tweeter or Facebook member). The purpose with which an organization is looking at social media data created by users (who are not part of or controlled by the organization) may be very different from the purposes that the creators had in mind. Social media data generated by users must hence be re-purposed by imposing a specific structure on it. Imposing a structure is not a trivial task. When combined with the fact that each new purpose may require a new and different structure, structuring social media data can be difficult and expensive. The positive aspect is that because it has no pre-defined structure, it can be, theoretically, re-purposed in many different ways (i.e., it is “liquid”).

Traditional methods for managing data quality rely on understanding the structure and semantics of the data ([25][36]). As social media data is structurally and semantically different from transactional data, traditional techniques for managing quality may not be applicable. There is a paucity of literature in data quality management addressing the quality of social media data. Our objectives in this paper are to understand data quality in the context of social media data. Specifically, the applicability of existing quality dimensions for managing the quality of social media data and the use of social media tools and data to manage data quality of both transactional data and social media data.

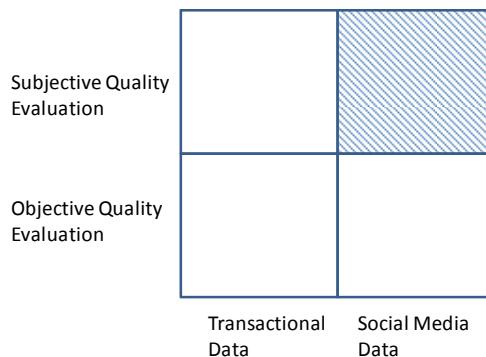
### ***Research Scope and Methodology***

Data quality is perceived as a multi-dimensional construct. Wang and Strong suggest defining data quality along multiple dimensions (such as accuracy, completeness, validity, and currency) to better reflect the concept of quality of transactional data for consumers [34]. Different metrics have been proposed for quantitatively measuring quality of transactional data along the different dimensions (e.g., [26][27]). Wang and Strong show that users view some quality dimensions as impartial - i.e., the perception of quality along these dimensions is based on the data itself, regardless of how that data is used [33]. Other dimensions are viewed as being contextual and the perception of quality along these depends on the decision context in which the data is used. Pipino et al., however, argue that the same dimension can be measured impartially and/or contextually, depending on the purpose the measurement serves [26]. As both impartial and contextual assessments contribute to the overall perception of data quality, it is important to address both when implementing data quality management solutions [9][26].

In this paper, we posit that the quality of social media data can only be assessed contextually for several reasons. First, social media data lacks a formal structure and users will have to interpret the structure based on the task for which the data is used. Since data quality assessment is dependent on the structure of the data, the quality must be gauged based on the interpreted structure. Second, the meaning of social media data is ambiguous and is interpreted based on the context the data is used. The users must assess quality based on their interpretation of data within the context in which that data is used. Third, the purpose for which the social media data was created (why was it tweeted, why was it blogged etc.) is known only to the creator, the users of that data will have to repurpose the data in the context in which the data is used. Therefore, the quality of the social media data needs to be interpreted within its context of use.

Methods that evaluate quality of social media data hence focus on a small subset of social media data,

semi-structured data. These focus on one dimension of quality, relative correctness, a proxy for accuracy. We summarize the research addressing the quality of social media data, for completeness. Link analysis and link-based methods have been shown to be successful for examining social media [28]. Ranking algorithms (such as PageRank [3] and HITS [21]) that use link-based techniques have been used to estimate the quality of question/answer portals in content-communities that allow users to post questions and other users to respond. Users are also allowed to rate questions as well as answers (e. g., Yahoo!Answers, Google Answers and Yedda). An extension of PageRank, ExpertiseRank [37], helps identify the quality of experts as well as identify experts in question/answer content-communities. Research has also studied the propagation of trust and distrust within the Epinions (<http://epinions.com>) users [14]. Su et al. (in [32]) and Jeon et al. (in [18]) have also looked at the quality of answers in question/answer portals. All of the above work treat quality as a single dimension and evaluate the quality of the answers based on length of answer and number of user-points received by each. They also use features such as fraction of best answers and the number of answers provided. This work is further extended by Agichtein et al. who include more features and evaluate the quality of the question in addition to the quality of answers [1]. Our research in this paper, like the above, examines the evaluation of the quality of social media data. Instead, we focus on how social media tools support the subjective evaluation of the data quality of social media data (see figure 1).



**Figure 1 : Research Scope**

Managing the quality of data is critical due to the importance of data as an organizational asset. As social media data is now a large part of organizational data, managing its quality is hence important. Owing to the significant differences between transactional and social media data quality management techniques that have been applied to the former may not be applicable to the latter. We examine this perspective by analyzing each dimension and evaluate its applicability to social media data. Given the nature of social media data, how can users and organizations gauge the quality of this data? Our second objective in this paper is to examine the tools offered by social media technologies to help users estimate the quality of the data generated within these social media technologies. We do so by adopting a bottom-up approach. We look at, using sample instances, how users evaluate quality of social media data and suggest ways to make this evaluation more useful, not only for evaluating quality but also for improving the quality of the user-generated social media data. As shown in figure 1, our research looks at the use of social media data and technologies to manage the quality data generated by users through the use of social media. We further develop a formal model for evaluating believability, a quality dimension that we posit will be critically important for social media data. We illustrate how this dimension may be gauged using social media tools and metrics for social media that have emerged in the recent past.

There is very little research and information available on how organizations use social media data. Furthermore, there is even less information available on how organizations manage quality of social media

data. Our objective was to understand the implications of quality dimensions for social media data and how organizations manage the quality of social media data along these specific dimensions. We were building theory through investigation. We hence adopted a methodology that would help us uncover how organizations use social media data. We relied on interpreting this data to better understand how quality dimensions are impacted by social media data and how data quality is managed using social media tools. Interviews are helpful when the researcher is trying to understand something from the subject's view point and interpret the underlying meaning based on this experience [22]. According to Kvale [22], interviews allow a subject to communicate his/her experience from his/her own perspective and in his/her own words. Interviews allow us to capture the subject's view on the research topic and interpret meaningful relations from it. We hence adopted a methodology that used interview as the research tool followed by the analyses of the interview-data to identify our findings.

To conduct these interviews, based on our understanding of the data needed, we first identified a plan for gathering the data. We also identified alternate questions that would help us probe further to get at additional details in accordance with the techniques described in [22]. We then identified a small group of subjects and conducted the interviews, either over phone or face-to-face. In all cases, with the approval of the subject, the conversations were recorded and subsequently transcribed. Each researcher independently analyzed the conversations and reached conclusions. These were then discussed by all three researchers and the conclusions were refined. The analyses of our findings from these interviews form the basis for our preliminary results on the applicability of traditional data quality dimensions for social media data, discussed next. These interviews also gave us insights, presented later on in this paper, into how organizations evaluate quality of social media data.

## **SOCIAL MEDIA DATA AND DATA QUALITY DIMENSIONS**

We present the findings from our analysis of quality dimensions and their applicability to social media data. We have described only those dimensions that have been shown as being important for practice and for organizations [23]. These dimensions are accuracy, completeness, consistency, believability, timeliness and accessibility.

**Accuracy** is defined as how correct a data value is, compared to some known baseline value [23][27][34]). It has been extensively addressed in data quality literature and is perceived as an important quality dimension for transactional data. Although literature states that accuracy is an intrinsic (or objective or context-independent) dimension [35], it has been shown that accuracy can be contextual – how accurate should the data be is determined by the context in which the data is used [26]. Accuracy is difficult to measure because the baseline value is unknown (at the time of measurement) or difficult to determine. With social media data, it is even more difficult to determine its accuracy. For instance, if a user tweets that he/she purchased a Spirit bicycle for \$600.00, there are several elements of this data that needs validation – did the user actually purchase a bike, was it a Spirit, or is \$600 the price the user paid for the product. In general, how does one accept whether the social media data is accurate? We have to rely on additional data from the social media community to infer the accuracy of the original data. Alternatively or simultaneously we could use other sources to arrive at some conclusion regarding the accuracy. In either case, it is difficult to verify the accuracy of the data and one can only gauge and subjectively arrive at some conclusion regarding the accuracy. Similar to transactional data, for social media data, accuracy is contextual. Depending on the task for which the data is to be used, users may decide how accurate they want the data to be. Hence, for social media data, we believe that accuracy is an important dimension and we need to define methods to estimate accuracy. Some methods for gauging accuracy are described in the next section.

**Completeness** is defined as the extent to which data elements are present (or included) in the data being

examined [2][26][27][31]). For transactional data, completeness is measured using three perspectives: schema completeness, column completeness and population completeness [23]. Schema defines the structure of a database. Schema completeness measures the extent to which all of the entities and attributes are present in the schema. Column completeness measures the extent to which the values of a specific attribute (or column) are present (i.e., if a column has missing values, it is considered incomplete). Population completeness measures the extent to which the population is represented in the database (e.g., in a table capturing the data on university students, if all records represent undergraduate students, the graduates are not represented). It must be evident from the above description of completeness that the measurement is based on data structure. In the case of social media data that lack structure, how does one define completeness so that it can be measured? Further, literature has shown that completeness is contextual – users perceive and measure completeness of data based on the task that data is to be used for. Given the contextual nature of social media data and its lack of structure, we do not believe that completeness is an applicable dimension to measure social media data.

**Consistency** is measured using two perspectives: value consistency and format consistency [26][34]. If the same attribute (say, customer name) in two different data sources or different parts of the same data source has different values for the same business entity (say, customer), we have inconsistent values. If the format is different (say, customer name as a single string in one case and split into last and first in another), we have inconsistency in format. Consistency is a context-independent measure and is intrinsic to the data. In transactional data, consistency is measurable because the values are extracted from a well-defined domain and because the data has structure. With social media data that is devoid of structure or formally defined value-domains, it is difficult or impossible to measure consistency. Particularly, social media allows the use of informal (but accepted) acronyms, but, does not insist on their use. Hence value and format consistencies will be present, but, are difficult to gauge without the use of software that can parse the “social media language”.

**Timeliness** is another context-dependent dimension of data quality. It is defined in literature as the extent to which data is up-to-date for use in the task or context that the data is to be used for [2][23][34]. Timeliness is important for transactional data because there could be a significant time lapse between the time the data was created (or captured) and the time the data is used. Further, in systems where data was manually captured and then digitized, it is important to understand the time elapsed between data capture and access to data. Hence, timeliness is considered to be a very important quality dimension for transactional data. With social media data, we argue that timeliness is even more important. In social media every data that is generated is time-stamped (with date and time). Data is captured and disseminated instantaneously. Further, social media data typically describes real-time events or actions. The data content can hence change rapidly, even within very small time intervals. Finally, time is a very important characteristic of context and since use of social media data is context-sensitive, timeliness will continue to be a critical dimension for measuring the quality of social media data.

**Accessibility** dimension measures the ease of attainability of data [12]. With transactional data, some data may be difficult to access due to a variety of reasons including privacy/security restrictions, sensitive nature of the data or difficulty with obtaining or capturing the data. With transactional data, the importance of this dimension diminished with the advent of mobile and wireless technologies that made access to data easier and quicker. By nature, social media data is not private, does not come with the same security restrictions that transactional data does and, the technology supporting social media makes access a breeze. This research does not examine the privacy issues surrounding social media data. Privacy is an important aspect of data and has not been treated as a *quality* dimension. We have hence not examined privacy-related issues of social media data in this paper. We agree that privacy is an important issue and, we have examined accessibility assuming that privacy is protected when referring to social

media data as easily accessible.

**Relevance** is yet another data quality dimension that measures the extent to which the data is relevant to the context. It is a context-dependent dimension of data quality and is estimated by the user based on the data and the context in which the data is used [12][34]. This dimension has not been examined closely by data quality literature, with respect to transactional data. However, there is no formal method proposed to measure relevance except for user assigned weights/scores. With social media data, we believe that relevance might take on a significant role as a quality dimension. Users are often asked whether they found the comment/blog “useful” - relevance is typically subsumed and is not explicitly assessed by “usefulness”.

**Believability** is a context-dependent dimension of data quality and is defined as the extent to which data is regarded as true and credible [34]. Literature observes that believability of data is determined by three factors: credibility of source, whether the data conforms to some internal or common-sense standard and the age of the data [34]. If the source of data is reputable or well-known, the data tends to be more believable. If the data is within a range of known or accepted values, the data is more believable. Finally, the older the data the less believable it tends to be because the more recent the data tends to be more relevant to the context that the data is used in. Research in data quality has not examined the believability dimension to the same extent as some of the other dimensions (e.g., accuracy, completeness and timeliness), giving the perception that believability is not as important. However, with social media data, believability might be a very important dimension. As the data may be generated by anyone, credibility of source becomes critical in gauging believability.

Quality Dimensions	Implications for Social Media Data
Accuracy	<ul style="list-style-type: none"> <li>- It is an important dimension in the context of social media data.</li> <li>- Needs to be gauged contextually.</li> <li>- Multiple different sources (of social media and traditional data) may be used to gauge/confirm accuracy.</li> <li>- Clear methods for managing accuracy of social media data are needed.</li> </ul>
Completeness	<ul style="list-style-type: none"> <li>- May be irrelevant in the context of social media data.</li> <li>- As there is no a priori structure, it is impossible to determine what is missing – assessing completeness is not possible.</li> </ul>
Consistency	<ul style="list-style-type: none"> <li>- May be difficult to gauge in social media data</li> <li>- There are no defined norms for representing data in social media.</li> </ul>
Believability	<ul style="list-style-type: none"> <li>- May be a critical data quality dimension for social media data</li> <li>- As source of data is often unknown, credibility of source may be a way to measure believability</li> <li>- Range of values is unspecified, hence it cannot be used to gauge believability (if the range is known and the data is within this range, the data is may be more believable)</li> </ul>
Timeliness	<ul style="list-style-type: none"> <li>- May continue to be a critical data quality dimension for social media data</li> <li>- Social media data tends to represent real-time events and/or opinions – both change with time.</li> </ul>
Accessibility	<ul style="list-style-type: none"> <li>- May not be as important for social media data as it is for transactional data</li> <li>- Social media tools are designed to support easy access to data!</li> </ul>
Relevance	<ul style="list-style-type: none"> <li>- May be a key dimension to assess quality of social media data</li> <li>- Is contextual and must be assessed by users</li> <li>- Today's tools assess “usefulness”, which does not help us understand whether the user found the data relevant.</li> </ul>

Table 1: Quality Dimensions and Implications for Managing Quality of Social Media Data

A summary of the above discussion is presented in table 1. It is not evident that all existing dimensions of data quality are applicable to social media data. While some are still applicable, it is difficult to measure these dimensions using the same measurement methods and instruments proposed for transactional data. Yet other dimensions may be irrelevant to deal with social media data. However, there are some dimensions such as relevance and believability that did not receive much attention in managing quality of transactional data, that we believe have the potential to be significantly important for managing quality of social media data. We examine one of these dimensions, believability, in more detail in the next section.

## EVALUATING QUALITY OF SOCIAL MEDIA DATA

Given the amount of social media data generated by users and the multiple different social media technologies (such as Facebook, Twitter and MySpace) that exist, how can an organization use social media data and technologies to manage quality? We next discuss how social media data and technologies can be used to manage data quality along the different dimensions of data quality described earlier.

To determine the accuracy of a data value, it is necessary to compare the data value to a baseline or a known correct value. Accuracy is difficult to measure because this baseline value is often unknown or indeterminable at the time of measurement. Hence accuracy is estimated by using statistical methods (see [25]) using historical data. In some cases, historical data is unavailable or is not useful [29]. In such instances, social media data and technologies can be used to obtain estimates of baseline data. One way is through a variation of crowd-sourcing – a way to outsource a task to a large undefined group of people, or a crowd [16]. Large organizations resort to this model to obtain estimates of data that is otherwise difficult to obtain [5] by using internal prediction markets (internal implies the markets that involve employee participation – the “crowd” is limited to the employees of the organization and hence, we refer to this as a variation of crowd sourcing). The authors argue that these models can provide insight into how organizations process information. The authors state that the prediction markets provide employees with incentives for truthful revelation and can capture changes in opinion at a much higher frequency than surveys. Following this model, organizations can use social media to solicit opinions, from recognized domain experts to obtain estimates of baseline data. This solution is inexpensive but it is important to offer some incentive in order to obtain genuine responses.

Social media technologies have helped improve timeliness of data. Organizations that listen to the social media have access to data instantaneously. However, more than the social media, it is the recent advances in mobile and wireless technologies that have addressed the problem of timeliness with transactional data [13]. Such technologies have ensured that there is no time lag between capture and dissemination of data and have significantly reduced data capture errors. Social media, per se, has not contributed significantly to managing timeliness of *transactional* data in organizations.

Social media can be used to create a proxy score for both the data and the source that generates that data. For example, if we enter the conversation within discussion groups on Amazon.com or Salesforce.com, applications are evaluated or questions are answered by a community of users. Each answer or assessment is rated by a reader on a 5 or 7 point scale on its usefulness to them. This method constantly validates the data and keeps it “refreshed”. If there are issues that can be corrected, either the source or a member of the community offers the corrections/additions. This improves the accuracy and timeliness of the data.

While the general idea of reviews and rating of experts is helpful, it is important to tune them to the data quality attribute we are trying to measure. The typical measure for gathering user reviews is based on “how useful is this data for you” and its variations. While “usefulness” is important, a response of “yes, it is useful” or “no, it is not useful” does not offer any additional insights into what should be done to improve the quality of the data evaluated. If the users were allowed to rate the usefulness based on data

quality dimensions, the data aggregator (the custodian) can gain insights into what should be done to improve the quality of the data.

Who is allowed to evaluate the data is also important – do we open it to employees, or our partners or the world at large. Take the example of Jigsaw that was recently acquired by Salesforce.com. Given the high mobility of knowledge workers, Jigsaw started with the premise that it was next to impossible for a company to keep the sales database complete. As opposed to fighting this trend with each organization, Jigsaw opened the database to the community and awarded points to each user that made changes to the database. Users could redeem these points within Jigsaw by freely accessing the data within Jigsaw, commensurate with their editing contributions. By opening it up to the community of users and by offering an incentive mechanism, Jigsaw was able to improve the completeness of the data within their system. By restricting the access to its community of users, Jigsaw solicited contributions from a knowledgeable and valid set of users, ensuring that the contributions were relevant. By allowing the community to award points to the users that made changes, Jigsaw was able to police the contributions.

Similar ideas have been applied in the governance of Wikipedia. The reliability of each article is improved by the many eyes that read and edit it. With the log traces of each edit being maintained, it is easy to find out when and who made any of the edits. The broader community is able to maintain a huge corpus of information much better than a closed group of editors.

A community of users can also be used to improve the reliability and accuracy of data. Take the example of the contest hosted by DARPA to locate the accurate position of balloons in the United States. DARPA created a contest that promised \$40,000 to the person(s) that could locate ten balloons that were randomly distributed in the US. The winning team made it into a contest by promising rewards to each person that helped in locating the balloons. The team was able to locate all 10 balloons in less than 48 hours.

When Netflix wanted to improve the quality of their recommendation data, they decided to involve the community and award a prize for the best quality achievable. They first created a goal or target for performance. This was followed by the release of the dataset against which to assess the performance. The community created small teams to tackle the problem and in a couple of years was able to improve the quality of the prediction data for NetFlix. While it is not evident what quality dimensions were targeted, it is clear that a community of connected users can improve the quality of even transactional data within an organization.

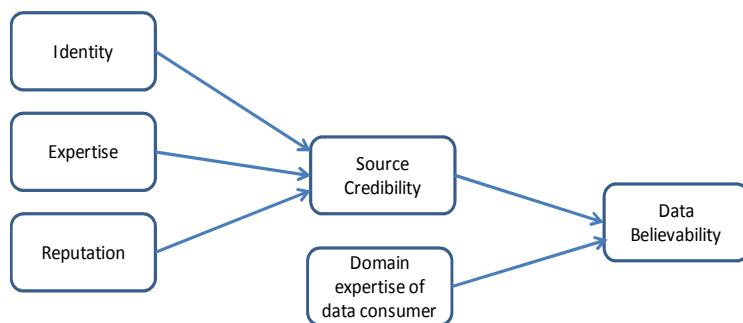
When Google decided to improve the quality of their project data and the data on the individual contributions of employees, they decided to publish both data on their intranet. Each employee enters their own data on project assignments and performance. Since this data is being viewed by all employees, inaccuracies are fixed immediately. Moreover, employees get assigned to their future projects based on their proclaimed availability and past performance. This creates checks and balances on the quality of information. In the context of data quality management, Total Quality Management (TQM), applied successfully by all manufacturing firms (e.g., Toyota) refers to managing quality at source. By assigning the responsibility of ensuring quality to the role/individual that generates the data and, by having the community of peers view and correct the data, Google was able to collect accurate and complete data on its projects and contributions.

### ***Believability and Social Media Data***

Social media data and technologies can also help improve the believability of organizational data. Organizations use data from both internal and external sources. The primary issue with believability of data is the credibility of the source. Data from a source that is more credible and/or better known is considered

more believable [12][34]). Literature addresses this issue using the terms data provenance or data lineage (e.g., [24]). Metadata describing the source is provided along with the data to allow users to gauge the credibility of the data source, when needed [36]. We believe that social media can be used to address the issue of source credibility. Recent insights reveal that 80% of people in the US gain trust about products and product-brands through Facebook. A company-page on Facebook is considered the biggest source for gaining trust [36]. The second factor in believability is the domain expertise of the data consumer. Crowd-sourcing may be used to inform and/or confirm the range when users are unsure of the acceptable range of values.

As discussed above, there are a variety of tools and measurement scores that can support the evaluation of data believability of social media data. We propose the model shown in figure 2 to suggest a method for gauging data believability.



**Figure 2: A model for assessing data believability**

Data believability is the extent to which a data consumer determines the data to be true and credible. Research has stated that source credibility is a key factor in determining believability. In addition, the domain expertise of the user also plays a role in determining believability. We posit that source credibility in the context of social media may be assessed using three constructs – identity, expertise and reputation. Each of these may be estimated using social media tools. When a data consumer is evaluating the believability of some social media data, the first construct is to gauge the identity of the data provider, the individual/organization that provided this data. Research has stated that knowing the identity of the source is a big part of gauging source credibility [23][36]). To establish identity of a data provider, data consumer can refer to the provider's profile information on **LinkedIn**. These profiles contain both historical and current information. Some of these can be verified using references given by other members of the community. In many instances, the referee may be a person from the data consumer's own network. There are two key issues with respect to gauging the identity of the provider. What if the provider's identity is unknown? We have assumed that if you have a profile on **LinkedIn** or on **Twitter/Facebook**, identity, in some form, is known. The bigger question is what if the provider has several “*avatars*” on the different social media platforms and we are unable to reconcile these different identities. While we do not have a concrete answer to this question (we know that several companies such as Acxiom are working solving this question from a social media marketing perspective), we believe that if a provider wants to be recognized as an “expert”, he/she will have a clearly identifiable profile as it is in their best interests. If identity is not determinable, then the data consumer will not trust that source whose provider is unidentifiable!

The second piece to gauging source credibility is to measure the level/degree of expertise of the data provider. It is not sufficient to know the identity, it is important to understand how knowledgeable the provider is, in the specific domain/area/topic. The degree of expertise can be evaluated by looking at the community's assessment of a person's expertise. These communities grant scores to experts based on the

number of questions they answered and the quality of their response. (Evaluating the quality of the question and its responses has been addressed by prior research – see section 2.2). Many applications exist to verify expertise of a provider. Some applications rely on self-reported content to evaluate the provider. In the self-reported category we have **LinkedIn** and **Branchout**. Others like **StackExchange** and **Smarterer** use the community to assign you grades for your skills. If a person is an expert programmer in Java or a supply chain expert, communities within **StackExchange** assign scores and badges for the person's actions within these communities. These scores go much beyond your knowledge of the topic and your capabilities within that domain/are. The scores are also based on your accessibility, interaction style and helpfulness. Some sites like **Identified** (see figure 3 for a sample) are able to generate a total score for your expertise, in the context of employability, based on peer ratings of various institutions that you have been affiliated to in your career.

The third construct is the provider's online reputation. Online reputation looks at how influential the provider is, within the social media arena. In addition to knowing who the provider is (identity), the domain expertise of the provider (expertise), reputation offers insights into how connected the provider is and to what extent the provider has influenced the his/her community in the social media. A reputation score can also be computed using applications like **Klout**. **Klout** measures an individual's influence on social media. It uses input from **Twitter**, **Facebook**, blogs, **Foursquare** and other applications to compute an influence score that ranges between 1 and 100. The higher the score, the wider and stronger is the individual's sphere of influence. Influence scores are computing based on how many people the data provider reaches or connects with, how much the data provider influenced them, and how influential they (the people the data provider influenced) are. Clearly, if the data provider can make an impact on "influential" members, it adds creditability to the data provider. Much of the input that goes into the computation of the **Klout** score is unstructured. However, the outcome is a measure of a person's influence (see figure 4 for a sample report).

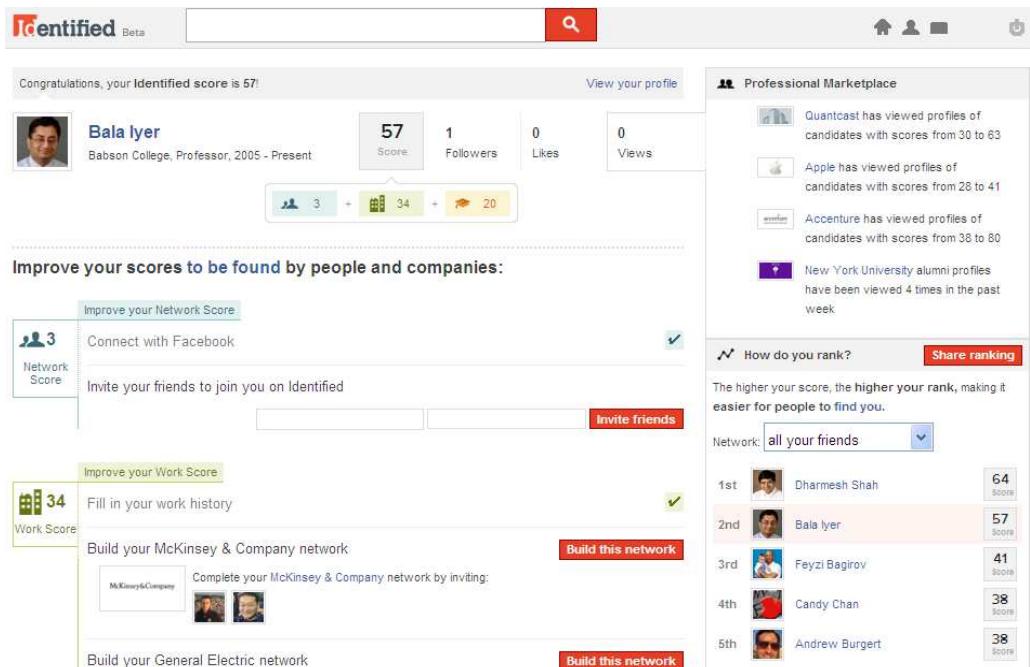


Figure 3: Expertise score on *Identified*

Property	Impact
Twitter	136,410 followers
OnStartups blog	30.000+ subscribers
OnStartups LinkedIn	270, 657+
OnStartups Answers	11,490+ users, 25,000+answers
Klout	60+

Figure 4: Overall Influence and Reputation report

In addition to a general influence score, **Klout** can identify a provider's influence on specific topic areas such as social media, search optimization and even data quality. Currently, a single place to see all these measurement scores together does not exist. However, one can aggregate this data by visiting these sites individually. When one reads blog posts or reviews on a certain topic, it is quite difficult to verify the reputation and expertise of the writer directly on the site. If we can connect the user identity to the sites that report on reputation and influence, we can use that information to make our own assessment of the believability of the data in the blog post/review.

We believe that the above tools and measurements that exist within the world of social media can help a user quantitatively gauge source credibility. With this assessment of source credibility, and combining it with a self-assessment of the data consumer's domain-expertise, the data consumer can subjectively evaluate the data believability.

Based on the above scenarios that describe how organizations gauge quality of social media data, we can observe the following:

- Communities of users are critical in assisting with several aspects of data quality. Communities are used to increase accuracy of data and keep data “refreshed”, thereby, increase timeliness. Further, communities of users are help gauge the credibility of the data source.
- The responder quality, as evaluated by the community, is important in evaluating the credibility of the data source. By allowing the community to rate the responder, users can benefit from the communities’ opinion of the responder. This way, one can form an opinion about the quality of the responder.
- Allowing the community to rate the quality of the data based on “usefulness” is not sufficient because it does not offer any insights to improve the quality of the data. By breaking “usefulness” down into specific quality dimensions (such as “is the data accurate?”, “is the data complete?”, “is the data believable?” etc.), the data custodian can get a better sense of what is wrong with the data in its current form and thus identify methods to improve its quality along one/more specific dimensions. Further, users must also be allowed to weight individual dimensions in terms of its importance to them. Based on this determination, they should assign points to the current state of the attribute along the dimensions.

We are in the process of collecting data on how decision-makers gauge data believability of social media

data using identity, degree of expertise and reputation (all describing the data provider). We are further examining and how this assessment, in turn, impacts the data consumer's perceived usefulness of the data and the impact on decision performance.

## CONCLUSION

In this paper, we have presented our observations based on our preliminary study of social media data and its impact on data quality. We first addressed the differences between social media data and transactional data. We then mapped the quality dimensions applied to manage quality of transactional data onto social media data to examine their applicability. Our mapping highlighted the fact that while some of the quality dimensions are still applicable to social media data, others are not due to the nature of social media data. It also led us to conclude that some dimensions such as believability and relevance will gain in stature as important quality dimensions for social media data. Additional work is required to identify new dimensions that may fit the social media data better. We also examined how quality of social media data may be managed. Our examination revealed that there are a number of measurements provided by social media tools that can be leveraged to manage quality of social media data. We described some of these measurements and proposed a model for evaluating data believability of social media data. We described the measurements and social media tools that can be used to evaluate data believability. We believe that we have presented a first step towards gaining a better understanding of how social media data can impact data quality and interesting ways to measure quality of social media data.

## REFERENCES

- [1] Agichtein, E., Castillo, C., Donato, D., Gionis, A. and Mishe, G. (2008), Finding High-Quality Content in Social Media, *WSDM 2008*, Palo Alto, CA, USA.
- [2] Ballou, D., Wang, R. Y., Pazer, H., and Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality, *Management Science*, 44(4), 462-484
- [3] Brin, S. and Page, L. The Anatomy of a Large-Scale, Hypertextual Web Search Engine, obtained in August 2011 from <http://infolab.stanford.edu/~backrub/google.html>.
- [4] Chengalur-Smith, I., Ballou, D. P. and Pazer, H. L. (1999) The Impact of Data Quality Informationon Decision Making: An Exploratory Study, *IEEE Transactions on Knowledge and Data Engineering* 11 (6).
- [5] Cowgill, B., Wolfers, J. and Zitzewitz, E. (2009). Using Prediction Markets To Track Information Flows: Evidence From Google, obtained in August 2011 from <http://www.bocowgill.com/GooglePredictionMarketPaper.pdf>
- [6] DeLone, W. H. and McLean, E. R. (1992). Information systems success: the quest for the dependent variable, *Information Systems Research*, 3(1), 60-95.
- [7] Eckerson, W. W. (2002). *Data Quality and the Bottom Line*, The Data Warehousing Institute, Seattle, WA.
- [8] English, L. (1999). *Improving Data Warehouse and Business Information Quality*. John Wiley & Sons, NY.
- [9] Even, A. and Shankaranarayanan, G. - Utility-driven assessment of Data Quality, *The DATA BASE for Advances in Information Systems* 38 (2) (2007) 76–93.
- [10] Even, A., Shankaranarayanan, G. and Berger, P. D., Economics-driven data management: an application to the design of tabular datasets, *IEEE Transactions on Knowledge and Data Engineering* 19 (6) (2007) 818–831
- [11] Fisher, C. W., Chengalur-Smith, I. and Ballou, D.P. (2003). The impact of experience and time on the use of data quality information in decision-making, *Information Systems Research*, 14(2), 170-188.
- [12] Fisher, C. W., Lauria, E., Chengalur-Smith, I. and Wang, R. Y. (2006) *Introduction to Information Quality*, Advances in Information Quality Book Series, MIT IQ Publications, Boston, MA.
- [13] Gaynor, M. and Shankaranarayanan, G. (2008) Implications of Sensors and Sensor-Networks for Data Quality Management, *International Journal of Information Quality*, Vol. 2, No. 1, pp. 75-93.

- [14] Guha, R., Kumar, R., Raghavan, P. and Tomkins, A. (2004) Propagation of Trust and Distrust, *Proceedings of the 13<sup>th</sup> International Conference on the World Wide Web (WWW'04)*, ACM Press, New York, NY.
- [15] Hernandez, M. A. and Stolfo, S. J. (1998). Real world data is dirty: data cleansing and the merge/purge problem, *Journal of Data Mining and Knowledge Discovery*, 2(1), 9-37.
- [16] Howe, J. (2006). *The Rise of Crowd-Sourcing*, Wired Magazine, obtained in August 2011 from <http://www.wired.com/wired/archive/14.06/crowds.html>
- [17] IDC Report (2010). The Digital Universe Decade – Are You Ready, ( indirect cite from <http://www.collectiveintellect.com/blog/wp-content/uploads/2011/04/IntegrationFINAL.pdf> accessed in June 2012). The original accessed at <http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm> (published May 2010, accessed June 2012).
- [18] Jeon, J., Croft, B. W., Lee, J. H. and Park, S. (2006) A framework to predict the quality of answers with non-textual features, *Proceedings of the 29<sup>th</sup> Annual ACM SIGIR Conference (SIGIR '06)*, ACM Press, New York, NY.
- [19] Kahn, B. K., Strong, D. M. and Wang, R. Y. (2002). Information quality benchmarks: product and service performance, *Communications of the ACM*, 45(4), 184-193.
- [20] Kaplan, A. M. and Haenlein, M. (2010) Users of the world, unite! The challenges and opportunities of Social Media, *Business Horizons* 53(1), Jan-Feb 2010, pp. 59-68
- [21] Kleinberg, J. M. (1999) Authoritative Sources in Hyperlinked Environment, *Journal of the ACM*, 46(5), pp. 604-632
- [22] Kvale, S. (1996) *Interviews: An Introduction to Qualitative Research Interviewing*. Sage Publications, Thousand Oaks, CA.
- [23] Lee, Y. W., Pipino, L. L., Funk, J. D. and Wang, R. Y. (2006). *Journey to Data Quality*. MIT Press, Cambridge, MA.
- [24] Madnick, S., Wang, R. Y. and Lee, . W. Overview and Framework for Data and Information Quality Research, *ACM Journal of Information and Data Quality*, vol. 1, 2009, pp. 1-22.
- [25] Morey, R. C. (1982) Estimating and improving the quality of information in the MIS, *Communications of the ACM* 25 (5) (1982) 337–342.
- [26] Pipino, L. L., Lee, Y. W. and Wang, R. Y. Data Quality Assessment, *Communications of the ACM* 45 (4) (2002).
- [27] Redman, T. C. (Ed.) (1996) *Data Quality for the Information Age*, Artech House, Boston, MA
- [28] Scott, J. P. (2000) *Social Network Analysis: A Handbook*, SAGE publications, Thousand Oaks, CA.
- [29] Shankaranarayanan, G., Ziad, M. and Wang, R. Y. (2003). Managing data quality in dynamic decision environment: an information product approach, *J. of Database Management*, 14, 14-32.
- [30] Shankaranarayanan, G. and Cai, Y. Supporting data quality management in decision making, *Decision Support Systems* 42 (1) (2006) 302–317.
- [31] Strong, D. M., Lee , Y. W. and Wang, R.Y. (1997), Data quality in context, *Communications of the ACM*, 40(5), 103-110.
- [32] Su, Q., Pavlov, D., Chow, J-H. and Baker, W. C. (2007) Internet-scale Collection of Human-Reviewed Data, *16<sup>th</sup> International Conference on the World Wide Web (WWW '07)*, ACM Press, New York, NY.
- [33] Wang, R.Y. (1998). A product perspective on total data quality management, *Communications of the ACM*, 41(2), 58-65
- [34] Wang, R. Y. and Strong D. M. (1996). Beyond accuracy: what data quality means to data consumers, *Journal of Management Information Systems*, 12(4), 5-34.
- [35] Watts, S., Shankaranarayanan, G. and Even, A. (2009) Assessing Data Quality in Context: A Cognitive Perspective, *Decision Support Systems*, 48, pp. 202-211.
- [36] Shankaranarayanan, G. and Watts, S. (2003) *A Relevant Believable Approach for Data Quality Assessment*, in the Proceedings of the International Conference on Information Quality (ICIQ), Boston, MA, U.S.A.
- [37] Webster, T. (2011) The Uneasy Relationship between Twitter and Social Media Measurement, (<http://brandsavant.com/the-uneasy-relationship-between-twitter-and-social-media-measurement>), accessed in June 2011

# **MEASURING INFORMATION QUALITY ON THE INTERNET A USER PERSPECTIVE**

(Research Paper)

**Olivier Blattmann**

University of Bern, Switzerland

[olivier.blattmann@iwi.unibe.ch](mailto:olivier.blattmann@iwi.unibe.ch)

**Patrick Kaltenrieder**

University of Bern, Switzerland

[patrick.kaltenrieder@iwi.unibe.ch](mailto:patrick.kaltenrieder@iwi.unibe.ch)

**Patrizia Haupt**

University of Bern, Switzerland

[patrizia.haupt@iwi.unibe.ch](mailto:patrizia.haupt@iwi.unibe.ch)

**Thomas Myrach**

University of Bern, Switzerland

[thomas.myrach@iwi.unibe.ch](mailto:thomas.myrach@iwi.unibe.ch)

**Abstract:** Research into information quality on the internet, in particular on websites, has become increasingly important in recent years. In this paper a research project is described in which a measurement instrument was developed that enables the information quality of websites to be determined and analyzed from the customer perspective. The measurement instrument was developed in several stages and on the basis of a methodical-theoretical approach. In a first step, previous research results and measurement instruments were systematically analyzed. In a second step, these results were adjusted and supplemented on the basis of a qualitative study. A quantitative test of the measurement instrument is planned.

**Key Words:** IQ Assessment, IQ in the Web, IQ Concepts, Metrics, Measures, and Models

## **INTRODUCTION AND BACKGROUND**

The concept of information quality (IQ) is not new. However, in recent years it has been enjoying increasing awareness in research. The work of Wang/Strong [41] is named [26] [27] [38] as the main cornerstone for this trend particularly in the English-speaking area. At the same time, conferences on IQ such as the “International Conference on Information Quality” at Massachusetts Institute of Technology (MIT) or the “German Information Quality Management Conference” of the German Society for Information and Data Quality (DGIQ) and many different national and international workshops have taken place. A few researchers have already analyzed and structured this variety of research projects on a meta-level, in an attempt to register the new research area of IQ, its roots and theoretical basis (cf. for example [10] or [26] on the importance of IQ as a separate research area).

As long ago as 1999, Wang et al. [42] wrote that there were few systematic approaches in existence for measuring IQ. A great many measurement instruments have come about in the meantime. These are most frequently intuitive, ad-hoc surveys of IQ aspects relevant from the perspective of a researcher [42]. In addition, it is possible to differentiate between three types of investigation and analysis of the multidimensional construct of IQ [25]. They may be collected empirically among information users (e.g. [37] or [41]) or, alternatively, by literature analyses of previous research projects on the subject (e.g. [3] or [22]). The final option is to focus on objectively or automatically measurable aspects of IQ (e.g. [18]).

Measurement criteria ascertained empirically among information users may contain inconsistencies, redundancies and/or omissions. This means that some of the identified aspects are dependent on one another (not orthogonally), are not generally recognized or are forgotten [35] [36]. To prevent or minimize such shortcomings, one possibility – besides the options mentioned by Lee et al. [25] – is a theoretical investigation of the construct of IQ (as conducted for example in [15] or [35]). Theory-based approaches, however, are also not completely free of shortcomings. For example, Wang/Strong [41] write that these are often better suited for optimizing the information preparation processes and less so for determining IQ from the user's perspective.

In the still relatively new discipline of IQ research, there is already an astonishing quantity and variety of measurement instruments for the many different domains (cf. for example the surveys in [8] or [24]). A majority of the research focuses on IQ in businesses, in which the information users are normally the employees. Only a small proportion deals specifically with IQ of websites, which (except in the special case of the intranet) are aimed at target groups outside companies. Thus a search for the keyword "Information Quality" in the "Business Source Premier" database (which contains the full text of over 3,600 academic journals with an economics background) since the year 2000 produces at least 94 hits (as at July 2010). However, only seven of these contain a reference to the internet. Even at the "International Conference on Information Quality", papers with an internet reference are sparsely represented. Of the 360 published papers from 13 such conferences held from 1996 to 2008, only 20 make any reference to the internet. The concept of "internet reference" was very broadly defined for this purpose and every published paper which contains terms from the internet environment (including e-commerce, online registrations, etc.) was counted. The number of articles which deal with the measurement of IQ on the internet, that have developed their own measurement approaches or have used existing known ones, is rather low in relation to the total quantity of papers published on the subject of IQ.

Nevertheless, a total of 28 academic papers were found in an intensive literature search. All of these contain a measurement instrument which is suitable for determining IQ on the internet. Five of them are generic instruments which are also suitable, according to their developers, for determining IQ on the internet (e.g. [40] or [41]). Not taken into account are papers which deal with the subject of measuring IQ on the internet (e.g. [39]) but do not contain a measurement instrument. Other instruments in the internet context which do not look at IQ from the customer perspective are likewise omitted (such as e.g. intranet-specific instruments [11]).

On three occasions, two published papers were summarized for the analyses. In the first case a published paper corresponds to a further development ([17] based directly on [41]). The two other cases are two papers published on the same subject in different publications ([5,38] and [14,15]). This leaves, as the basis for the literature analysis, a total of 25 studies which contain a measurement instrument for determining IQ on the internet [1] [3] [5,38] [7] [9] [12] [14,15] [17,42] [19] [20] [21] [22] [23] [28] [30] [31] [32] [33] [37] [40] [43] [44] [45] [48] [49].

From the analysis it emerges that 20 of the 25 studies have developed or derived the instruments on the basis of literature. Only one contains attributes that have been put together intuitively or on an ad-hoc basis [1]. A few instruments establish a theoretical reference, but only one is developed on a completely theoretical basis [14,15]. At least three studies contain their own empirical investigations for developing a suitable instrument [33] [37] [41]. What is interesting is that this is already a somewhat older instrument.

In total, ten of the publications found refer in various forms to Wang/Strong [41]. One research project translated the measurement instrument into German [5,38]. Others used some elements [19] [23] [32]

[48] or even the whole instrument [20] [21] [44] for their own studies.

Caro et al. [3] use the categories of Wang/Strong [41] for structuring their 33 identified attributes. Knight/Burn [22] use them for the description and implementation of their 20 attributes. This indicates the importance of Wang/Strong [41], Kahn/Strong/Wang [17] and all other works by these researchers for this area of research.

## RATIONALE AND PURPOSE

For information users, the importance of IQ is not only in decision-making [16]. Whether or not information users are satisfied with the quality of information provided also influences their attitude and behavior [9]. Thus the user becomes the center of interest when the requirement for IQ is being established [44]. However, so that IQ can be systematically measured and optimized for better fulfillment of customer needs, a suitable measurement instrument is essential.

The object of the work is to develop a literature-based and theoretically, methodically and empirically founded measurement instrument for determining IQ on the internet from the user's perspective. In doing so, previous research projects will be worked on systematically and methodically, and supplemented by means of a qualitative investigation. The basis of this is described in the following section.

## METHODS

Candid discussions with internet users reveal that evaluation criteria for determining the quality of websites exist on different levels of abstraction. Thus the statements "*The website should convey a pleasant shopping experience.*", "*The website should be easy to use*", "*Adequate product information should be available*" or "*I think it is important for a website to have a sitemap*" are becoming increasingly specific in their detail. So-called means-end chains [47] have been found useful for resolving such difficulties. Thus correlations of individual or multiple *concrete cues* or functions (means) on websites and the *perceptual attributes* are described. These attributes are put into groups and assembled into *dimensions*. Several dimensions produce a *higher-order abstraction*. The quality of websites from the customer's perspective is formed from several such higher-order abstractions. In a behavior-oriented perspective, the perceived quality ultimately has an influence on the attitude, behavioral intention and behavior of the customer (end) [46]. Thus the quality ultimately also influences the success of a website [4] [6].

This method enables a great many different statements from interviews or results of literature analyses on various aspects of website quality to be integrated into a measurement instrument. A means-end chain was also used as the methodical basis for the extensive study to develop the E-S-QUAL approach, an instrument for determining service quality in online purchases [34].

The functioning of means-end chains for determining IQ evaluation criteria can be explained using a winemaker's website as an example. For the sake of simplification, only the "Description of the actual wine year" (e.g. climate, quality of grapes, progress of development of the vineyard, etc.) and "Actual list of prices" (e.g. prices of different vintages) are considered as *concrete cues* for the purposes of the example. In the example, "The information contained on the web-site is up-to-date" (consisting of two concrete cues) and "The structure of the website is logical" (with only one concrete cue) are named as possible *perceptual attributes*. Several attributes can be combined into *dimensions*. In the example, the dimension "Timeliness" consists of just one attribute, whereas the "Representational consistency" dimension consists of two attributes. One or more dimensions can finally be formed into *higher-order abstractions*. All higher-order abstractions together make up the "perceived IQ" of the website.

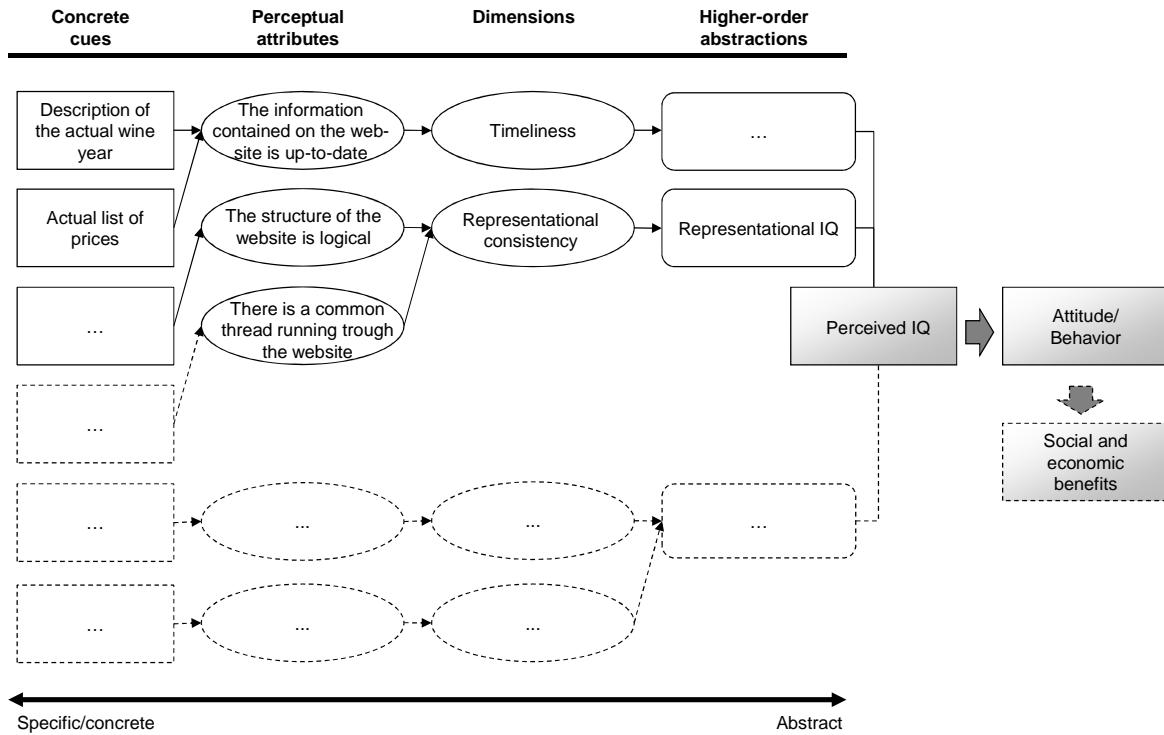


Figure 1: Example of a means-end chain.

The dimension “Representational consistency” is assigned to the higher-order abstraction “Representational IQ”. Further dimensions, such as e.g. “Concise presentation”, are possible depending on the object and purpose of an investigation. In the example, the higher-order abstraction in which the dimension “Timeliness” is categorized is left open. With such means-end chains it is possible for a complex construct such as IQ on the internet to be systematically broken down and analyzed.

If a user rates the information provided by a wine producer on a website in a specific context as being positive with regard to “Timeliness” and “Representational consistency” (and other possible dimensions), then the perceived IQ of the website positively influences the attitude and also ultimately the behavior of the customer. Negative effects can be expected if the criteria are not fulfilled. If the considerations from the research into attitude and behavior are also taken into account, then social and economic effects will also ultimately be determined by the behavior.

These methodical-theoretical principles were used in the work to systematically analyze all of the measurement instruments found. The basis provided the initial analysis of the existing 25 measurement instruments, resulting in a total of 254 concrete cues, 271 attributes, 93 dimensions and 29 higher-order abstractions. The subsequent systematic analysis enabled to condense this vast amount of dimensions and attributes to a total of 21 dimensions and 134 attributes of IQ. The challenge lied in the fact that scholars differ in their understanding of the various attributes and terms, which leads to an inconsistent use of the terms in research [3]. Moreover, some authors use specific terms as attributes while others refer to the same term as a dimension or list it as a criteria. This impedes a direct comparison of the proposed measurement instruments. Nevertheless, the same four higher-order abstractions as already used in Wang/Strong [41] were provisionally used. This now formed the basis for the analysis of the qualitative study in which the results of the literature analysis were reviewed and supplemented (cf. [2] for full doc-

umentation of the study).

An extensive qualitative research design modeled on Mayring [29] was used as the basis for the qualitative study in the Swiss wine market. Eight carefully selected users of Swiss wine producer websites were surveyed. The data collection, processing, analysis, communicative validation and interpretation took place between May 2009 and March 2010.

The interviewees were confronted in problem-centered interviews first with open and then with closed questions. It was only in the closed part of the interview that the findings and dimensions of IQ resulting from the literature analysis played an important role.

In total almost 26 hours of data material was recorded in the eight interviews and the communicative validations. On the basis of this, 185 pages – in other words, 2,304 paragraphs i.e. changes of speaker, 80,462 words or 496,016 characters (including spaces) were transcribed. For the purposes of the qualitative content analysis in MaxQDA, these texts were provided with 387 memos, i.e. sometimes relatively detailed explanations and commentaries, and coded with a total of 762 different codes. Between 157 and 185 passages of text per interviewee, totaling 1,370 overall, were coded using these codes.

Each of the total of eight interviews was initially individually coded and studied to analyze its content. A rule-based, content-structured content analysis was carried out for the evaluation. In order to check and improve the category system and the coding rules, the first three interviews were each analyzed by two researchers. Different codings were discussed and arguments for a specific allocation were balanced against one another. The coders finally each agreed on a specific coding, whereby the category system and the coding rules were purposefully improved in the first three interviews. This enabled both methodical knowledge as well as understanding of the investigation object to be enhanced. The coding system developed on the basis of the first three interviews was only expanded later if completely new categories and codings came to mind. The analysis and/or coding of the remaining interviews were carried out by a researcher working alone. A second coder was called upon to assist in cases of uncertainty or difficult codings. To ensure that the results of the interviews are consistent despite the coding system which could be easily adapted and supplemented on an ongoing basis, the interviews were analyzed again with the help of the categorization system and the coding rules before and after the communicative validation process.

When discussing a dimension of IQ, the interviewees frequently mentioned one or more attributes that belong to another dimension according to the coding system. The main reason for this is that the interviewees only learned about the dimensions during the closed part of the interview and new dimensions also resulted from the data material in the qualitative content analysis. They therefore could not know that a separate dimension exists for the perceptual attribute they mentioned. Secondly, it became clear that the understanding of concepts in theory or literature differs in certain points from that of the interviewee. A special process therefore needed to be developed for the closed interview part. If a mentioned attribute was a new attribute, it was provisionally assigned during the evaluation to the dimension in which it was mentioned by the interviewees, contrary to the coding system. If it was a known attribute which was correctly assigned elsewhere according to the coding system, a duplicate of it was prepared and this was provisionally assigned to the dimension in which it was mentioned by the interviewees. In both cases the attribute was clearly identified for further work. In the communicative validation the interviewees were confronted with the analysis and asked why they had mentioned the attribute in this particular dimension. In most cases these attributes could be assigned according to the coding system with the agreement of the interviewees. In the case of a few attributes, however, the explanation by the interviewees led to a better understanding of a dimension and to adjustments in the investigation results.

A further challenge for the analysis of the interviews are statements by interviewees which are made on different levels of abstraction. For example, the interviewees alternated during the interview between talking about concrete cues, perceptual attributes, dimensions and sometimes even higher-order abstractions and consequences of IQ. For this reason, and in order to achieve a systematic analysis, it was decided to use the means-end chain method for this too. This process enabled statements to be assigned to different levels of abstraction and to be further used accordingly. The clear definitions of the individual components of the means-end chain proved useful when delimitation was difficult. To enable the interviewees to understand the results of the study, a means-end chain was presented and explained to them in the communicative validation. So that the interviewees would not be too heavily influenced, none of the higher-order abstractions were named (in contrast to the means-end chain shown in figure 1).

The communicative validation is an essential element in quality assurance of qualitative research [13]. Consequently, a few weeks after the first interview the results of the structuring content analysis were visualized with the help of MAXMaps (MAXMaps is a component of MAXQDA) and presented to the interviewee. The aim of this process was to check the content and make any adjustments to the prepared and analyzed statements.

To permit more in-depth analysis of the closed part of the interview, i.e. the individual dimensions of IQ, a special form of communicative validation, the so-called structure-laying technique, is used [13]. It enables concepts to be structured in a form similar to the theory. A central element of it is that the interviewees carry out this structuring and graphical illustration of their statements themselves. Thus the interviewees were requested to put the IQ dimensions written on cards in three to a maximum of five groups. The aim of this process is that largely similar dimensions will be contained within the groups at the end. The interviewees also had to give each group what they felt to be a suitable name which matched the higher-order abstractions of the perceived IQ. After all interviews were completed the individual results of the structure-laying technique were compared with one another and finally with the results of the theoretical analyses.

A similar procedure was also used in earlier projects to investigate IQ. For example, in a subsidiary project of their investigations, Wang/Strong [41] asked test subjects to group and sort their identified IQ dimensions according to certain criteria. However, they called their process a sorting study because its focus was more on sorting and grouping than on the structuring and communicative validation of the dimensions.

After the analysis of the individual interviews and the adjustments from the communicative validation were completed, a summary content analysis across all the interviews was carried out using the Z rules [29]. The generalized, short paraphrases of text modules with significant content were already assigned a unique code during the content analysis (Z1 and Z2). The first and second reduction (Z3 and Z4) were carried out again by two researchers working together. The ensuing results and their interpretation are contained in the following section.

## **RESULTS**

The qualitative content analyses finally result in 5 higher-order abstractions, 20 dimensions and 100 attributes of the IQ of websites. The attributes from the qualitative content analysis exist in the form of an implementation of the respective dimension and can be used in this form as variables and items for quantitative studies. In contrast, the attributes from the literature analysis often came from just an individual word, cannot always be clearly interpreted and are not suitable for quantitative studies. A comparative analysis on the basis of attributes is therefore not possible. Instead, the respective lists of attributes were

compared analogously and searched for similarities and differences. Certain dimensions from the literature analysis contain attributes which were assigned to a new dimension during the qualitative content analysis. The analysis shows that the dimensions of the literature analysis and the empirical study correspond analogously in 15 out of the total 26 cases. Of the remaining eleven dimensions, four correspond partially (Concise presentation, Availability/Accessibility, Added value, Completeness) and two are completely different (Personalization, Security). A total of five new dimensions were added in the empirical work. For two or three (explicitly: Adequate presentation, Aesthetics; implicitly: Authenticity) there are attributes in the literature-based implementation that were assigned to other dimensions (Usability Ease of Use, Representational consistency und Traceability) in the literature analysis. The other two (Emotionality, Entertainment value) are completely independent of previous dimensions and attributes.

There are two main reasons for these differences: the first is associated with the sources on which the literature analysis is based. Even though measurement instruments were also consciously taken into account in the literature analysis for determining the IQ of websites, most measurement instruments are focused on objects other than websites. For the users of winegrowers' websites, however, other and sometimes completely new aspects of IQ are relevant. Within the dimensions that are already known, there are also other aspects in the foreground than is the case, for example, for users of (company-internal) information systems. These mainly concern the new dimensions that could only be identified in the empirical study and the two dimensions Personalization and Security (both dimensions turned out to be irrelevant in the context of winegrowers' websites). Secondly, the absolutely essential in-depth discussion to define the meaning of the individual dimensions of IQ of websites for the purposes of qualitative content analysis leads to an improved understanding of the terms and to clearer delimitation of the individual dimensions. For example, the differences in implementations in the dimensions of Concise presentation, Availability/Accessibility, Added value and Completeness can be traced primarily to improved understanding of the terms.

The following table shows the 20 dimensions and 100 attributes of IQ on the internet that were produced from the analysis. It has to be noted at this point that the study was originally conducted in Switzerland, i.e. the dimensions and their attributes were formulated in German. In order to present this study at ICIQ 2012, they have been translated into English. Yet it proves challenging to accurately account for all linguistic nuances and subtleties. It would therefore require an international study verifying the translated dimensions and attributes across different languages and countries.

<b>Dimensions</b>	<b>No.</b>	<b>Perceptual Attributes</b>
Timeliness	1	The information contained on the website is up-to-date.
	2	Information that may become obsolete, is updated.
	3	As soon as new information about the company or its products are known, they are published on the website.
	4	Upcoming events / activities are announced in advance.
Adequate presentation	5	The design of the website appears to be professional.
	6	The layout of the website is suitable for the presentation of the information.
	7	The information is presented in an original and surprising way.
	8	The information is presented in an appropriate and readable font (size and colors).
	9	Informative elements such as pictures, photos, etc. are of high resolution and quality.
	10	Various multimedia elements (text, image, audio, video, animation, etc.) are combined usefully.
Appropriate	11	The provided information is focused on the essentials.

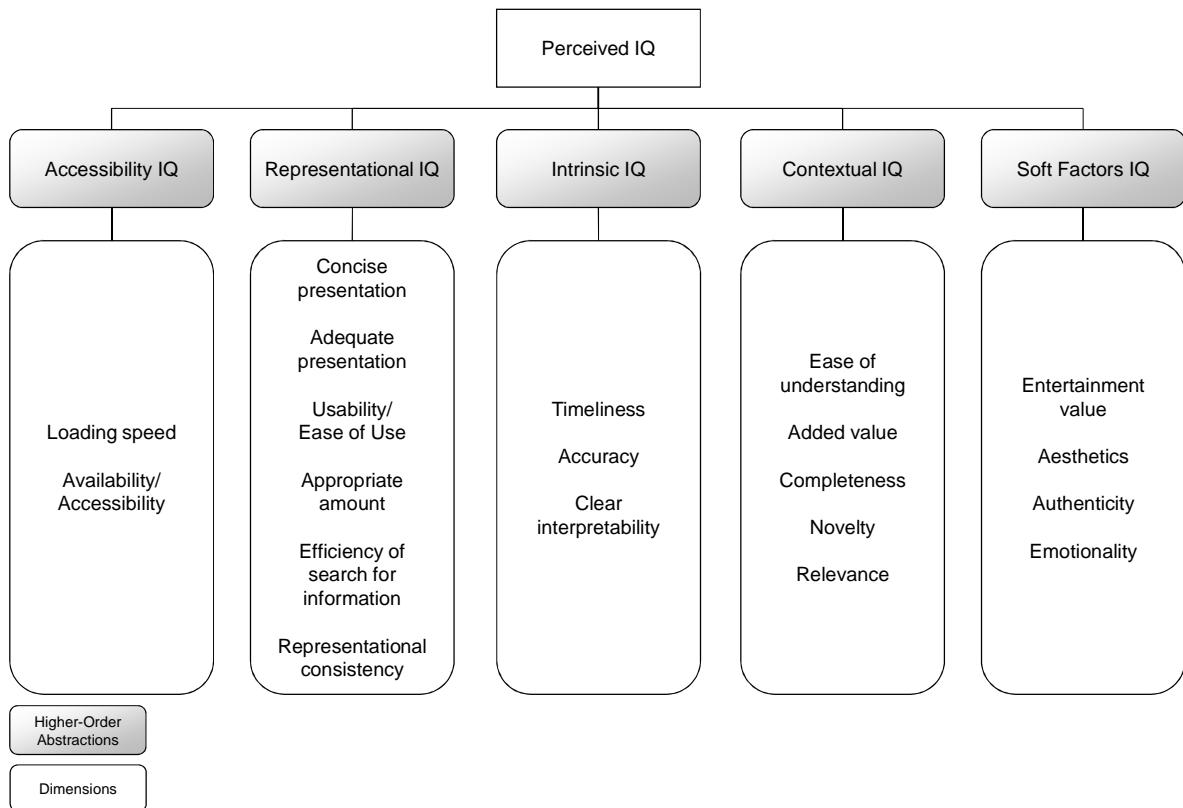
amount	12	The amount of information on every single page is appropriate.
	13	The provided information is offered at a reasonable depth.
	14	The website is not overloaded.
	15	There is not too much information on the website.
	16	There is not too little information on the website.
Aesthetics	17	The design of the website is appealing.
	18	The information is presented in an appealing way.
	19	The information is presented in a visually attractive, i.e. aesthetic way.
Authenticity	20	The elements that make up a company (i.e. making it unique and distinctive) can be perceived on the website.
	21	The identity of the company is clearly visible.
	22	The website fits the company.
	23	The website reflects the company/ the personality of the producer.
	24	The unique signature of a company, that is felt on its labels, product packaging, and all other means of communication, is also visible on the website.
	25	Emotions associated with the product/company can be perceived on the website.
Usability/ ease of use	26	The use of the website works the way I'm used to.
	27	The use of the website is simple.
	28	The use of the website is consistent.
	29	The use of the website is intuitive.
	30	The keywords provided in the navigation give an overview of the content that can be expected.
	31	The information is easy to find.
	32	The menu navigation is consistent throughout the site.
	33	The navigation is clear and understandable.
	34	The navigation between different pages/content is easy.
	35	It is always clear, where you are currently located on the website.
Efficiency of search for information	36	The effort to search for information is reasonable.
	37	The information sought is found quickly.
	38	The menu helps you find the information quickly and efficiently.
	39	Frequently requested information, i.e. the most interesting information on the website is easy to find.
	40	New information is immediately apparent.
Clear interpreta- bility	41	The meaning of the information is clear.
	42	The information contains no ambiguities.
	43	The information is unequivocal.
Concise presenta- tion	44	The design of the various pages is uniform and consistent.
	45	The amount of fonts, sizes, and colors is appropriate.
	46	There is a common thread running through the design of the pages.
Emotionality	47	The visit of the website is a "sensory experience".
	48	The information triggers positive feelings (e.g. grace, sympathy, etc.).
	49	The information is prepared and presented with care.
	50	Emotions behind the product can be perceived.
	51	Besides the design (matching color schemes, emotional imagery, etc.) the information content is not neglected (balance of emotion and information).

Availability/ Accessibility	52	Access to information is simple.
	53	The website and thus the information offered are easy to find.
	54	The website and thus the information offered are available and accessible at any time.
	55	The web pages are displayed correctly.
	56	The navigation between different content on the website is working properly.
Accuracy	57	The information on the website is free of contradictions.
	58	The information is current, that is valid.
	59	The information is error-free, i.e. true with regards to the content.
	60	The language is correct and free of grammatical and spelling errors.
	61	Translations are accurate.
Loading speed	62	Offered contents are displayed quickly (short loading time).
	63	The time needed to display the information on the website is appropriate for me.
	64	The web pages load quickly.
Added value	65	Due to the information contained on the website I save time searching for information about a company and its products.
	66	The information provided by the website facilitates my search for information about a company and its products.
	67	The information offered is beyond my expectations.
	68	The information expands my knowledge, is new to me, and improves my level of information.
	69	The information is useful for me, help me.
	70	The website also contains information, that is of real added value to me.
	71	New information, i.e. news, can be found on the website.
Novelty	72	There are always new and useful information to find.
	73	I find the information I seek on the website.
	74	The content of the website is relevant to me.
	75	The information offered meet my information needs.
Relevance	76	The information is pertinent.
	77	The structure of the website is logical.
	78	There are an appropriate number of navigation levels (main category, sub-categories, sub-sub-categories, etc.).
	79	The information can be found where I expect them to be.
Representational consistency	80	The website is clear.
	81	The website is structured similarly to what I am familiar with.
	82	The structure supports the search for information and the users' orientation.
	83	There is a common thread running through the website.
	84	It is clear where the information sought can be found.
	85	Pages with a lot of information are well structured.
	86	Pages with similar content are built/structured in the same way.
	87	The information on the website contributes to the user's entertainment.
	88	The information is presented in an entertaining way.
	89	Besides the entertainment, the information content is not neglected.

Ease of understanding	90	The form of expression is appropriate.
	91	The chosen language is understandable to a broad audience.
	92	The information is easily understandable.
	93	The information is concisely formulated.
	94	Foreign words and technical terms are avoided where possible or used efficiently where inevitable (as few as possible, as many as necessary).
	95	Information that is comprehensible only for a professional audience, is provided separately from general information (e.g. technical data about the products).
	96	Translations are understandable.
Completeness	97	The information offered is complete.
	98	The information contained on the website complete the company's overall offer of information (e.g. in addition to e-mails, newsletters, brochures, pamphlets, letters, etc.).
	99	The website contains the information that I expect.
	100	No essential information is missing.

At the level of the higher-order abstractions there are scarcely any differences between the empirical and the literature-based result. Thus only the higher-order abstraction of Soft Factors was added by the interviewees. Greater differences were to be found at the dimensions level. As already described, six dimensions were deleted and five new ones added. In addition, five dimensions were moved within the higher-order abstractions: Timeliness from Contextual to Intrinsic IQ, Clear interpretability from Representational to Intrinsic IQ, Appropriate amount from Contextual to Representational IQ, Efficiency of search for information from Contextual to Representational IQ and Usability/Ease of Use from Accessibility to Representational IQ. An overview of the higher-order abstractions and their corresponding dimensions is provided in figure 2.

For the interviewees, a piece of information is not intrinsically correct if it is not up to date (timely) and cannot be clearly interpreted. These two dimensions were therefore moved to Intrinsic IQ. It may further be concluded that Representational IQ has a different meaning on the internet than in the context of traditional information systems. For example, Usability/Ease of Use is considered by the interviewees to be an important design element of a website. It is also important that the design enables information to be searched for efficiently and prevents information overflow. Viewed in this way, Representational IQ might also be renamed “design-related” IQ in the context of the internet. The further dimensions contained in the higher-order abstraction however relate rather to the presentation of information in the narrower sense, which is why the name is not changed. The new dimensions Entertainment value, Aesthetics, Authenticity and Emotionality were assigned to the new higher-order abstraction of Soft Factors IQ.



**Figure 2: Higher-order abstractions and dimensions of IQ.**

## DISCUSSION

The implications and limitations of this work are discussed in the following two sections.

### ***Implications***

The contribution of this work for research can be evaluated from methodical, content-related (conceptual and empirical) and theoretical perspectives.

From the methodical viewpoint, a high quality was achieved for the qualitative-explorative research project and its results through the consistent orientation to qualitative effectiveness criteria. The methodical basis proved to be an essential support for the complex analyses, and especially during difficult phases of the research project. Future researchers can use this basis as the starting point for their own projects. From the content-related viewpoint, the basis for what is – from an academic and practical perspective – a complex and interesting phenomenon, was created systematically and methodically.

Thus all previously found measurement instruments and empirical studies in the field of perceived IQ on the internet were assembled in the first instance. This, according to the results of the literature analysis, constitutes the most extensive inventory of research work in German and English on the subject to date. The previous research studies for measuring perceived IQ on the internet using a methodical basis were then consolidated and structured. Only a few of the previous studies and compilations of measurements demonstrate a clear methodical basis. Even rarer in previous works is the use of self-developed empirical findings. A few even assemble their criteria in an ad-hoc way and on an intuitive basis. Using the means-end chain method, all of these aspects of perceived IQ found in the literature search were analyzed me-

thodically and were checked and supplemented by means of own empirical investigations, in which concrete cues, perceptual attributes, dimensions and higher-order abstractions of perceived IQ were determined and systematized. In addition to the methodical and content-related findings described, in the theoretical context the paper also contributes to the permeation of the phenomenon of IQ on the internet. As already mentioned above, previous studies sometimes lacked any clear reference to existing literature, methodology or theory.

### **Limitations**

The results obtained essentially document the “current state of error” of the authors of this work. This statement incorporates two core elements. Firstly, the entire process of the qualitative-explorative research was oriented toward qualitative effectiveness criteria. The work therefore endeavored to capture the phenomenon under investigation as objectively as possible. Nevertheless, it is possible – indeed even probable – that certain aspects were overlooked or overestimated because of the subjective perspectives and due to knowledge that was lacking or already available. Secondly, the findings and the measurement instrument reflect only a snapshot, and so it is possible that the measurement instrument will be developed further and revised in future investigations. A further restriction results from the fact that empirical analyses carried out during the study are limited only to the Swiss wine industry. It is clear that the results cannot be generalized without further investigations. For generalization to be possible, the analyses would need to be wider – i.e. carried out in different sectors, countries, and languages. Last but not least, the measurement instrument lacks a quantitative test with which it may be further tested and thereby improved.

## **CONCLUSION AND FURTHER RESEARCH**

As explained above, the measurement instrument is now being further developed and refined by means of quantitative studies. This will enable, for example, the attributes, dimensions and higher-order abstractions of perceived IQ to be tested. It would also make sense to reduce the number of attributes (and possibly also dimensions and higher-order abstractions) using suitable methods, in order to facilitate practical and pragmatic analyses. Furthermore, it will be fascinating to investigate the consequences of perceived IQ on the internet. Reference is also made for this purpose to eleven of the 25 previous studies, which have already carried out quantitative empirical investigations.

In this sense, this paper can be used as the basis for further research projects and publications on the phenomenon of IQ on the internet, the significance of which – owing to the increasing importance of the online channel – will continue to grow still further in the future.

## **REFERENCES**

- [1] Alexander, J. E., Tate, M. A. *Web Wisdom: How to Evaluate and Create Information Quality on the Web*. Lawrence Erlbaum Associates. Mahwah, NJ., 1999.
- [2] Blattmann, O. *iQual - Informationsqualität im Internet: Eine Analyse am Beispiel der Schweizer Weinbranche*. Südwestdeutscher Verlag für Hochschulschriften. Saarbrücken, 2012.
- [3] Caro, A., Calero, C., Caballero, I., Piattini, M. “A Proposal for a Set of Attributes Relevant for Web Portal Data Quality.” *Software Quality Journal*, 16 (4). 2008. pp. 513-542.
- [4] DeLone, W. H., McLean, E. R. “Measuring E-Commerce Success: Applying the DeLone & McLean Information Systems Success Model.” *International Journal of Electronic Commerce*, 9 (1). 2004. pp. 31-47.
- [5] Deutsche Gesellschaft für Informations- und Datenqualität e.V. *Informationsqualität - Definitionen, Dimensionen und Begriffe*. 2007. [http://www.dgiq.de/\\_data/pdf/IQ- Definition/IQ- Definitionen.pdf](http://www.dgiq.de/_data/pdf/IQ- Definition/IQ- Definitionen.pdf) [2008-06-01].
- [6] Doll, W. J., Torkzadeh, G. “The Measurement of End-User Computing Satisfaction: Theoretical and

- Methodological Issues." *MIS Quarterly*, 15 (1). 1991. pp. 5-10.
- [7] Eppler, M. J. "A Generic Framework for Information Quality in Knowledge-Intensive Processes." *International Conference on Information Quality*. Cambridge. 2001. pp. 329-346.
- [8] Eppler, M. J. *Managing Information Quality*. Springer Verlag. Berlin, Heidelberg, New York, 2006.
- [9] Eppler, M. J., Algesheimer, R., Dimpfel, M. "Quality Criteria of Content-Driven Websites and Their Influence on Customer Satisfaction and Loyalty: An Empirical Test of an Information Quality Framework." *International Conference on Information Quality*. Cambridge. 2003. pp. 108-120.
- [10] Eppler, M. J., Helfert, M., Gasser, U. "Information Quality: Organizational, Technological, and Legal Perspectives." *Studies in Communication Sciences*, 4 (2). 2004. pp. 1-15.
- [11] Eppler, M. J., Muenzenmayer, P. "Information Quality on Corporate Intranets: Conceptualization and Measurement." *International Conference on Information Quality*. 1999. pp. 162-175.
- [12] Eppler, M. J., Muenzenmayer, P. "Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology." *International Conference on Information Quality*. Cambridge. 2002. pp. 187-196.
- [13] Flick, U. *Qualitative Sozialforschung - Eine Einführung*. Rowohlt Taschenbuch Verlag. Reinbek bei Hamburg, 2007.
- [14] Gräfe, G. "Incredible Information on the Internet: Biased Information Provision and a Lack of Credibility as a Cause of Insufficient Information Quality." *International Conference on Information Quality*. Cambridge. 2003. pp. 133-146.
- [15] Gräfe, G. *Informationsqualität bei Transaktionen im Internet*. Deutscher Universitäts-Verlag, GWV Fachverlage GmbH. Wiesbaden, 2005.
- [16] Gräfe, G. "Informationsqualität in Informations- und Entscheidungsprozessen." *IS Report*, 11 (5) special IQ report No. 2. 2007. pp. 32-34.
- [17] Kahn, B. K., Strong, D. M., Wang, R. Y. "A Model for Delivering Quality Information as Product and Service." *International Conference on Information Quality*. Cambridge. 1997. pp. 80-94.
- [18] Kargar, M. J., Ramli, A. R., Ibrahim, H., Azimzadeh, F., Noor, S. B. "Assessing Quality of Information on the Web Towards a Comprehensive Framework." *Iranian Journal of Engineering Science*, 1 (1). 2007. pp. 37-49.
- [19] Katerattanakul, P., Siau, K. "Measuring Information Quality of Web sites: Development of an Instrument." *International Conference on Information Systems*. Charlotte. 1999. pp. 279-285.
- [20] Klein, B. D. "User Perceptions of Data Quality: Internet and Traditional Text Sources." *Journal of Computer Information Systems*, 41 (4). 2001. pp. 9-15.
- [21] Klein, B. D. "Internet Data Quality: Perceptions of Graduate and Undergraduate Business Students." *Journal of Business & Management*, 8 (4). 2002. pp. 425-432.
- [22] Knight, S., Burn, J. "Developing a Framework for Assessing Information Quality on the World Wide Web." *Informing Science Journal*, 8 (3). 2005. pp. 159-172.
- [23] Kopsco, D., Pipino, L., Rybolt, W. "The Assessment of Web Site Quality." *International Conference on Information Quality*. Cambridge. 2000. pp. 97-108.
- [24] Kremer, H. *Informationsmanagement*. Springer Verlag. Berlin, Heidelberg, New York, 2005.
- [25] Lee, Y. W., Strong, D. M., Kahn, B. K., Wang, R. Y. "AIMQ: A Methodology for Information Quality Assessment." *Information & Management*, 40 (2). 2002. pp. 133-146.
- [26] Levis, M., Helfert, M., Brady, M. "Information Quality Management: A Review of an Evolving Research Area." *International Conference on Information Quality*. Cambridge. 2007. pp. 1-16.
- [27] Lima, L. F. R., Maçada, A. C., Vargas, L. M. "Research into Information Quality: A Study of the State-of-the Art in IQ and its Consolidation." *International Conference on Information Quality*. Cambridge. 2006. pp. 1-16.
- [28] Matheus, A. "Web Design Quality Versus Web Information Quality." *International Conference on Information Quality*. Cambridge. 2004. pp. 89-98.
- [29] Mayring, Ph. *Qualitative Inhaltsanalyse*. Deutscher Studienverlag. Weinheim, 2003.

- [30] McLaughlin, J., Pavelka, D., McLaughlin, G. "Assessing the Integrity of Web Sites Providing Data and Information on Corporate Behavior." *J. of Education for Business*, 80 (6). 2005. pp. 333-337.
- [31] Moustakis, V. S., Litos, C., Dalivigas, A., Tsironis, L. "Website Quality Assessment Criteria." *International Conference on Information Quality*. Cambridge. 2004. pp. 59-73.
- [32] Naumann, F., Rolker, C. "Assessment Methods for Information Quality Criteria." *International Conference on Information Quality*. Cambridge. 2000. pp. 148-162.
- [33] Oliver, K. M., Wilkinson, G. L., Bennett, L. T. "Evaluating the Quality of Internet Information Sources." *The Annual Convention of the Association for the Advancement of Computing in Education*. Calgary. 1997. pp. 2-9.
- [34] Parasuraman, A., Zeithaml, V. A., Malhotra, A. "E-S-QUAL: A Multiple-Item Scale for Assessing Electronic Service Quality." *Journal of Service Research*, 7 (3). 2005. pp. 213-233.
- [35] Price, R., Shanks, G. "A Semiotic Information Quality Framework." *International Conference on Decision Support Systems*. Melbourne. 2004. pp. 658-672.
- [36] Price, R., Shanks, G. "Empirical Refinement of a Semiotic Information Quality Framework." *Hawaii International Conference on System Sciences*. Hawaii. 2005. pp. 1-10.
- [37] Rieh, S. Y. "Judgment of Information Quality and Cognitive Authority in the Web." *Journal of the American Society for Information Science and Technology*, 53 (2). 2002. pp. 145-161.
- [38] Rohweder, J. P., Kasten, G., Malzahn, D., Piro, A., Schmid, J. *Informationsqualität - Definitionen, Dimensionen und Begriffe*. In: Hildebrand, K., Gebauer, M., Hinrichs, H., Mielke, M. (Eds.) *Daten- und Informationsqualität*. Vieweg+Teubner Verlag, GWV Fachverlage GmbH. Wiesbaden, 2008. pp. 25-45.
- [39] Segev, A. "On Information Quality and the WWW Impact: A Position Paper." *International Conference on Information Quality*. Cambridge 1996. pp. 16-23.
- [40] Stvilia, B., Gasser, L., Twidale, M. B., Smith, L. C. "A Framework for Information Quality Assessment." *Journal of the American Society for Information Science and Technology*, 58 (12). 2007. pp. 1720-1733.
- [41] Wang, R. Y., Strong, D. M. "Beyond Accuracy: What Data Quality means to Data Consumers." *Journal of Management Information Systems*, 12 (4). 1996. pp. 5-34.
- [42] Wang, R. Y., Strong, D. M., Kahn, B. K., Lee, Y. M. "An Information Quality Assessment Methodology." *International Conference on Information Quality*. Cambridge. 1999. pp. 258-265.
- [43] Wang, Y. S., Tang, T. I., Tang, J. E. "An Instrument for Measuring Customer Satisfaction toward Web Sites that Market Digital Products and Services." *Journal of Electronic Commerce Research*, 2 (3). 2001. pp. 89-102.
- [44] Xu, H., Koronios, A. "Understanding Information Quality." *E-Business Journal of Computer Information Systems*, 45 (2). 2004. pp. 73-82.
- [45] Yang, Z., Cai, S., Zhou, Z., Zhou, N. "Development and Validation of an Instrument to Measure User Perceived Service Quality of Information Presenting Web Portals." *Information & Management*, 42 (4). 2005. pp. 575-589.
- [46] Zeithaml, V. A., Berry, L. L., Parasuraman, A. "The Behavioral Consequences of Service Quality." *Journal of Marketing*, 60 (2). 1996. pp. 31-46.
- [47] Zeithaml, V.A., Parasuraman, A., Malhotra, A. "A Conceptual Framework for Understanding E-Service Quality: Implications for Future Research and Managerial Practice." *Marketing Science Institute Cambridge*, Working Paper Nr. 00-115. 2000.
- [48] Zhang, P., von Dran, G.M., Blake, P., Pipithsukkunt, V. "A Comparison of the Most Important Website Features in Different Domains: An Empirical Study of User Perceptions." *Americas Conference on Information Systems*. Long Beach. 2000. pp. 1367-1372.
- [49] Zhu, X., Gauch, S. "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web." *International ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens. 2000. pp. 288-295.



## AUTHOR INDEX

<i>Last Name, First Name</i>	<i>Proceedings Page</i>	<i>Last Name, First Name</i>	<i>Proceedings Page</i>
Ayad, Sarah	70	Malik, Piyush	14
Batini, Carlo	212	Mercorio, Fabio	163
Blattmann, Olivier	326	Mezzanzanica, Mario	163
Bonner, Joseph	111	Myrach, Thomas	326
Borek, Alexander	126	Nahm, Meredith	46, 111
Boselli, Robert	163	Nargaraj, Chitra Kagathur	14
Caballero, Ismael	31, 260	O'Riain, Sean	58
Cappiello, Cinzia	31	Palmonari, Matteo	212
Caro, Angélica	31	Panse, Fabian	278
Cesarini, Mirko	163	Parlikad, Ajith Kumar	126, 193
Curry, Edward	58	Parmet, Yisrael	229
Decker, Hendrik	139	Parsons, Jeffrey	206
Even, Adir	99, 229	Pekkola, Samuli	1
Feldman, Michael	229	Pierce, Elizabeth	14
Floridi, Luciano	178	Pipino, Leo	151
Friedrich, Steffen	278	Reed, Philip L.	111
Gao, Jing	193	Ritter, Norbert	278
George, Elaine	193	Rodríguez, Alfonso	31
Guerra, César	260	Rybolt, William	151
Haider, Abrar	243	Shankaranarayanan, G.	311
Haupt, Patrizia	326	Si-Said Cherfi, Samira	70
Howard, Kit	111	Stoddard, Donna	311
Illari, Phyllis	178	Tanaka, Yasuhiro	297
Iyer, Bala	311	ul Hassan, Umair	58
Iznaga, Yonelbys	260	Vilminko-Heikkinen, Riikka	1
Kaltenrieder, Patrick	326	Viscusi, Gianluigi	212
Kodate, Akihisa	297	Wechsler, Alisa	99
Koronios, Andy	193	Wingerath, Wolfram	278
Laine, Sami	85	Woodall, Philip	126, 193
Lee, Snag Hyun	243	Yonke, C. Lwanga	14
Lukyanenko, Roman	206		

## ICIQ 2012 SPONSORS

### Platinum Sponsors



### Gold Sponsors



### Silver Sponsors

