ICEIS 2015

17th International Conference on Enterprise Information Systems

PROCEEDINGS

Volume 1

Barcelona, Spain

27 - 30 April, 2015

www.iceis.org

SPONSORED BY:
INSTICC

LOGISTICS PARTNER:
SCITEVENTS
SCIENCE AND TECHNOLOGY EVENTS

PAPERS AVAILABLE AT:
SCITEPRESS
DIGITAL LIBRARY

ICEIS 2015

Proceedings of the
17th International Conference on
Enterprise Information Systems

Volume 1

Barcelona, Spain

27 - 30 April, 2015

Sponsored by
INSTICC – Institute for Systems and Technologies of Information, Control and Communication

In Cooperation with
ACM SIGMIS – ACM Special Interest Group on Management Information Systems
ACM SIGAI – ACM Special Interest Group on Artificial Intelligence
IEICE / SWIM – IEICE Special Interest Group on Software Interprise Modelling
AEPIA – Spanish Association of Artificial Intelligence
AAAI – Association for the Advancement of Artificial Intelligence

In Collaboration with
IRC – Informatics Research Center

Edited by Slimane Hammoudi, Leszek Maciaszek and Ernest Teniente

Printed in Portugal
ISBN: 978-989-758-096-3
Depósito Legal: 390330/15

http://www.iceis.org
iceis.secretariat@insticc.org

BRIEF CONTENTS

INVITED SPEAKERS IV

ORGANIZING AND STEERING COMMITTEES V

SENIOR PROGRAM COMMITTEE VI

PROGRAM COMMITTEE VI

AUXILIARY REVIEWERS XII

SELECTED PAPERS BOOK XII

FOREWORD XIII

CONTENTS XV

INVITED SPEAKERS

George Giaglis

Athens University of Economics and Business

Greece

Witold Staniszkis

Rodan Development

Poland

Martin Mocker

MIT, USA and

Reutlingen University, Germany

ORGANIZING AND STEERING COMMITTEES

CONFERENCE CO-CHAIRS

Olivier Camp, MODESTE/ESEO, France

José Cordeiro, Polytechnic Institute of Setúbal / INSTICC, Portugal

PROGRAM CO-CHAIRS

Slimane Hammoudi, ESEO, MODESTE, France

Leszek Maciaszek, Wroclaw University of Economics, Poland and Macquarie University, Sydney, Australia

Ernest Teniente, Polytechnic University of Catalonia, Spain

PROCEEDINGS PRODUCTION

Marina Carvalho, INSTICC, Portugal

Helder Coelhas, INSTICC, Portugal

Bruno Encarnação, INSTICC, Portugal

Lúcia Gomes, INSTICC, Portugal

Ana Guerreiro, INSTICC, Portugal

Raquel Pedrosa, INSTICC, Portugal

Vitor Pedrosa, INSTICC, Portugal

Sara Santiago, INSTICC, Portugal

José Varela, INSTICC, Portugal

CD-ROM PRODUCTION

Pedro Varela, INSTICC, Portugal

GRAPHICS PRODUCTION AND WEBDESIGNER

André Lista, INSTICC, Portugal

Mara Silva, INSTICC, Portugal

SECRETARIAT

Vitor Pedrosa, INSTICC, Portugal

WEBMASTER

Susana Ribeiro, INSTICC, Portugal

SENIOR PROGRAM COMMITTEE

Balbir Barn, Middlesex University, U.K.

Senén Barro, University of Santiago de Compostela, Spain

Albert Cheng, University of Houston, U.S.A.

Jan Dietz, Delft University of Technology, The Netherlands

António Figueiredo, University of Coimbra, Portugal

Nuno Guimarães, ISCTE-University Institute of Lisbon, Portugal

Kecheng Liu, University of Reading, U.K.

Pericles Loucopoulos, Manchester University, U.K.

Yannis Manolopoulos, Aristotle University, Greece

Masao Johannes Matsumoto, Solution Research Lab, Japan

Alain Pirotte, Université catholique de Louvain, Belgium

Matthias Rauterberg, Eindhoven University of Technology, The Netherlands

Colette Rolland, Université Paris 1 Panthéon-Sorbonne, France

Abdel-Badeeh Mohamed Salem, Ain Shams University, Egypt

Bernadette Sharp, Staffordshire University, U.K.

Alexander Smirnov, SPIIRAS, Russian Federation

Ronald Stamper, Measur Ltd, U.K.

François Vernadat, European Court of Auditors, Luxembourg

PROGRAM COMMITTEE

Miguel Angel Martinez Aguilar, University of Murcia, Spain

Adeel Ahmad, Laboratoire d'Informatique Signal et Image de la Côte d'Opale, France

Rama Akkiraju, IBM Research, U.S.A.

Antonia Albani, University of St. Gallen, Switzerland

Patrick Albers, ESEO - Ecole Superieure D'Electronique de L'Ouest, France

Abdullah Alnajim, Qassim University, Saudi Arabia

Mohammad Al-Shamri, Ibb University, Yemen

Rainer Alt, University of Leipzig, Germany

Andreas S. Andreou, Cyprus University of Technology, Cyprus

Tamara Babaian, Bentley University, U.S.A.

Cecilia Baranauskas, State University of Campinas - Unicamp, Brazil

Ken Barker, University of Calgary, Canada

Rémi Bastide, ISIS - CUFR Jean-François Champollion, France

Bernhard Bauer, University of Augsburg, Germany

Lamia Hadrich Belguith, ANLP Research Group, MIRACL, University of Sfax, Tunisia

Orlando Belo, University of Minho, Portugal

Domenico Beneventano, DII - Università di Modena e Reggio Emilia, Italy

Jorge Bernardino, Polytechnic Institute of Coimbra - ISEC, Portugal

Mikael Berndtsson, University of Skovde, Sweden

Frederique Biennier, INSA Lyon, France

Sandro Bimonte, Irstea, France

Foreword

This book contains the proceedings of the 17th International Conference on Enterprise Information Systems (ICEIS 2015), which was sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC), held in cooperation with the Association for the Advancement of Artificial Intelligence (AAAI), IEICE Special Interest Group on Software Interprise Modelling (SWIM), ACM SIGMIS - ACM Special Interest Group on Management Information Systems, ACM SIGAI - ACM Special Interest Group on Artificial Intelligence and in collaboration with the Informatics Research Center (IRC). This year ICEIS was held in Barcelona, Spain from 27 - 30 April, 2015.

The purpose of the 17th International Conference on Enterprise Information Systems is to bring together researchers, engineers and practitioners from the areas of “Databases and Information Systems Integration”, “Artificial Intelligence and Decision Support Systems”, “Information Systems Analysis and Specification”, “Software Agents and Internet Computing”, “Human-Computer Interaction” and “Enterprise Architecture”, interested in the advances and business applications of information systems.

ICEIS 2015 received 327 paper submissions from 52 countries in all continents, which represents a growth in relation to last year and makes ICEIS one of the largest conferences in the World in the area of Information Systems, thus demonstrating the success and global dimension of this conference. From these, 73 papers were published and presented as full papers (30min oral presentation), 83 papers reflecting work-in-progress were accepted for short presentation and another 40 papers were presented in a poster session. These numbers, leading to a full-paper acceptance ratio of 22%, show the intention of preserving a high quality forum for the next editions of this conference.

The high number and high quality of the received papers imposed difficult choices in the selection process. To evaluate each submission, a double blind paper review was performed by the Program Committee, whose members are highly qualified researchers in ICEIS topic areas.

All presented papers will be available at the SCITEPRESS Digital Library and will be submitted for indexation by Thomson Reuters Conference Proceedings Citation Index (ISI), INSPEC, DBLP, EI (Elsevier Index) and Scopus.

Additionally, a short list of presented papers will be selected to be expanded into a forthcoming book of ICEIS 2015 Selected Papers to be published by Springer in the LNBIP Series.

The technical program of the conference included a panel and 3 invited talks delivered by internationally distinguished speakers, namely: George Giaglis (Athens University of Economics and Business, Greece), Witold Staniszkis (Rodan Development, Poland) and Martin Mocker (MIT, USA and Reutlingen University, Germany). Their participation positively contributes to reinforce the overall quality of the Conference and to provide a deeper understanding of the fields addressed by the conference.

Moreover, ICEIS 2015 had a special session on Security in Information Systems, a Doctoral Consortium on Enterprise Information Systems and 3 tutorials. We are thankful to the chairs for their dedication and hard work in organizing these events.

We sincerely thank all the authors for their submissions and participation in ICEIS 2015. Furthermore, we would like to thank all the members of the program committee and reviewers, who helped us with their expertise, dedication and time. We would also like to thank the invited speakers for their excellent contribution in sharing their knowledge and vision and the workshop/special session chairs whose collaboration with ICEIS 2015 was much appreciated. Finally, we gratefully acknowledge the professional support of the ICEIS 2015 team for all organizational processes.

We hope that all colleagues find this a fruitful and inspiring conference. We hope to contribute to the development of the Enterprise Information Systems community and look forward to having additional research results presented at the next ICEIS, to be held in Rome, Italy.

Slimane Hammoudi
ESEO, MODESTE, France

Leszek Maciaszek
Wroclaw University of Economics, Poland and Macquarie University, Sydney, Australia

Ernest Teniente
Polytechnic University of Catalonia, Spain

CONTENTS

INVITED SPEAKERS

KEYNOTE SPEAKERS

Money-over-IP - From Bitcoin to Smart Contracts and M2M Money
George Giaglis IS-5

Empowering the Knowledge Worker - End-User Software Engineering in Knowledge Management
Witold Staniszkis IS-7

Complexity in the Digital Age - How can IT Help, not Hurt
Martin Mocker IS-21

DATABASES AND INFORMATION SYSTEMS INTEGRATION

FULL PAPERS

Analysing Business-IT Alignment in the IT Outsourcing Context - A Case Study Approach
Ekaterina Sabelnikova, Claudia-Melania Chituc and Jos Trienekens 5

Applying Ensemble-based Online Learning Techniques on Crime Forecasting
Anderson José de Souza, André Pinz Borges, Heitor Murilo Gomes, Jean Paul Barddal and Fabrício Enembreck 17

A Centroid-based Approach for Hierarchical Classification
Mauri Ferrandin, Fabrício Enembreck, Julio César Nievola, Edson Emílio Scalabrin and Bráulio Coelho Ávila 25

Techniques for Effective and Efficient Fire Detection from Social Media Images
Marcos V. N. Bedo, Gustavo Blanco, Willian D. Oliveira, Mirela T. Cazzolato, Alceu F. Costa, Jose F. Rodrigues Jr., Agma J. M. Traina and Caetano Traina Jr. 34

Tracking and Tracing of Global Supply Chain Network - Case Study from a Finnish Company
Ahm Shamsuzzoha, Michael Ehrs, Richard Addo-Tengkorang and Petri Helo 46

TIDAQL - A Query Language Enabling on-Line Analytical Processing of Time Interval Data
Philipp Meisen, Diane Keng, Tobias Meisen, Marco Recchioni and Sabina Jeschke 54

Decision Guidance Analytics Language (DGAL) - Toward Reusable Knowledge Base Centric Modeling
Alexander Brodsky and Juan Luo 67

Knowledge Management Framework using Wiki-based Front-end Modules
Catarina Marques-Lucena, Carlos Agostinho, Sotiris Koussouris and João Sarraipa 79

Using Petri Nets to Enable the Simulation of Application Integration Solutions Conceptual Models
Fabricia Roos-Frantz, Manuel Binelo, Rafael Z. Frantz, Sandro Sawicki and Vitor Basto-Fernandes 87

ROBE - Knitting a Tight Hub for Shortest Path Discovery in Large Social Graphs
Lixin Fu and Jing Deng 97

Conceptual Framework of Anything Relationship Management
Jonathan Philip Knoblauch and Rebecca Bulander 108

GOTA - Using the Google Similarity Distance for OLAP Textual Aggregation
Mustapha Bouakkaz, Sabine Loudcher and Youcef Ouinten 121

PM-DB: Partition-based Multi-instance Database System for Multicore Platforms
Fang Xi, Takeshi Mishima and Haruo Yokota 128

A Hybrid Memory Data Cube Approach for High Dimension Relations
Rodrigo Rocha Silva, Celso Massaki Hirata and Joubert de Castro Lima 139

Access Prediction for Knowledge Workers in Enterprise Data Repositories
Chetan Verma, Michael Hart, Sandeep Bhatkar, Aleatha Parker-Wood and Sujit Dey 150

ERP in Healthcare
Martin Mucheleka and Raija Halonen 162

Implementing Multidimensional Data Warehouses into NoSQL
Max Chevalier, Mohammed El Malki, Arlind Kopliku, Olivier Teste and Ronan Tournier 172

What if Multiusers Wish to Reconcile Their Data?
Dayse Silveira de Almeida, Carmem Satie Hara and Cristina Dutra de Aguiar Ciferri 184

Collaborative Teaching of ERP Systems in International Context
J ̄anis Grabis, Kurt Sandkuhl and Dirk Stamer 196

On the Discovery of Explainable and Accurate Behavioral Models for Complex Lowly-structured Business Processes
Francesco Folino, Massimo Guarascio and Luigi Pontieri 206

SHORT PAPERS

Risk Management in Project of Information Systems Integration During Merger of Companies
E. Abakumov, D. Agulova and A. Volgin 221

Streaming Networks Sampling using top-K Networks
Rui Sarmento, Mário Cordeiro and João Gama 228

A Framework for Analysing Dynamic Communities in Large-scale Social Networks
Vítor Cerqueira, Márcia Oliveira and João Gama 235

On-premise ERP Organizational Post-implementation Practices - Comparison between Large Enterprises and Small and Medium-Sized Enterprises
Victoria Hasheela 243

Function-based Case Classification for Improving Business Process Mining
Yaguang Sun and Bernhard Bauer 251

Context-sensitive Indexes in RDBMS for Performance Optimization of SQL Queries in Multi-tenant/Multi-application Environments
Arjun K. Sirohi and Vidushi Sharma 259

Graph-based ETL Processes for Warehousing Statistical Open Data
Alain Berro, Imen Megdiche and Olivier Teste 271

Genetic Mapping of Diseases through Big Data Techniques
Julio Cesar Santos dos Anjos, Bruno Reckziegel Filho, Junior F. Barros, Raffael B. Schemmer, Claudio Geyer and Ursula Matte 279

An IFC4-based Middleware for Data Interoperability in Energy Building Operation
José L. Hernández, Susana Martín and César Valmaseda 287

Dimensionality Reduction for Supervised Learning in Link Prediction Problems
Antonio Pecli, Bruno Giovanini, Carla C. Pacheco, Carlos Moreira, Fernando Ferreira, Frederico Tosta, Júlio Tesolin, Marcio Vinicius Dias, Silas Filho, Maria Claudia Cavalcanti and Ronaldo Goldschmidt 295

An Empirical Study of Recommendations in OLAP Reporting Tool
Natalija Kozmina 303

Sharding by Hash Partitioning - A Database Scalability Pattern to Achieve Evenly Sharded Database Clusters
Caio H. Costa, João Vianney B. M. Filho, Paulo Henrique M. Maia and Francisco Carlos M. B. Oliveira 313

Modelspace - Cooperative Document Information Extraction in Flexible Hierarchies
Daniel Schuster, Daniel Esser, Klemens Muthmann and Alexander Schill 321

Internet of Things Applications in Production Systems
A. Boza, B. Cortés, L. Cuenca and F. Alarcón 330

Towards Data Warehouse Schema Design from Social Networks - Dynamic Discovery of Multidimensional Concepts
Rania Yangui, Ahlem Nabli and Faiez Gargouri 338

Discretization Method for the Detection of Local Extrema and Trends in Non-discrete Time Series
Konstantinos F. Xylogiannopoulos, Panagiotis Karampelas and Reda Alhajj 346

Discovering Business Models for Software Process Management - An Approach for Integrating Time and Resource Perspectives from Legacy Information Systems
C. Arevalo, I. Ramos and M. J. Escalona 353

Natural Language Processing Techniques for Document Classification in IT Benchmarking - Automated Identification of Domain Specific Terms
Matthias Pfaff and Helmut Krcmar 360

The mqr-tree for Very Large Object Sets
Wendy Osborn and Marc Moreau 367

Towards Principled Data Science Assessment - The Personal Data Science Process (PdsP)
Ismael Caballero, Laure Berti-Equille and Mario Piattini 374

Mining Big Data - Challenges and Opportunities
Zaher Al Aghbari 379

A Hybrid Genetic based Approach for Real-time Reconfigurable Scheduling of OS Tasks in Uniprocessor Embedded Systems
Ibrahim Gharbi, Hamza Gharsellaoui and Sadok Bouamama 385

CORE - A Context-based Approach for Rewriting User Queries
Antonio Mendonça, Paulo Maciel, Damires Souza and Ana Carolina Salgado 391

Graph Database Application using Neo4j - Railroad Planner Simulation
Steve Ataky Tsham Mpinda, Luis Gustavo Maschietto, Marilde Terezinha Santos Prado and Marcela Xavier Ribeiro 399

Entity Identification Problem in Big and Open Data
J. G. Enríquez, Vivian Lee, Masatomo Goto, F. J. Domínguez-Mayo and M. J. Escalona 404

Towards Principled Data Science Assessment
The Personal Data Science Process (PdsP)

Ismael Caballero[1], Laure Berti-Equille[2] and Mario Piattini[1]
[1]Institute of Technologies and Information Systems, University of Castilla-La Mancha,
Paseo de la Universidad 4, Ciudad Real, Spain
[2]Qatar Computing Research Institute, Doha, Qatar
{ismael.caballero, mario.piattini}@uclm.es, lberti@qf.org.qa

Keywords: Data Scientist, Maturity Model, Data Governance, Data Management, Data Quality Management.

Abstract: With the Unstoppable Advance of Big Data, the Role of Data Scientist Is Becoming More Important than Ever before, in This Position Paper, We Argue That Scientists Should Be Able to Acknowledge the Importance of Data Quality Management in Data Science and Rely on a Principled Methodology When Performing Tasks Related to Data Management, in Order to Quantify How Much a Data Scientist Is Able to Perform the Core of Data Management Activities We Propose the Personal Data Science Process (PdsP), Which Includes Five Staged Qualifications for Data Science Professionals, the Qualifications Are based on Two Dimensions: Personal Data Management Maturity (PDMM) and Personal Data Science Performance (PDSPf), the First One Is Defined According to Dgmr, a Data Management Maturity Model, Which Include Processes Related to the Areas of Data Management: Data Governance, Data Management, and Data Quality Management, the Second One, PDSPf, Is Grounded on PSP (Personal Software Process) and Cover the Personal Skills and Knowledge of Data Scientist When Participating in a Data Science Project, These Dimensions Will Allow to Developing a Measure of How Well a Data Scientist Can Contribute to the Success of the Organization in Terms of Performance and Skills Appraisal.

1 INTRODUCTION

Data scientist is considered the sexiest job in the world of the 21st century (Davenport and Patil, 2012). With the growth of Big Data and the increasing demand of Big Data professionals, it is becoming more important than ever to describe the required skills that any worker should have in order to perform successfully the functions related to Data Science. (Partnership, 2014) classifies these skills into two groups: Hard Skills (Subject matter expertise, math and statistics knowledge and data and technical skills) and soft skills (problem solving, storytelling, collaboration, creativity, communication and curiosity) (Partnership, 2014)

Due to the great impact that any decision taken on data could have for the organization, it is paramount to make available data with adequate levels of quality for the tasks at hand. This is not only a matter of reactively cleaning the data, but to make sure that data is adequately managed through its entire data lifecycle, from the sources up to the targets (Redman, 2013). Data quality is often understood as *fitness for use* (Strong et al., 1997). According to this, the stakeholders using the data should be able not only to specify when the data is adequate for a given task but also to specify some specific data quality requirements for all of the stages of the lifecycle of the data science projects. This does not mean that the stakeholders should be in charge for the activities related to data management. Literature describes a set of roles in charge of specific activities related to data quality management: Chief Data Officer (Yang et al. ,2014), or data stewards (Plotkin 2013), or data governors (Seiner, 2014). Data scientists should be expert in the analysis of data, but not necessarily in data management as both disciplines are core skills for data management (Pryor and Donnelly, 2009). However, this does not mean that they should not know how to better perform the data management activities in order to obtain better insights from the data.

Thus, we argue that data scientists should be integrated in the data management processes since they are the most relevant source of requirements when it comes to extract the highest value and key performance indicators from the data. Therefore, we pose that data scientists' ability of being involved in any given data project management should be somehow measured in order to let organizations know how to choose the most adequate professionals for a task.

The way in which we propose the measurement of this ability covers two dimensions: on one hand, it covers the data management expertise of the data scientist; and on the other hand, his/her efficiency when performing specific data science tasks.

To define the measure of the first dimension – named as **personal data management maturity** - we ground our proposal on an existing data management maturity model: **dgmr** (Caballero et al., 2013) which is further described in Section 2. Similarly, the measure for the second dimension – named as **personal data science performance**- is grounded on the Personal Software Process (PSP) described in (Humphrey, 2000) (1997) as "*a set of methods, forms, and scripts that show software engineers how to plan, measure, and manage their work*"

With these two dimensions, we propose the *Personal Data Scientist Process* as a structured set of process descriptions, measurements, and methods that can help data scientists to improve their personal performance and their ability to act and decide on the various steps of the lifecycle of data used to conduct the various analyses.

To the best of our knowledge, no one has ever proposed any principled methodology for data scientist (self-) appraisal. The main rationale for this proposal is to develop a universal recognition of the skills and capabilities of professional working on Data Science. In this sense, organizations can select the most valuable professionals for their projects, and data scientists can self-assess themselves against a common reference framework.

The remainder of the paper is structured as follows: Section 2 introduces the most important concepts underlying our position paper. Section 3 describes the PdsP. Section 4 introduces an illustrative example to describe the framework. Finally, Section 5 provides conclusions and future work.

2 STATE OF THE ART

In this section, we introduce the most relevant concepts to better ground the basis of our proposal.

2.1 Required Skills for Data Scientists

Data scientists usually have a strong educational background in Mathematics, Statistics, Computer Science or Engineering. They can acutely understand the business problems and needs of the industry they are working in and fluently translate their technical findings to a non-technical team, such as the Marketing or Sales departments. Along with strong technical skills in Analytics (mastering R or SAS) data scientists should have skills in Computer Science for big data management and experience with Hadoop platform, Pig or Hive and also be able to write complex SQL queries. Their goal is to arm the business and decision makers with quantified insights for their decision-making process and technical skills to tame, clean, and analyse the data appropriately.

2.2 Dgmr Framework

This section briefly introduces dgmr, which is a framework containing three main elements:

- A process reference model, describing the processes related to data management (DM), data quality management (DQM) and data governance (DG). These processes are described as ISO 12207 does. See Table 1.
- A maturity model, in which the processes previously described, has been arranged in five levels, according to what organizations should perform in order to maintain the highest levels of quality and availability for data. See Figure 1.
- An assessment methodology, which enables the assessment of the level of organizational data management maturity.

3 PdsP

The PdsP describes the concepts and processes that any data scientist should learn and follow to get a better job when analysing data. In this context, "a better job" means not only getting more reliable results but also more repeatable results in a more productive way.

The design of PdsP is based on analogous principles as PSP (Humphrey, 2000):

1. Every data scientist is different; to be most effective, data scientists should be able to plan their work and they should base their plan on their own personal data.
2. In order to improve their performance, data scientists should follow well-defined and measured processes.
3. High quality data analysis must be achieved by highly motivated and responsible data scientist.

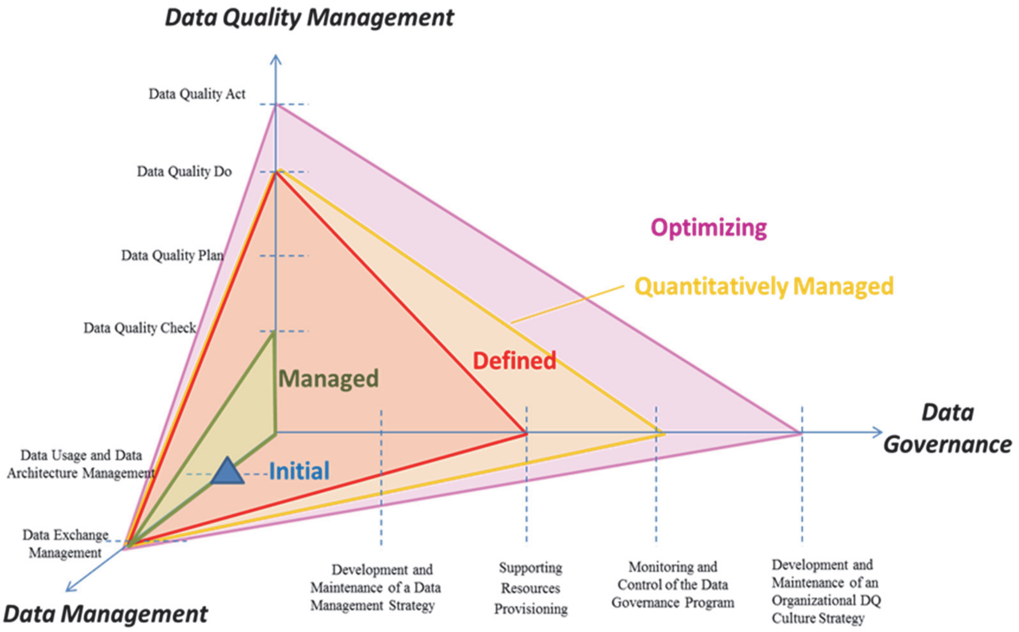
Data Quality Management
Data Quality Act
Data Quality Do
Data Quality Plan
Data Quality Check
Optimizing
Quantitatively Managed
Defined
Managed
Initial
Data Governance
Data Usage and Data Architecture Management
Data Exchange Management
Data Management
Development and Maintenance of a Data Management Strategy
Supporting Resources Provisioning
Monitoring and Control of the Data Governance Program
Development and Maintenance of an Organizational DQ Culture Strategy

Figure 1: Data Management Maturity Model associated to dgmr.

Table 1: Dgmr.

Disci	Areas
DM	Data Usage and Data Architecture Management
	Data Exchange Management
DQM	Data Quality Plan
	Data Quality Do
	Data Quality Check
	Data Quality Act
DG	Development and Maintenance of a Data Management Strategy
	Development and Maintenance of an Organizational DQ Culture Strategy
	Supporting Resources Provisioning
	Monitoring and Control of the Data Governance Program

4. Data scientists must know the context of the data used in the data science project as well as the whole data lifecycle and the data science project lifecycle.
5. It is usually cheaper to find and defects in data analysis earlier than later.
6. It is more efficient to prevent defects than to find and fix them.
7. The right way is always the fastest and cheapest way to do a job

3.1 Dimensions of PdsP

How well data scientists perform in these principles is measured by using two dimensions, namely:

- **Personal Data Management Maturity (PDMM).** It measures the extent to which a data scientist can understand the data management foundations. This data management foundations include data management (referring to data management itself), data quality and data governance. This extent is aligned to dgmr. Coherently, we define five possible values to represent the skills and knowledge of the data scientists: initial, managed, defined, quantitatively managed and optimizing. These values reflect how well a data scientist can address his/her tasks within the data lifecycle. This is important from the business and IT point of view because enable data scientist to better contextualize the data he/she is using and to understand the business value of the data for the task. See Figure 1.
- **Personal Data Science Performance (PdsPf).** It represents the extent to which a data scientist is disciplined when they conduct the various data analysis. In this sense this dimensions measures how much the data scientist follow the best-practices in order to produce high quality results in a predictable way and within schedule and budget. In order to quantify this dimension, and analogously as PSP does, we define four sets of processes containing the best practices for data analysis. The sets are listed in growing order. Each immediately higher level includes the best

practices of the previous one and some other new ones which implies a higher personal maturity for the data scientist when performing data analysis. See Table 2 for the description of the sets. It is important to realize that the performance of data scientist is measured in terms of time, size and quality measures.

Therefore, depending on the ability of the data scientist of meeting the best practices of a specific set, this will be his/her measure for PdsPf.

Table 2: Best practices for each PdsPf levels.

Set	Best practices addressed
PdsPf0	Meeting requirements of the task Performing basic measurement about the execution of the task Using Coding standards Process improvement proposal Size measurement
PdsPf1	Size Estimating Test Report Task Planning Schedule planning
PdsPf2	Data reviews Design review Design and use of templates
PdsPf3	Cyclic development

3.2 How to Determine the PdsP Level

According to the two dimensions previously explained, we identify five qualifications for professionals – PdsP level, which can be inferred by plotting these two dimensions across (please, see Figure 2).

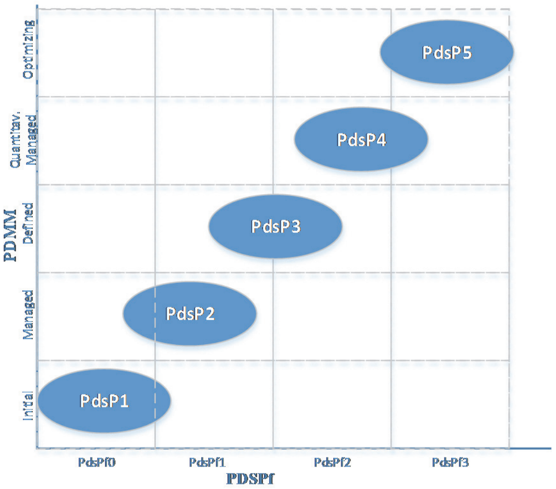

Figure 2: Values for PdsP (Caballero et al., 2013).

The five ideal levels of qualification we propose in the PdsP framework are:

- **PdsP1** corresponds to a junior profile. A data scientist of this type will be able to perform easy tasks and he/she can understand the role of the data in the data analysis and how to interact with the data architecture to recover and store data, but maybe he/she is not still ready enough to become involved in any data management task.
- **PdsP2**, in which a data scientist is able to perform basic data analysis and can slightly contribute with some requirements related to data management. He/she can understand the importance of data quality for data used in the analysis, but he/she cannot describe how to achieve it.
- **PdsP3** is an intermediate profile of somebody who can contribute efficiently to tasks related to data management and data quality management and is able to perform complex analysis on the data. He/she fully understand the meaning of data quality and how to achieve it by following data management strategy that he/she will take into account during the design of the data analysis tasks.
- **PdsP4** corresponding to data scientist profile who can perform very complex data analysis and provides requirements for optimizing the analysis and monitoring data management, data quality management, and data governance.
- **PdsP5** corresponds to a senior profile. The Data Scientist is not only able to lead data science projects, but he/she can also provide end-to-end data governance requirements.

However, it is possible to find profiles of data scientist with a very low PDMM and a very high PDSPf. Professionals in this situation are supposed to know very well the data in the organization although they are not able to face up with any managerial task.

4 CONCLUSIONS

This paper has presented the Personal Data Scientist Process, which can help Data Scientist to get more mature. In fact, the benefits that any Data Scientist could expect from the application of the framework are:

- It can help the data scientists in better developing high quality data products from the corresponding analysis
- It can better guide data scientists for personal improvement
- It gives data scientists the command over the work

they are performing

- It gives not only the necessary confidence on herself to perform a better job but also to improve the ways of doing added-value activities.

As future work, we will work on the following concerns: define and adapt the scripts provided in PSP to PDSPf, define and test equivalent metrics for size and quality in data science project. Also we want to conduct some case studies to better delimit the scope of each PdsP level and to introduce the framework to different organizations in order to validate it.

ACKNOWLEDGEMENTS

This work has been partially funded by the VILMA project (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-048-P) and by the GEODAS-BC project (Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional FEDER, TIN2012-37493-C03-01)

REFERENCES

Caballero, I., S. Serrano and M. Piattini (2013). *Technical Report*: dgmr: A Data Management Maturity Model, University of Castilla-La Mancha.

Davenport, T. H. and D. Patil (2012). "Data Scientist." *Harvard Business Review 90*: 70-76.

Humphrey, W. S. (2000). Introduction to the Personal Software Process (PSP). Pittsburg (Pennsylvania, USA), *Addison Wesley*.

Partnership, T. (2014). Big Data Analytics: Assessment of Demand for Labout and Skills 2013-2020. *T. Partnership*. London (UK), SAS UK & Ireland.

Plotkin, D. (2013). Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance, *Newnes*.

Pryor, G. and M. Donnelly (2009). "Skilling up to do data: whose role, whose responsibility, whose career?" *International Journal of Digital Curation 4*(2): 158-170.

Redman, T. (2013). Data Quality Management Past, Present, and Future: Towards a Management System for Data. *Handbook of Data Quality*. S. Sadiq, Springer: 15-40.

Seiner, R. S. (2014). "Non-invasive data governance".

Strong, D. M., Y. W. Lee and R. Y. Wang (1997). "Data quality in context." *Communications of the ACM 40*(5):103-110.

Yang, L., S. Madnick, R. Wang, F. Wang and Z. Hongyun (2014). "A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data." *MIS Quarterly Executive 13*(1): 1-13.