ICIQ 2009

The 14th International Conference on Information Quality

Co-located with the German Information Quality Management Conference (GIQMC)

November 7-8, 2009.

Hasso Plattner Institute for IT Systems Engineering

University of Potsdam, Germany

The 14th International Conference on Information Quality, 2009

Co-located with the German Information Quality Management Conference (GIQMC)

http://www.iciq2009.org

November 7-8

Hasso Plattner Institute for IT Systems Engineering

Potsdam, Germany

Contents

Welcome Message 5
Hasso Plattner Institute at a Glance 6
Program 8
Organization 10
Conference Venue 12
Keynote Talks 16
Research and Poster Papers 20
Location 35
Banquet on Saturday 46
Essential German Words and Expressions 47

Welcome Message

Felix Naumann

Dear participants of the 2009 edition of ICIQ,

welcome to the 14th International Conference on Information Quality 2009 (ICIQ'09). Since 1996 the ICIQ conferences were held annually at MIT in Cambridge. Now ICIQ has gone abroad: The Hasso Plattner Institute (HPI) in Potsdam, on the border of Berlin, is proud to host the first edition away from MIT. The HPI is privately funded and dedicated to teaching and research in the area of IT systems engineering. It is affiliated with the University of Potsdam. This year HPI was ranked among the top 4 university computer science institutes in Germany.

ICIQ provides a forum to exchange IQ knowledge and ideas and learn from each other. Practitioners and researchers will present findings and experience on topics such as IQ concepts, IQ management, case studies, best practices, cost/benefit analysis, IQ and data warehousing, IQ and e-business, policies and standards. In addition to the official program, there is plenty of opportunity for informal discussions. Highlights of this year's program include two keynote talks and a dinner talk at the conference banquet. The program features 18 talks grouped into six sessions and a poster session featuring six posters.

On behalf of all the organizers, I wish you a pleasant, insightful, and interesting stay during ICIQ at HPI. We have put together this booklet to provide helpful and interesting information about the conference, the venue, and the location at a glance. Please do not hesitate to contact me or any conference organizer for assistance and questions.

Enjoy your visit,

Felix Naumann

Hasso Plattner Institute at a Glance

A Profile of a Pioneering Institute

Hasso Plattner Institute for IT Systems Engineering (HPI) at the University of Potsdam is unique in Germany because it has been the first institute in Germany financed entirely by private funds, offering a unique degree in "IT Systems Engineering" as an alternative to conventional computer science programs.

Since 1999, HPI has awarded over than 200 Bachelor and more than 50 Master's degrees. The institute teaches more than 330 students in the design, development and control of complex IT systems. There are currently more than 50 professors, lecturers, and staff working at HPI.

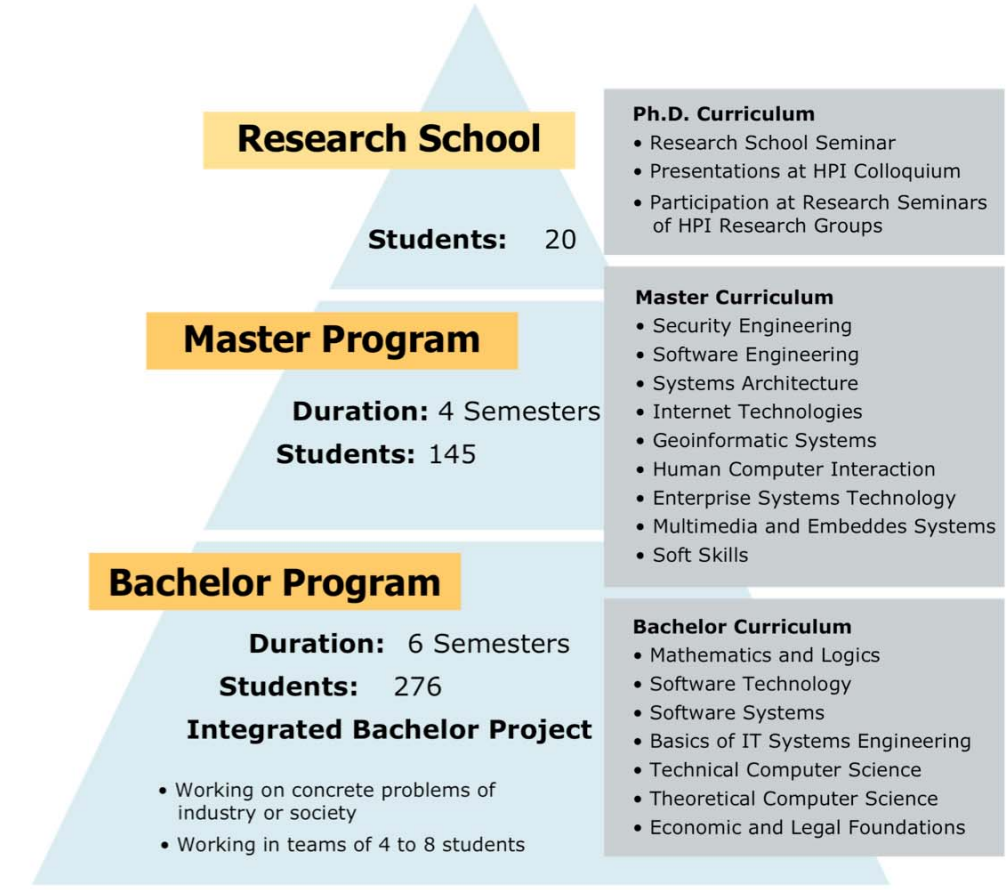
The HPI achieved excellent results in two major evaluations of IT schools in Germany, undertaken by the well-recognized Karriere magazine and the acknowledged Center for Higher Education Development (CHE). Both in the CHE-ranking of May 2006, and in the June 2006 edition of Karriere, HPI was ranked on the fourth place out of more than 100 institutions. Karriere refers to HPI as the "Shooting-Star in computer sciences".

HPI provides an outstanding student-teacher ratio, supported by latest technical and structural equipment. As a result, this pioneering Potsdam institute provides an entire new generation of scientific elites with the best study and research conditions available today.

One of the HPI's greatest priorities is the personal support of each individual student. Each year up to 80 of the best qualified applicants are accepted to the HPI where, in their first semester, they attend a seminar designed to guide them throughout their studies. A maximum of 40 students are accepted into the Master's Program each year. Every student is assigned a professor as a personal mentor; a sufficient number of computer workstations are available in the labs and seminar rooms. Tuition fees are not required.

In October 2005, the HPI started its Research School on "Service-Oriented Systems Engineering", a graduate school based on the model of the DFG (German Research Foundation) "Graduiertenkolleg" (Graduate School), which has more than 25 members.

The School of Design Thinking at the Hasso-Plattner-Institute began its program in the winter term of 2007/2008. Under the directorship of Prof. Ulrich Weinberg, this unique complementary program in Design Thinking (inventive development) is the first of its kind at a German university. Modelled on the famous d.school at Stanford University in California (USA), the one-year program in Design Thinking will enable students to develop particularly user-friendly IT-based products and services in multidisciplinary teams.



Research Groups

- Enterprise Platform and Integration Concepts
- Internet Technologies and Systems
- Operating Systems and Middleware
- Business Process Technology
- Human Computer Interaction
- Computer Graphics Systems
- System Analysis and Modeling
- Software Architecture
- Information Systems
- School of Design Thinking

A total of 71 Ph.D. Students in HPI Research Groups
20 graduates completed a Ph.D.

Status: 11.06.2009

Saturday, 7. Nov. 2009

Time \ Location	Lecture Hall (HS 2)	Lecture Hall (HS 3)
9:00-13:00	GIQMC	
13:00-14:00	Lunch	
14:00-15:00	**Keynote talk Jeff Jonas "Macro Trends in Data and Sensemaking" (HS 1)**	
15:00-15:30	Coffee break	
15:30-17:00	Parallel sessions	
	Session 1 "Master Data Management" Chair: Andy Koronios, University of Southern Australia	**Session 2 "Modeling & Metadata"** Chair: Ahmed Elmagarmid, Purdue University
	Master Data Management: Products and Research	Flexible and Generic Data Quality Metadata Exchange
	Information Management along the life cycle of data and application systems - challenges and solution approaches	Extending BPMN to Support the Modeling of Data Quality Issues
		Data Quality through Conceptual Model Quality - Reconciling Researchers and Practitioners through a Customizable Quality Model
17:00-17:15	Short break	
17:15-18:45	Parallel sessions	
	Session 3 "Techniques for Improving IQ" Chair: Marcus Gebauer, DGIQ	**Session 4 "IQ Metrics & Assessment"** Chair: Elizabeth Pierce, University of Arkansas at Little Rock
	How to Screen a Data Stream - Quality-Driven Load Shedding in Sensor Data Streams	Identification of Business Oriented Data Quality Metrics
	SOG: A Synthetic Occupancy Generator to Support Entity Resolution Instruction and Research	A Framework for Economic-driven Assessment of Data Quality Decisions
	Computing Uncertain Key Indicators from Uncertain Data	A SOA-based Data Quality Assessment Framework in a Medical Science Center
18:45-19:15	Transfer to Exploratorium	
19:15-22:00	Conference banquet (GIQMC & ICIQ)	
	Dinner Speech Richard Wang "Challenges in Advancing Information Quality"	

Sunday, 8. Nov. 2009

Time \ Location	Lecture Hall (HS 2)	Lecture Hall (HS 3)
8:00-9:00	Breakfast	
9:00-9:30	Award Ceremony (HS1)	
9:30-10:30	**Keynote talk Niels Weigel "Lean and Agile Enterprise Information Management" (HS 1)**	
10:30-11:00	Coffee break	
11:00-13:00	Parallel sessions	
	Session 5 "Poster Flash" Chair: John Talburt, University of Arkansas at Little Rock	**Session 6 "Case Studies"** Chair: Boris Otto, University of St. Gallen
	Is DQ/IQ the Quality of Information? Two Views	Information Quality Issues in the Mortgage Banking Industry
	A Meta-model for Data Quality Management Simulation	The Quality of Monitoring Data in Civil Engineering Works
	Schema Based Deduplication	
	A Multi-Dimensional Model for Assessing the Quality of Answers in Social Q&A Sites	Data Quality Evaluation in an E-Business Environment: A Survey
	Multi-Source Object Identification With Constraints	
	ExplainIE - Explaining Information Extraction Systems	Multidimensional Management and Analysis of Quality Measures for CRM Applications at EDF
13:00-14:00	Lunch	
14:00-15:30	Parallel sessions	
	Session 7 "IQ in Web 2.0" Chair: Rolf Wigand, University of Arkansas at Little Rock	**Panel**
	An Empirical Study on Criteria for Assessing Information Quality in Corporate Wikis	
	Towards Assessing Information Quality in Knowledge Management in the Enterprise 2.0	
	Web-based Affiliation Matching	
15:30-15:45	Closing Remarks (HS 1)	

Organization

General Chair

Felix Naumann: Hasso Plattner Institute for IT Systems Engineering, Germany.

Program Chairs

- **Paul Bowen**: Florida State University, USA.
- **Ahmed K. Elmagarmid**: Purdue University, USA.
- **Hubert Oesterle**: University of St Gallen, Switzerland.
- **Kai-Uwe Sattler**: Ilmenau University of Technology, Germany.

Program Committee

- **Carlo Batini:** University of Milan, Italy.
- **Laure Berti-Equille:** University of Rennes, France.
- **Mokrane Bouzeghoub:** Universite de Versailles, France.
- **Ismael Caballero:** University of Castilla-La Mancha, Spain.
- **Tiziana Cartaci:** University of Roma La Sapienza, Italy.
- **Tamraparni Dasu:** AT&T Labs, USA.
- **Helena Galhardas:** Technical University of Lisbon, Portugal.
- **Floris Geerts:** University of Edinburgh, UK.
- **Michael Gertz:** University of Heidelberg, Germany.
- **Eberhard Hechler:** IBM, Germany.
- **Markus Helfert:** Dublin City University, Ireland.
- **Theodore Johnson:** AT&T Labs, USA.
- **Barbara Klein:** University of Michigan-Dearborn, USA.
- **Andrea Maurino:** Universita di Milano Bicocca, Italy.
- **Helinä Melkas:** Lappeenranta University of Technology, Finland.
- **Paolo Missier:** University of Manchester, UK.
- **Boris Otto:** University of St. Gallen, Switzerland.
- **Mourad Ouzzani:** Purdue University, USA.
- **Elizabeth Pierce:** University of Arkansas, Little Rock, USA.
- **Leo Pipino:** University of Massachusetts Lowell, USA.
- **Erhard Rahm:** University of Leipzig, Germany.
- **Friedrich Roithmayr:** University of Linz, Austria.
- **Monica Scannapieco:** University of Roma La Sapienza, Italy.
- **Yasuki Sekiguchi:** Hokkaido University, Japan.
- **Valerie Sessions:** Charleston Southern University, USA.
- **Kai Simon:** Gartner, Germany.
- **Divesh Srivastava:** AT&T Labs, USA.
- **Dan Suciu:** University of Washington, USA.
- **Panos Vassiliadis:** University of Ioannina, Greece.
- **Vassilios Verykios:** University of Thessaly, Greece.
- **Anette Weisbecker:** Fraunhofer Institute, Germany.

ICIQ Steering Committee

- **Felix Naumann:** Hasso Plattner Institute, Germany, (Chair).
- **Bruce Davidson**: Cedars-Sinai Health System, USA.
- **Jing Gao**: University of South Australia.
- **Marcus Gebauer**: DGIQ, Germany.
- **Andy Koronios**: University of Southern Australia.
- **John Talburt**: University of Arkansas at Little Rock, USA.
- **Richard Wang**: MIT, USA.

Proceedings Chair

- **Chistoph Böhm**: Hasso Plattner Institute for IT Systems Engineering, Germany.

Publicity Chair

- **Mohammed AbuJarour**: Hasso Plattner Institute for IT Systems Engineering, Germany.

Local Arrangements Chair

- **Tobias Vogel**: Hasso Plattner Institute for IT Systems Engineering, Germany.

EXTENDING BPMN TO SUPPORT THE MODELING OF DATA QUALITY ISSUES

(Research-in-Progress)

Noelia Sánchez-Serrano, Ismael Caballero, Félix García
Alarcos Research Group, U. of Castilla La Mancha (UCLM), Spain
noelia.SanchezSerrano@gmail.com , {Ismael.Caballero, Felix.Garcia} @uclm.es

Abstract: Many companies have just realized about the increasing importance of dealing with data which have a suitable level of quality for their business operations, thus avoiding the errors and major pitfalls that could make a negative impact their performance. Various works agree that the means to achieve this goal is through the reengineering of organizational business processes, focusing attention on the most critical points, and bringing the data and information quality requirements to fruition in the most explicit manner possible. BPMN provides a suitable notation with which to represent business process issues, although it lacks the means to cope with specific data quality requirements. IP-MAP, on the other hand, does have these means, but is not as widely accepted and used to represent business process as is BPMN; in addition, BPMN is supported by different tools, which makes it more usable in practice. The main aim of this paper is to analyze BPMN's capability for represent required Data and Information Quality issues for business processes by providing certain proposals of extensions based on IP-MAP. We discuss how to extend BPMN to the support the data quality provided by IP-MAP in order to obtain a sufficiently rich notation. An example of the application of this extension is also presented.

Key Words: Data Quality, Information Quality, Business Process, BPMN, IP-MAP

INTRODUCTION

Currently, organizations need to manage a great deal of data in order to be as more competitive as possible [3]. This data must be conveniently gathered, transformed and stored according to a data model by following a series of technical processes involved in the organizational use of the information [8]. Moreover, data and information quality management processes and specific business manufacturing processes should be executed jointly. The aforementioned processes therefore represent the cycle of information within any company's Information System, and permit certain data (considered as raw data by Wang in [10]) to be transformed into the so-called data product.

However, in order to obtain information from these data products, we must be sure that the latter have appropriate quality levels which assure that they are fit for use in the tasks at hand. Many researchers agree that these levels of quality can be assured by introducing classical quality management tools and techniques when using data [1, 10]. Such management issues must be derived from the data quality user requirement specification, as parts of the respective business processes. Hence, the importance of adequately representing data quality requirements as parts of the data life cycle within a business process diagram. In this respect, Caballero et al. in [4] propose the joint provision of the technical processes for data treatment and those processes related to the management of that data's quality level. In particular, they have suggested the use of IP-MAP [9], a specific notation with which to represent *information products maps* (described in Section II) or BPMN[6]. IP-MAP permits the specification of business processes by means of a conceptual map, in which the activities corresponding to the data quality management process are properly addressed. Unfortunately, however, IP-MAP does not have all of the advantages of BPMN (an international OMG standard which is currently widely accepted and used, and which provides tools to support modeling, and which is also extensible [6]). Conversely, BPMN lacks the

specific focus on data quality that IP-MAP has.

Having observed the increasing demand for tools and techniques with which to manage data quality requirements [3], we became aware of the need to make certain proposals available to fill this gap in both communities of academics and practitioners: we believe that it is necessary to provide business analysts with the foundations and suitable tools for the treatment of specific data quality requirements, thus enabling them to model and represent their work by using a common, widely-used and easily-integrated notation. The main aim, and hence the main contribution of this research paper is, therefore, to show the work that we are carrying out to extend BPMN with those data quality issues that can already be modeled by using IP-MAP. We thus present a more complete language with which to cover this increasing need.

This paper is structured as follows: Section II presents a mapping between the concepts of IP-MAP and BPMN, and an underlying discussion of how to extend BPMN is also represented. Section III offers an example of the mapping between IP-MAP and BPMN. Finally, some conclusions and future work are presented in Section IV.

A MAPPING BETWEEN IP-MAP AND BPMN

Bearing in mind what is represented by the main constructs of IP-MAP (as shown in Table 1), and the expressive capability of BPMN, this section aims to establish a mapping between these two notations in order to make it possible to identify BPMN's lacks with regard to IP-MAP, and to thus adequately model data quality issues. To guarantee that the semantics is not lost during the mapping, we have made the comparison by matching the metadata associated with each construct in IP-MAP according to the potential equivalent in BPMN, by identifying the equivalences if they exist, or by extending the notation if they do not.

CONSTRUCTS OF IP-MAP			
CONSTRUCTS	REPRESENTS	CONSTRUCTS	REPRESENTS
<Name>	Data Source /Data Vendor /Point-of-Origin	Quality Criteria	Quality/Evaluation/Check
<Process Identifier>	Process	<Current / New System Names>	Information System Boundary
<Storage Name>	Data/Information Storage	<Current / New Org/dept names>	Organizational Boundary
<Criteria>	Decision	<Name>	Data Sink/Consumer Block/Point-of-Destination

Table 1. Constructs of IP-MAP.

Thanks to the extension capabilities of BPMN [6], as an first approach, we have added a new star-shaped symbol called "DQDim" (which stands for *Data Quality Dimensions*) that represents the data quality dimensions proposed in [5]. With this new symbol, we want to cover the lack of data quality issues in BPMN in order to extend the BPMN notation used by many business analysts. Thus, BPMN will also support the data quality issues required in business processes. This new star symbol can modify some BPMN elements by providing them with the quality that they lacks.

Now, we shall briefly discuss each construct of IP-MAP with a brief explanation, and we shall introduce the mapping of its metadata to BPMN symbols and a small visual example of the translation. For these, we use small examples, and lastly, we propose pieces of the detailed example of the IP-MAP diagram shown by Pierce in [7]. The following section shows the referred mapping.

Data Source Block

This construct aims to represent the source of each initial raw (input) data, providing data for the information manufacturing process. In BPMN, it could be mapped to a *pool*, a *lane* and a *start message event*. Table 2 summarizes how to maintain the required metadata for this construct in the mapping, and Table 3 shows an example of the translation of a data source block in a BPMN notation.

METADATA ASSOCIATED WITH DATA SOURCE BLOCK	
Data Source Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name is equivalent to the name of the BPMN *pool* element.
<Department / Role>	This is represented by the *pool* inside the *lane*.
<location>	Could be specified in the description of the *pool/lane*.
<Business Process>	This description of the set of rules associated with the raw data is represented by a *start event*.
<Composition>	Explained in the description of the *start event.*
<Base System>	Specified within the description of the *start event*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 2. Metadata associated with a Data Source Block.

IP-MAP			**BPMN**		
Block	Blocks metadata	**Example**	Blocks metadata	Translation	Translation with DQDim
Alumni notify office of change	Name/Type	Alumni/(DS1)	Alumni are student participants and correspond to a lane. /2pool+1lane+1 start message event		
	Departament /Role	Former BSU/ students	Name of Pool / Name of Lane		
	Location	Depends on the individual	In the description of the lane.		
	Business Process	Alumni send in their data changes	Name of the start event		
	Composed of	Name, Address, and other Status Changes in a free format	In the description of the start event		
	Base System	By mail, phone, in-person or through electronic correspondence	In the description of the start event		
	Quality Issues	Typos or misinformation sometimes occur	Represented by the dimensions "accuracy" and "completeness" through "DQDim".		

Table 3. Example of mapping of a Data Source Block.

Data Process Block

This block is used to represent the changes to be made to the raw data during its processing or the transformation according to the business requirements. The mapping to BPMN is directly related to the *task* construct. Table 4 summarizes how to maintain the required metadata for this construct in the mapping, and Table 5 shows an example of the translation of a data process block in a BPMN notation.

METADATA ASSOCIATED WITH A PROCESS BLOCK	
Process Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the BPMN *task* element.
<Department / Role>	Represented by the *pool* inside the *lane*
<location>	Could be specified in the description of the *pool/lane*.
<Business Process>	This description of the set of rules associated with the raw data is specified in the description of the *task*.
<Composition>	Explained in the description of the *task*.
<Base System>	Specified in the description of the *task*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 4. Metadata associated with a Data Process Block.

IP-MAP			**BPMN**		
Block	Blocks metadata	**Example**	Blocks metadata	Translation	Translation with DQDim
Update Alumni Database with Name Changes	Name/Type	Update Alumni Database with Name Changes /(P3N)	Name of the task / 1 pool + 1lane+ 1task		
	Departament/ Role	BSU Alumni Affairs /Secretary	Name of the pool / Name of the lane		
	Location	204 Breezedale Hall	In the description of the lane.		
	Business Process	Secretary loads CD1N (name updates) into Alumni Database	In the description of the task		
	Composed of	Input: CD1N	In the description of the task		
	Base System	Visual Basic Application	In the description of the task.		
	Quality Issues	Only 10% of alumni self-report name changes	Represented by the dimension "currentness" through "DQDim".		

Table 5. Example of mapping of a Data Process Block.

Data Storage Block

This block represents the storage of data items in files or databases that can be made available for future processing. It could be represented in BPMN by a *lane/pool* (which represents the storage system). Table 6 summarizes how to maintain the required metadata for this construct in the mapping, and Table 7 shows an example of the translation of a data storage block in a BPMN notation.

METADATA ASSOCIATED WITH DATA STORAGE BLOCK	
Data Storage Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the BPMN *pool* element.
<Department / Role>	This is represented by the *pool* inside the lane.
<location>	Could be specified in the description of the *pool/lane*.
<Business Process>	This description of the set of rules associated with the raw data is specified in the description of the *pool/lane*.
<Composition>	Specified inside the description of the *pool/lane*.
<Base System>	Specified inside the description of the *pool/lane*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 6. Metadata associated with a Data Storage Block.

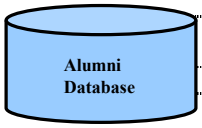
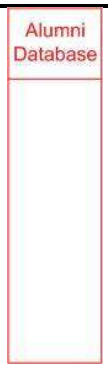
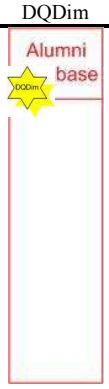
IP-MAP		**Example**	**BPMN**		
Block	Blocks metadata		Blocks metadata	Translation	Translation with DQDim
	Name/Type	Alumni Database /(STO2)	Name of the pool / 1 pool		
	Departament/ Role	BSU Alumni Affairs /Secretary	In the description of the pool.		
	Location	204 Breezedale Hall	In the description of the pool.		
	Business Process	Secretary maintains data on alumni	In the description of the pool.		
	Composed of	Alumni Mailing Information, Other Alumni Data. See data dictionary for a complete list of fields	In the description of the pool.		
	Base System	Oracle Database	In the description of the pool.		
	Quality Issues	Data becomes obsolete as it sits in the database. See component description for obsolescence rates.	Represented by the dimension "currentness" through "DQDim".		

Table 7. Example of mapping of a Data Storage Block.

Decision Block

Sometimes, depending on the value of the data, it may be necessary to direct the data to a different set of blocks for further processing. In such cases, a decision block is used to evaluate the different conditions of incoming data. Decision Block is directly mapped to the *complex gateway* construct, which is represented in a *pool/lane*. Table 8 summarizes how to maintain the required metadata for this construct in the mapping, and Table 9 shows an example of the translation of a decision block in a BPMN notation.

METADATA ASSOCIATED WITH A DECISION BLOCK	
Decision Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadat in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the BPMN *complex gateway*.
<Department / Role>	This is represented by the *pool* which contains the *complex gateway*.
<location>	Could be specified in the description of the *pool/lane* which contains the *complex gateway*.
<Business Process>	Specified in the description of the *complex gateway*
<Composition>	Specified in the description of of the *complex gateway*.
<Base System>	Specified in the description of the *complex gateway*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 8. Metadata associated with a Decision Block.

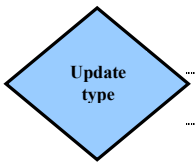
IP-MAP		Example	BPMN		
Block	Blocks metadata		Blocks metadata	Translation	Translation with DQDim
	Name/Type	Update Type/(D1)	Name of the Complex Gateway / 1 pool +1lane +1complex gateway		
	Departament/Role	BSU Alumni Affairs /Secretary	Name of the pool/ Name of the lane		
	Location	204 Breezedale Hall	In the description of the lane.		
	Business Process	Secretary reviews and sorts various types of update requests into separate folders. (RD1N, RD1A, RD1D, RD1O)	In the description of the gateway.		
	Composed of	Input: RD1 – Correspondence from alumni (either by mail, e-mail, phone, or in-person)	In the description of the gateway.		
	Base System	Paper-based system.	In the description of the gateway.		
	Quality Issues	-	In this example, the decision block has no quality issues. But if it had any quality issues, they would be represented by a "DQDim"		

Table 9. Example of mapping of a Decision Block.

Data Quality Block

This construct represents the checks for data quality on those data items required to produce a "defect-free" IP. It has no equivalence in BPMN: This is the major deficiency of BPMN if we wish to represent a full map made in IP-MAP. However, thanks to the fact that BPMN permits the addition of marks and new symbols to its current elements of representation, our idea is to extend BPMN by adding the new symbol "DQDim" which allows this data quality control to be represented through an evaluation of a series of quality dimensions.

Information System Boundary Block

This construct reflects the changes to the data items as they move from one information system to another. It could be represented in BPMN by a *task* in a *pool/lane*. This *task* only carries out the translation of the kind of data that the information system boundary block represents. For example: the translation from paper to electronic form. Table 10 summarizes how to maintain the required metadata for this construct in the mapping, and Table 11 shows an example of the translation of an information system boundary block in a BPMN notation.

METADATA ASSOCIATED WITH AN INFORMATION SYSTEM BOUNDARY BLOCK	
Information System Boundary Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the BPMN *task*.
<Department / Role>	This is represented by the *pool* inside the *lane* which contains the *task*.
<location>	Could be specified in the description of the *pool/lane* which contains the *task*.
<Business Process>	Specified in the description of the *task*.
<Composition>	Specified in the description of the *task*.
<Base System>	Specified in the description of the *task*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 10. Metadata associated with an Information System Boundary Block.

IP-MAP		**Example**	**BPMN**		
Block	Blocks metadata		Blocks metadata	Translation	Translation with DQDim
Convert Name Changes to Electronic Form	Name/Type	Convert name change to electronic form /(SB1)	Name of the task / 1 pool +1lane +1task		
	Departament/Role	BSU Alumni Affairs /Secretary	Name of the pool/ Name of the lane		
	Location	204 Breezedale Hall	In the description of the lane.		
	Business Process	Secretary records update request in an online update file	In the description of the task.		
	Composed of	Input: RD1N	In the description of the task.		
	Base System	Visual Basic Application	In the description of the task.		
	Quality Issues	Secretary may commit a typo when recording	Represented by the dimension "accuracy" through "DQDim".		

Table 11. Example of mapping of an Information System Boundary Block.

Organizational Boundary

This block represents the exchange of data items from one business unit to another. It is therefore used to specify the movement of the data across departmental or organizational boundaries. This construct could be represented in BPMN by a *message flow* and a *data object* associated with this. This *message flow* goes from one *pool* to another. Table 12 summarizes how to maintain the required metadata for this construct in the mapping, and Table 13 shows an example of the translation of an organizational boundary block in a BPMN notation.

METADATA ASSOCIATED WITH AN ORGANIZATIONAL BOUNDARY	
Information System Boundary Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	It does not appear in the diagram.
<Department / Role>	This is represented by a *pool/lane* of origin and a destination *pool/lane*.
<location>	Could be specified in the description of the *pool/lane* of origin and in the description of the destination *pool/lane*.
<Business Process>	This description of the set of associated rules is specified in the description of the *data object* associated with a *flow* that goes from the *pool* of origin to the destination *pool*.
<Composition>	Could be specified in the description of the *data object* associated with a *flow* that goes from the *pool* of origin to the destination *pool*.
<Base System>	Could be specified in the description of the *data object* associated with a *flow* that goes from the *pool* of origin to the destination *pool*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 12. Metadata associated with an Organizational Boundary.

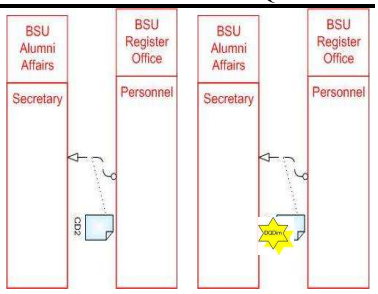
IP-MAP		Example	BPMN		
Block	Blocks metadata		Blocks metadata	Translation	Translation with DQDim
Transfer from Registrar Office to Alumni Affairs Office	Name/Type	Transfer from Registrar Office to Alumni Affairs Office /(BB1)	Does not appear in the diagram / 1 pool +1lane+ 1pool+ 1lane+ 1flow+ 1data object		
	Departament/Role	BSU Alumni Affairs /Secretary	Name of the pool/ Name of the lane (of origin), Name of the pool/ Name of the lane (of destination)		
	Location	222 Carpenter Hall, 204 Breezedale Hall	In the description of the lane (origin), In the description of the lane (destination)		
	Business Process	Registrar Office (Personnel) with help from the BSU IT Staff to extract Alumni Affairs after each semester.	In the description of the data object.		
	Composed of	Input: CD2	In the description of the data object.		
	Base System	File is sent via FTP.	In the description of the data object.		
	Quality Issues	-	In this example, the organizational boundary block has no quality issues. But if it had any quality issues, they would be represented by a "DQDim"		

Table 13. Example of mapping of an Organizational Boundary.

Customer (output) Block

This construct represents the consumer of the final data element obtained in the process. The mapping to BPMN is realized with a *task* and an *end event*. Both are in a *pool/lane*. Table 14 summarizes how to maintain the required metadata for this construct in the mapping, and Table 15 shows an example of the translation to a customer (output) block in a BPMN notation.

METADATA ASSOCIATED WITH A CUSTOMER BLOCK	
Customer Block = {<Name>, <Department/Role>, < location>, <Business Process>, <Composition>, <Base System>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the BPMN *pool/lane*.
<Department / Role>	This is represented by the *pool* in the *lane* that it belongs to.
<location>	Could be specified in the description of the *pool/lane*.
<Business Process>	Specified in the description of the last *task*. This last *task* is just before the *end event*.
<Composition>	Specified in the description of the last *task*.
<Base System>	Specified in the description of the *task*. This description states the kind of system (paper or electronic) in which the final data element is stored.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 14. Metadata associated with a Customer Block.

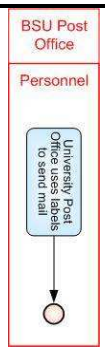
IP-MAP		Example	BPMN		
Block	Blocks metadata		Blocks metadata	Translation	Translation with DQDim
University Post office uses labels to send mail	Name/Type	University Post Office /(CB1)	Name of the task / 1 pool +1lane+ 1pool+ 1task+ 1flow+ 1 end event		
	Departament/Role	BSU Post Office /Personnel	Name of the pool/Name of the lane		
	Location	111 Sullivan Hall	In the description of the lane		
	Business Process	Mailing labels are used to address alumni publications	In the description of the task.		
	Composed of	IP1	In the description of the task.		
	Base System	Set of paper labels	In the description of the task.		
	Quality Issues	-	In this example, the customer block has no quality issues. But if it had any quality issues, they would be represented by a "DQDim".		

Table 15. Example of mapping of a Customer Block.

Raw and Component Data

Raw data could be defined as the data items that enter the information manufacturing system at the beginning, and component data could be described as the association of certain different data items anywhere in the information manufacturing system. They could be represented by a message/sequence flow with an associated data object. Table 16 summarizes how to maintain the required metadata for this construct in the mapping, and Table 17 shows an example of the translation of raw/component data in a BPMN notation.

METADATA ASSOCIATED WITH RAW AND COMPONENT DATA	
Raw and Component Data = {<Name>, <Data element>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the *data object* associated with the BPMN *flow*.
<Data elements >	Specified in the description of the *data object*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 16. Metadata associated with a Raw and a Component Data.

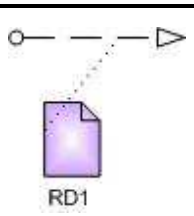
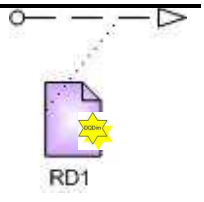
IP-MAP		Example	BPMN		
Flow	Flow metadata		Blocks metadata	Translation	Translation with DQDim
RD1	Name / Type	RD1 /(Raw)	Name of the data object / 1 message flow + 1 data object		
	Data elements	Update information in free format (phone, mail, e-mail, in person)	In the description of the data object.		
	Quality Issues	A typo may infrequently occur.	Represented by the dimension "accuracy" through "DQDim".. See Table		

Table 17. Example of mapping of Raw and Component Data.

Information Product

The information product can be defined as the final data item that leaves the information manufacturing system. This element could be represented in BPMN by a *message/sequence flow* with the last associated *data object* that appears in the business process diagram. It is usually associated with the last *task* just

before the *end event*. Table 18 summarizes how to maintain the required metadata for this construct in the mapping, and Table 19 shows an example of the translation of an information product in a BPMN notation.

METADATA ASSOCIATED WITH AN INFORMATION PRODUCT	
Information Product = {<Name>, <Data element>, <Quality Issues>}	
Metadata in IP-MAP	Proposal for Representation using BPMN
<Name>	The name matches the name of the *data object* associated with a BPMN *flow*.
<Data elements >	Specified in the description of the *data object*.
<Quality Issues>	Represented by the new symbol "DQDim".

Table 18. Metadata associated with an Information Product.

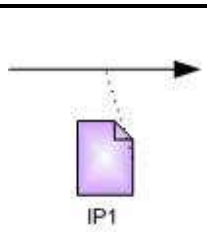
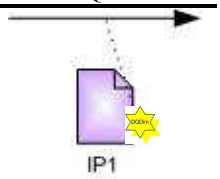
IP-MAP			**BPMN**		
Flow	Flow metadata	**Example**	Blocks metadata	Translation	Translation with DQDim
IP1	Name / Type	IP1 /(Information Product)	Name of the data object / 1 message flow + 1 data object		
	Data elements	Name, Address, City, State, Zip Code in U.S. Postal Service Standard Address Format	In the description of the data object.		
	Quality Issues	-	In this example, the information product has no quality issues. But if it had any quality issues, they would be represented by a "DQDim"		

Table 19. Example of mapping of Information Product.

AN EXAMPLE ILLUSTRATING THE USE OF BPMN IN DATA QUALITY

This section presents an example of how to use the extended BPMN notation to represent a business process with its corresponding data quality requirement issues. The example is adapted from the one borrowed from [7], in which Pierce describes a fictitious data quality campaign in the department of the Office of Alumni Affairs at school called Big State University (BSU), see Figure 1. In this department, Pierce has identified mailing labels as an important information product, but it is supposed that the 'staff' has realized that incorrect mailing labels are a problem for alumni affairs. BSU has therefore decided to create an IP-MAP diagram to represent the manufacturing process of its alumni mailing labels. The process is briefly described as follows: by the end of each semester, the data regarding graduating seniors is sent from the BSU's student database to the Alumni Affairs' alumni database. In order to maintain this data updated, Alumni Affairs advises alumni to inform them of any corrections made to name, address…by phone, e-mail, in person… Also, every quarter, Alumni Affairs sends a list of its mailing labels to a Change of Address Service. This department has a master list of addresses, so the department carries out the comparison of that list against the address of the alumni and then informs whether those addresses have changed. Moreover, once a year, BSU sends its active mailing list to an Obituary Service. The Obituary Service also maintains a master list that it compares with the BSU's list to obtain alumni who are now dead. Finally, the department of Alumni Affairs creates a list of mailing labels by following a determinate format for the labels that will be pasted onto the outgoing publication when BSU needs to send out an alumni publication. These labels are sent to the University Post Office for mailing.

Figure 2 shows the translation we have made using BPMN for the working example expressed by using IP-MAP and shown in Figure 1. Firstly, we followed the aforementioned described mechanisms to translate each construct block of the Figure 1 to the corresponding blocks of BPMN. Then, we joined all these blocks to obtain Figure 2. Moreover, more details of certain subtasks are specified, such as the

process to extract alumni data from Student Database to Alumni Database (Table 20), the process to check alumni data with the Change of Address Service (Table 21), and the process to check alumni data with the Obituary Service (Table 22). Figure 2, Table 21 and Table 22 therefore show the result of a mapping BPMN diagram and how the quality issues have been extended by adding the aforementioned star symbol in that elements that correspond to an IP-MAP construct that has quality issues in the description of its metadata. How this is done was illustrated above with an example of mapping to each construct of IP-MAP. Finally, elements of BPMN that have the star symbol have their meaning modified by quality issues.

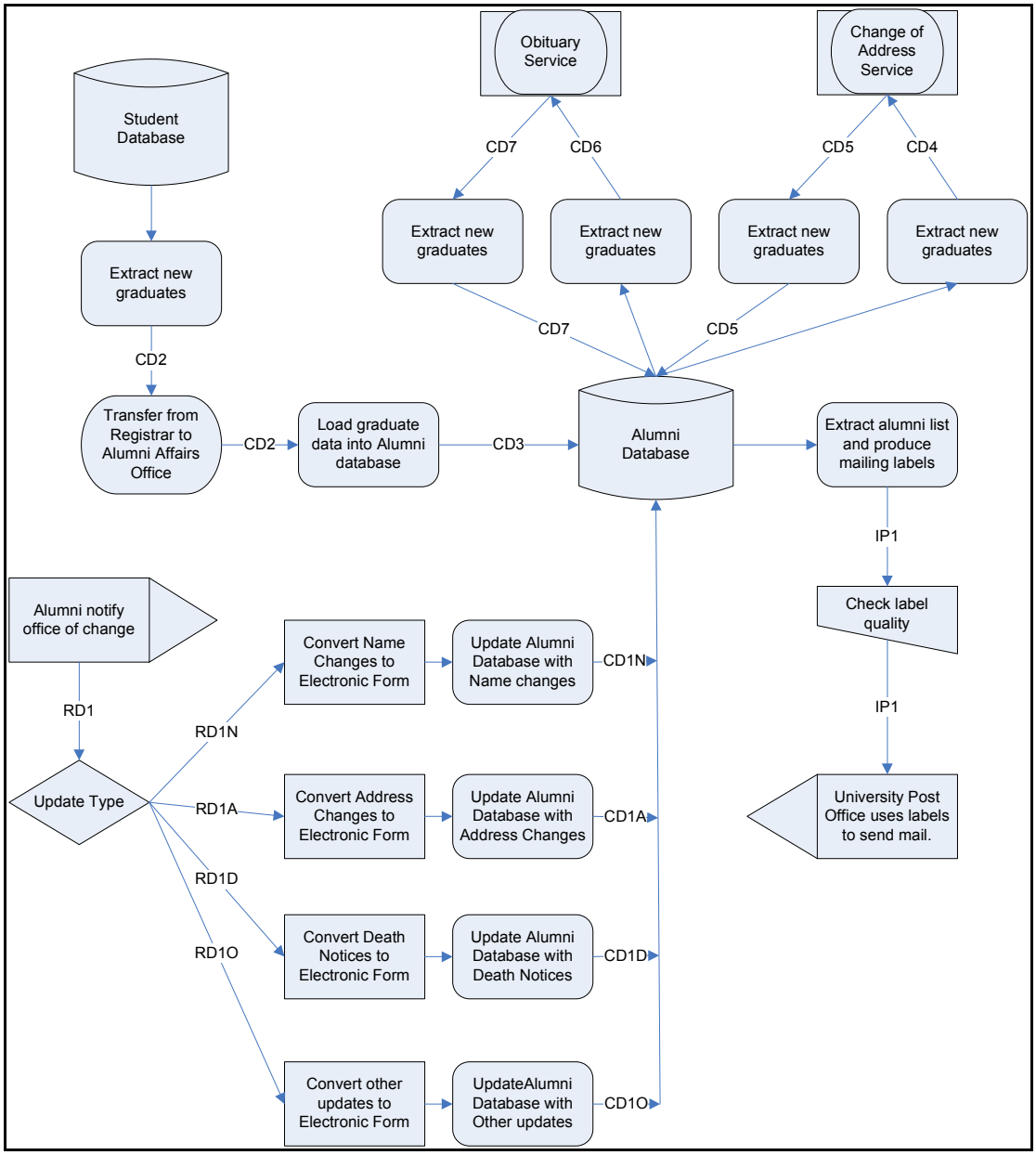


Figure 1. IP-MAP for production of alumni mailing labels by E. Pierce.

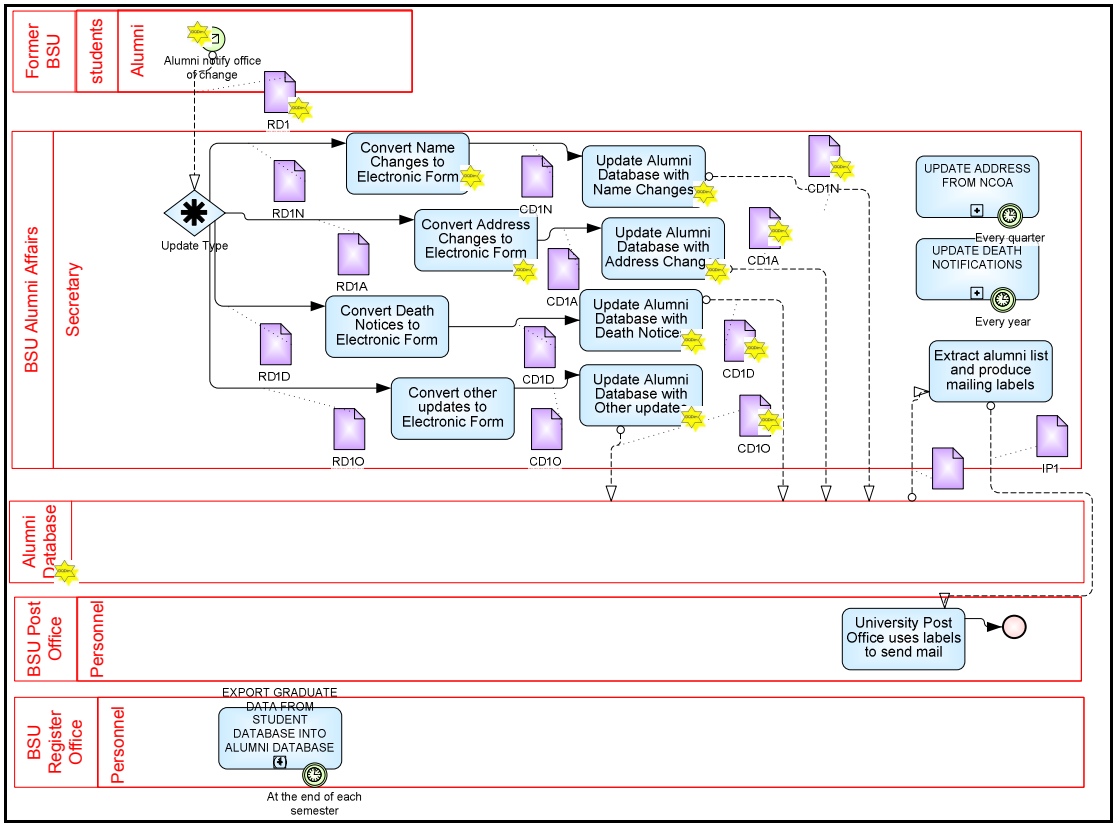


Figure 2. A translation to BPMN for production of alumni mailing labels.

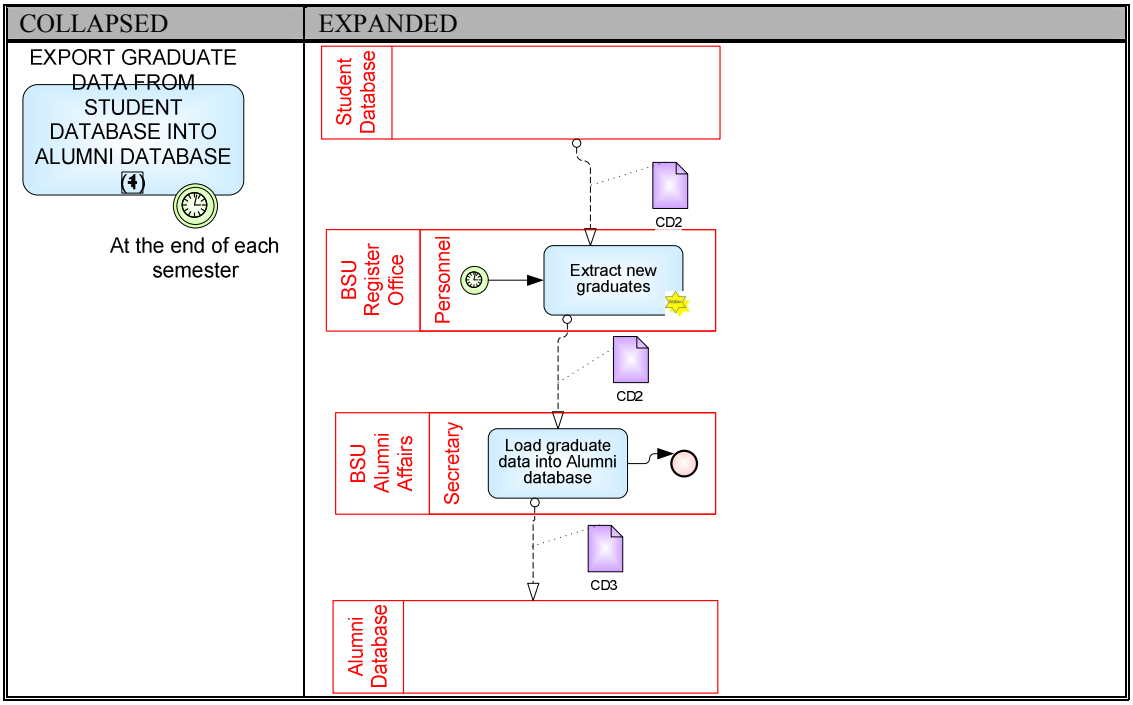


Table 20. A subtask for extracting alumni data from Student Database to Alumni Affairs.

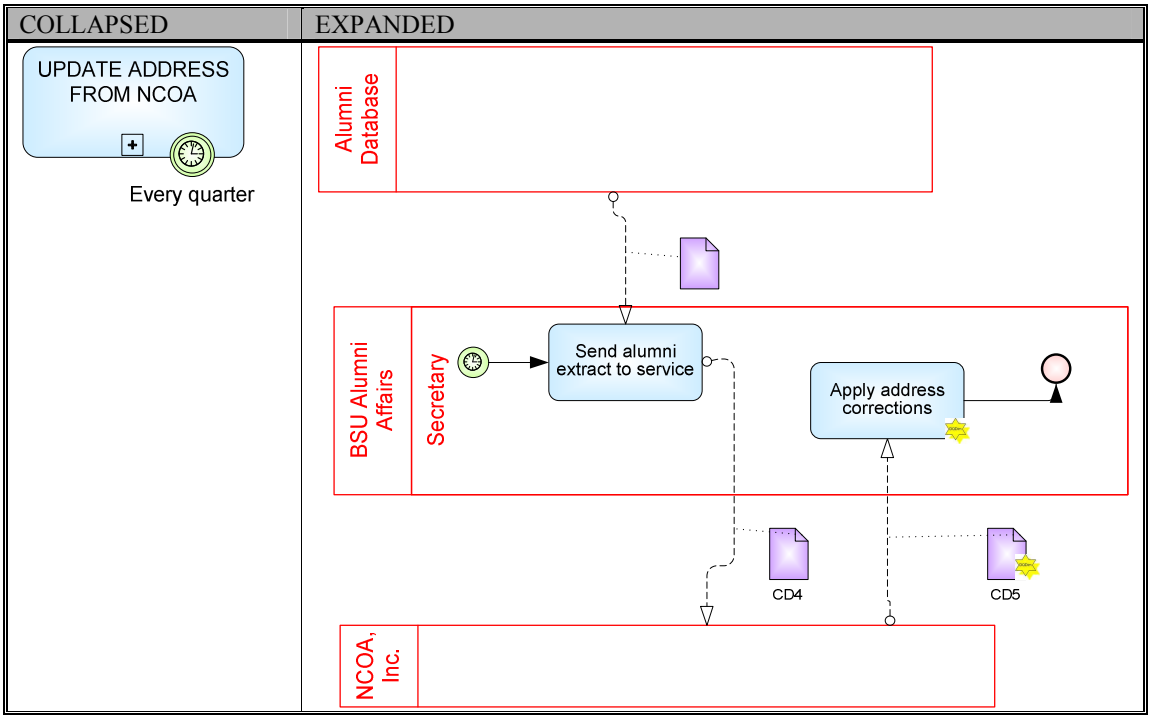


Table 21. A subtask for checking alumni data with Change of Address Database.

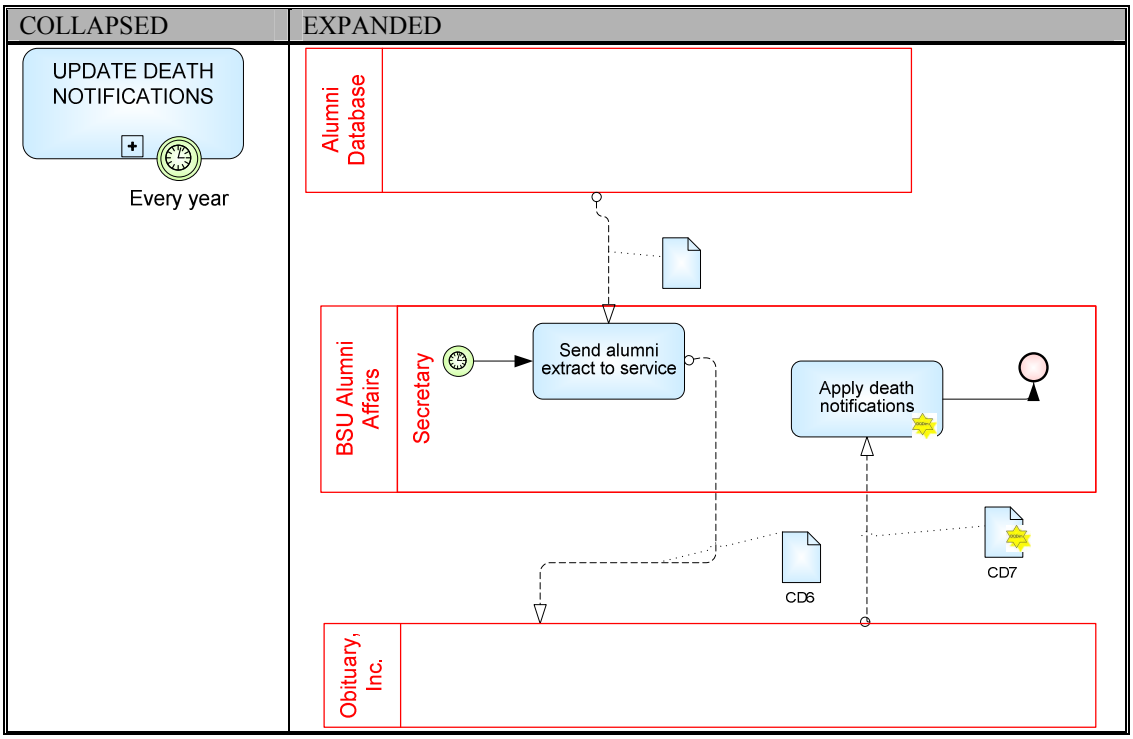


Table 22. A subtask for checking alumni data with Obituary Service.

CONCLUSION

This paper presents how to model the corresponding issues of data and information quality by using BPMN. Our starting point is IP-MAP, a specific notation developed by Shankaranarayanan[2].

We have analyzed the main constructs of IP-MAP, and we have mapped them to those provides by BPMN. From this comparison, we have identified which part of BPMN must be extended in order to better support business analysts' tasks, but with new features which satisfy the data quality criteria. To do this, we have proposed a first approach consisting in adding a new star-shaped symbol, denominated as "DQDim", which represents the data quality dimensions controlled in a business process. This new symbol modifies the elements of BPMN, in which it remains in order to add quality issues to these elements. We are currently performing different study cases so that we can enhance our approach in order to obtain a stronger BPMN notation to represent business processes and the quality issues involved in these aforementioned processes.

Given that BPMN is currently supported by several industrial tools, as part of our future work, we will intend to also introduce the data quality foundations described in the papers by means of MDA technologies.

ACKNOWLEDGE

This research is part of the projects DQNet (TIN2008-04951-E) supported by the Spanish Ministerio of Educación y Ciencia, IVISCUS (PAC08-0024-5991) supported by the JCCM, the IQMF-Tool supported by the University of Castilla-La Mancha.

REFERENCES

[1] Ballou, D.P. and Tayi, G.K. *Managerial Issues in Data Quality*. in *First International Conference on Information Quality (ICIQ'96)*. 1996. MIT, Cambridge, MA, USA.

[2] Ballou, D.P., Wang, R.Y. and Pazer, H., *Modelling Information Manufacturing Systems to Determine Information Product Quality.* Management Science, 1998. **44**(4): p. 462-484.

[3] Batini, C., Capiello, C., et al., *Methodologies for data quality assessment and improvement.* ACM Computing Surveys, 2009.

[4] Caballero, I., Caro, A., et al., *IQM3: Information Quality Management Maturity Model.* Journal on Universal Computer Science, 2008. **14**(22): p. 3658-3685.

[5] ISO-25012, *ISO/IEC 25012: (SQuaRE) - Data Quality Model* 2008.

[6] OMG, *Business Process Model Notation Specification 1.2.* 2009, Object Management Group.

[7] Pierce, E.M., *Integrating IP-MAPS with Business Process Modelling*, in *Information Quality Management: Theory and Applications*, Al-Hakim, L., Editor. 2007, Idea Group Publishing: Hershey, PA, USA. p. 145-167.

[8] Redman, T.C., *Data Quality for the Information Age*. 1996, Boston, MA, USA: Artech House Publishers.

[9] Shankaranarayan, G., Wang, R.Y. and Ziad, M. *IP-MAP: Representing the Manufacture of an Information Product*. in *Fifth International Conference on Information Quality*

(ICIQ'2000). 2000. MIT, Cambridge, MA, USA.
[10] Wang, R.Y., *A Product Perspective on Total Data Quality Management.* Comm. of the ACM, 1998. **41**(2): p. 58-65.