Requirements Elicitation in Data Mining for Business Intelligence Projects

Paola Britos[1], Oscar Dieste[2] and Ramón García-Martínez[3]

[1]Software and Knowledge Engineering Center. Buenos Aires Institute of Technology, AR

[2]Empirical Software Engineering Research Group, Polytechnic University of Madrid, ES

[3]Intelligent Systems Laboratory. Engineering School. University of Buenos Aires, AR

{pbritos, rgm}@itba.edu.ar, odieste@fi.upm.es

Abstract: There are data mining methodologies for business intelligence (DM-BI) projects that highlight the importance of planning an ordered, documented, consistent and traceable requirement's elicitation throughout the entire project. However, the classical software engineering approach is not completely suitable for DM-BI projects because it neglects the requirements specification aspects of projects. This article focuses on identifying concepts for understand DM-BI project domain from DM-BI field experience, including how requirements can be educed by a proposed DM-BI project requirements elicitation process and how they can be documented by a template set.

1. Introduction

A Data Mining for Business Intelligence (DM-BI) methodology seeks to organize the pattern discovery process in the data warehouse of an organization. These methodologies consider requirements specification as one of the early activities of the project (Chapman et al. 2000; Pyle 2003). Similarly, requirements are an important phase in software engineering methodologies (IEEE, 1993; Winter & Strauch 2004; Maiden et al. 2004, 2007; Solheim et al. 2005; Jiang & Eberlein 2007).

Several authors (Winter & Strauch 2002; Silva & Freire 2003; Yang and Wu, 2006) have addressed the need to improve DM-BI methodologies, but they focuses on DM-BI goals definition and DM-BI tasks specification as exploratory data analysis and develop tools for DM-BI process documentation, model-building, and pattern-finding. The DM-BI community has neglected the requirements specification aspects of projects, failing to identify any technique to elicit necessary knowledge or suggest any template for systematic documentation of requirements.

In order to explore how to minimize the impact of the presented problems this research focuses on an approach based on: understanding the DM-BI project's domain, knowing the DM-BI project's data domain, understanding the DM-BI project's scope, identifying the needed human resources, and selecting the appropriate DM-BI tool. The approach also looks to specify documentation tools for required information of DM-BI projects.

In this paper we present related research on DM methodologies addressing the problem (section 2); a solution approach is introduced (section 3); a proposed method for requirements engineering in DM-BI projects is developed (section 4); focusing on process and products, an example of a real case based use of templates is drawn (section 5); a discussion of the strengths of the proposed process is presented (section 6); and some conclusions are drawn (section 7).

2. Current Methodologies for DM-BI

The DM-BI literature on requirements elicitation identifies concepts related to how to extract, transform, aggregate, and discover business patterns in organization data. Moreover, these activities should be performed based on a concise dimensional schema. In this context, stakeholders and requirements engineers work together to identify what and where to look within organization data sources, in order to provide the bases for discovering business patterns for business improvement. The requirement elicitation process is addressed by most commonly used data mining (DM) methodologies (Chapman *et al*, 2000; Pyle, 2003, SAS, 2008). DM methodologies state the necessity of business understanding as the starting point for any DM project.

The CRISP (cross industry standard for data mining) methodology (Chapman *et al*, 2000) consists of four levels of abstraction, hierarchically organized from general tasks to specific cases. The process is divided into six phases, each one having many general tasks of second level or sub phases. General tasks are projected to specific ones, where the actions that must be developed for specific situations are described. As a consequence, we find a general task "cleaning data" in second level; then in third level, those tasks that must be developed for a specific case, as for example "cleaning numerical data" or "cleaning categorical data". In the fourth level, groups of actions, decisions and results about the specific data mining project are collected. The CRISP-DM methodology presents two different documents as a tool for assisting during the development of the data mining project: the reference model and the user's guide. The document model of reference describes, in a general way, the phases, general tasks and exit-points of a data mining project. The user's guide brings detailed information about practical application of the model of reference to specific data mining projects; it also gives advices and check-lists about each phase's tasks.

The methodology P^3TQ (Product, Place, Price, Time, and Quantity) consists of two parts (Pyle, 2003): [a] Modeling (PI): provides a step-by-step guide to develop and to build a model to address a business problem or opportunity. Modeling depends very much on the business circumstances that prompt the modeling in the first place, as indicated by the five different entry scenarios to PI. Largely, PI provides lists of actions that must be completed, depending on circumstances; and [b] Data Mining (PII): provides a step-by-step guide to mining the data to produce the required model as identified in PI. Data Mining consists of a series of stages that have to be completed in order. Unlike modeling in which several tasks may take place at the same time, mining has to proceed from stage to stage. Each part is based on four types of "activity boxes"; *action boxes*: indicate one or more required "next steps" for you to take; *discovery boxes*: provide exploratory actions that you need to take to decide what to do next; *technique boxes*: provide supplemental information about the recommended steps to be described in the action or discovery boxes; and *example boxes*: gives a detailed description of how to use a specific technique, along with pointers to an excel worksheet.

SEMMA (Sample, Explore, Modify, Model and Assess) is a methodology oriented to select, explore and model a great amount of data; looking to discover business patterns in the data (SAS, 2008). The process begins with the extraction of sample data on which analysis is going to be applied. Once the sample is selected, the methodology proposes to explore the data in order to simplify the model. The third phase involves entailing data to DM tool. The fourth phase involves running the DM tool on the selected data. The last phase consists of evaluating results by analyzing the model by contrast with statistical models or new samples.

One assumption behind approaches to requirements engineering in DM-BI is that sufficient knowledge of the requirements already exits. It is quite well known that in normal situations, customers and users are 'speaking another language' than the development team (Maiden et al. 2007). The task of translating customers' and users' ideas into the development teams' language is done by requirements engineers and business-analysts using different notations (Jiang & Eberlein 2007). However, this is increasingly flawed because of the breadth of expertise that is needed to specify complex systems and the number of humans that may be involved in the process. Thus, current requirements elicitation methodologies fail in that they do not provide adequately coverage of concepts needed to elicit requirements, nor do they support corresponding documentation or cross referencing.

3. Framework for Requirements Elicitation in DM-BI

The need to adapt traditional requirements engineering process for DM-BI systems is based on the premise that the requirements analysis for these types of systems differ substantially from requirements analysis for conventional information systems. Evidence of this assertion is found in a wide range of DM-BI project domains: mobile telephony (Grosser *et al*, 2005), health policies (Felgaer *et al*, 2006), agro-industry (Cogliati *et al*, 2006), and criminal intelligence (Valenga *et al*, 2008). In each of these cases, the DM-BI methodologies had difficulty dealing with some common requirements problems, such as the customer doesn't understand the technical lexis used by DM-BI group, the customer were not clear about the goals and capabilities of the DM-BI project or what it could achieve, or the models defined by DM-BI group were different from the ones the customer envisioned. A complete list of the identified problems is shown in Fig. 1. This field experience has taught the authors the necessity of defining a list of concepts to be educed during the business understanding phase. The list of needed concepts and its relation to the detected problems is shown in Fig. 1.

PROBLEM	CONCEPTS TO BE EDUCED
[a] The customer doesn't understand the technical lexis used by DM-BI group [b] DM BI group can't understand the lexis of the customer's information domain [c] The DM-BI group found it hard to understand how they could help the customer because they didn't know the project domain	Definitions, acronyms and abbreviations
[d] The customer was not sure what the DM-BI project could do or achieve [e] Models defined by DM- BI group were different from the ones the customer envisioned	Project objectives Successful criteria of the project Project expectations Project suppositions
[f] The customer was an unpredictable group (not so concerned with the project)	Human resource involved
[g] The customer did not know the needed organizational information and its condition	Project restrictions Project risks Contingency planning
[h] Data identified by requirements were not the right ones	Requirements goal The requirement information or data source Attributes related on requirements
[i] When DM-BI project was in modeling phase (requirements solutions) and DM-BI group detected problems in data, (i.e., data identified by requirements were not the right ones), it was necessary to redefine requirements	Requirement results suppositions Requirement restrictions Requirement risks Requirement contingencies plan
[j] Requirements of DM-BI project misunderstandings resulted in selection of the wrong modeling tool	Evaluating DM-BI tools

Fig. 1. Relation among identified problems in the field and the Concepts needed to be educed

To solve these problems, we have needed to educe specific information in each DM-BI project. This information may be modeled by a list of concepts that are educed in the listed below:

- *Definitions, acronyms and abbreviations:* It is necessary to identify definitions, acronyms and abbreviations for establishing lexis to be shared among all persons related to the DM-BI project. This addresses problems: [a], [b] and [c] (see Fig. 1).
- *Project objectives*: It is necessary to identify the objective of the DM-BI project and its motivation to characterize what customer needs. This addresses problems: [d] and [e].
- *Successful criteria of the project*: It is necessary to identify the criteria which turn the project into a successful one. The criteria must be described in terms of expected achievements of the DM-BI project. This addresses problems: [d] and [e].
- *Project expectations*: It is necessary to identify what is expected to be achieved by the DM-BI project and to confirm that they fulfill the customer's expectations. The expectations must be aligned with the objectives and the project success criteria. This addresses problems: [d] and [e].
- *Project's suppositions*: It is important to identify the suppositions that must be assumed as true ones in order to start the DM-BI project. The project's suppositions become the start point of the requirement elicitation process. This addresses problems: [d] and [e].

- *Project restrictions*: In order to specify the DM-BI project context, it is necessary to identify the limits previously established for the project: *related to organization*: political, legal and data quantity, *related to data*: to access sources of information, and data quality, *related to human and technical resources*: the size of the data sources related to hardware and software handlers, hardware and software limitations, human resources; and *related to the project*: those activities which affect the project and its security (access to documentation about the project, without any possibility of a backup). This address problem: [g].
- *Project risks*: Identify risks for the DM-BI project by looking continuously at what might be wrong in the organization (related to the DM-BI project) and determining which risks are important to be solved. Risks identification is needed to define contingency plans to be applied to mitigate risk. This address problem: [g].
- *Contingency plans*: It is necessary to define contingency plans to be applied to off-set risk. This address problem: [g].
- *Human resource involved*: It is important to identify the different roles in the DM-BI project and the human resources that will fill these roles. The roles are in the areas of exploring data and business domain expertise. This address problem: [f].
- *Requirement goals*: The project's objectives are decomposed in requirement goals. The requirement goals are needed in conjunction with project's suppositions to define the DM-BI processes to be applied. This address problem: [h].
- *The requirement information or data source*: It is necessary to establish which information or data source are going to be used and where is it in order to accomplish a requirement's goal. This address problem: [h].
- *Requirement results suppositions*: It is necessary to identify the suppositions about requirement results in order to have guidance to act to accomplish the requirement goal. It must be consistent with the project goal, its expectations and suppositions. This addresses problem: [i].
- *Requirements restrictions*: In order to specify each DM-BI project requirement context, it is necessary to identify requirement limits which must be consistent with those described in other parts of the elicitation document: *related to data*: to access sources of information, and data quality, *related to human and technical resources*: the size of the data sources related to hardware and software handlers, hardware and software limitations, human resources; and *related to the project security*. This address problem: [i].
- *Attributes related on requirements*: Establish which attributes are going to be used in order to accomplish a requirement goal. This address problem: [h].
- *Requirement risks*: It is important to identify DM-BI project requirement risks by looking continuously for what might be wrong in the requirement (related to the DM-BI project) and determining which risks are important to be solved. Requirement risks identification is needed to define contingency plans to be applied when needed. This address problem: [i].
- *Requirement contingency plans*: It is necessary to define contingency plans to be applied when an occurrence warrants. This address problem: [i].
- *Evaluating DM-BI tools*: It is necessary to evaluate available DM-BI tools to establish which are the best ones to accomplish the project's objectives. This address problem: [j].

Based on these concepts, to address the problems identified in Fig. 1, we propose a method for DM-BI project requirements elicitation process next.

4. Proposed Method

The proposed method of five steps is described in section 4.1, and the process products and their relation with the process steps is shown in section 4.2.

4.1. Process

Once the needed concepts have been identified, it is necessary to establish the steps to educe those concepts. The proposed structure is similar to those proposed by Software Engineering that allows progressing over the needed concepts to maintain their natural order. In the business understanding phase of any DM-BI methodology, we propose a DM-BI project requirements elicitation process of five steps that is shown in Fig. 2.

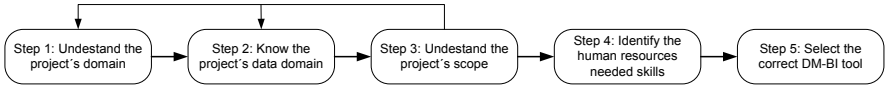


Fig. 2. Process of requirements elicitation

The purpose of the step "understand the project's domain" consists of establishing communication channels in ordinary language among persons involved into the DM-BI project. The purpose of the step "know the project's data domain" consists of establishing the project's requirements; the data needed for those requirements and its location, risks involved in the data and the requirements´ development, the data and requirements' restrictions, and finally its suppositions. The purpose of the step "understand the project's scope" is to achieve the DM-BI projects objective, its limitations, expectations and risks. The purpose of the step "identify the human resources needed skills" consists of knowing the list of human resources involved, its restrictions, risks and responsibilities. The purpose of the step "select the correct DM-BI tool" is to select an adequate tool according to the information obtained in the earlier steps.

To know the project's data domain in terms of requirements goal, the requirements information of data source information, requirements results suppositions, requirements restrictions, attributes involved in requirements, risks and contingency plans; it is necessary to understand the project's domain in terms of definitions, acronyms and abbreviations. To understand the project's scope in terms of project objectives, successful criteria of the project, project expectations, project suppositions, restrictions, risks, and contingency plans; it is necessary to know the project's data domain in terms of requirements goal, the requirements information of data source information, requirements results suppositions, requirements restrictions, attributes involved in requirements, requirements risks and requirements contingency plans. To identify the human resources needed in terms of defining human resources involved; it is necessary to understand the project's scope in terms of project objectives, project successful criteria, project expectations, project suppositions, project restrictions, project risks, and contingency plans. To identify the human resources needed skills in terms of defining human resources involved; it is necessary to select the correct DM-BI tool in terms of tools evaluation.

The conceptual dependency among the needed concept is shown in Fig. 3.

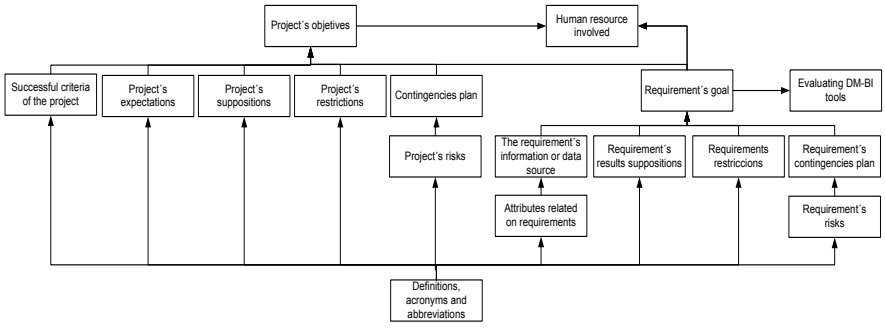


Fig. 3. Cross references of educed concepts represented by the templates

4.2 Products

We have defined a set of templates. Each template is associated to each concept. These templates have a detailed description of the concepts to be educed (see examples in section 5). The templates allow the concept evolution through the requirements elicitation process. The relation between the educed concepts as products and the steps of the proposed process (see section 4.1) to generate them is shown in Fig. 4.

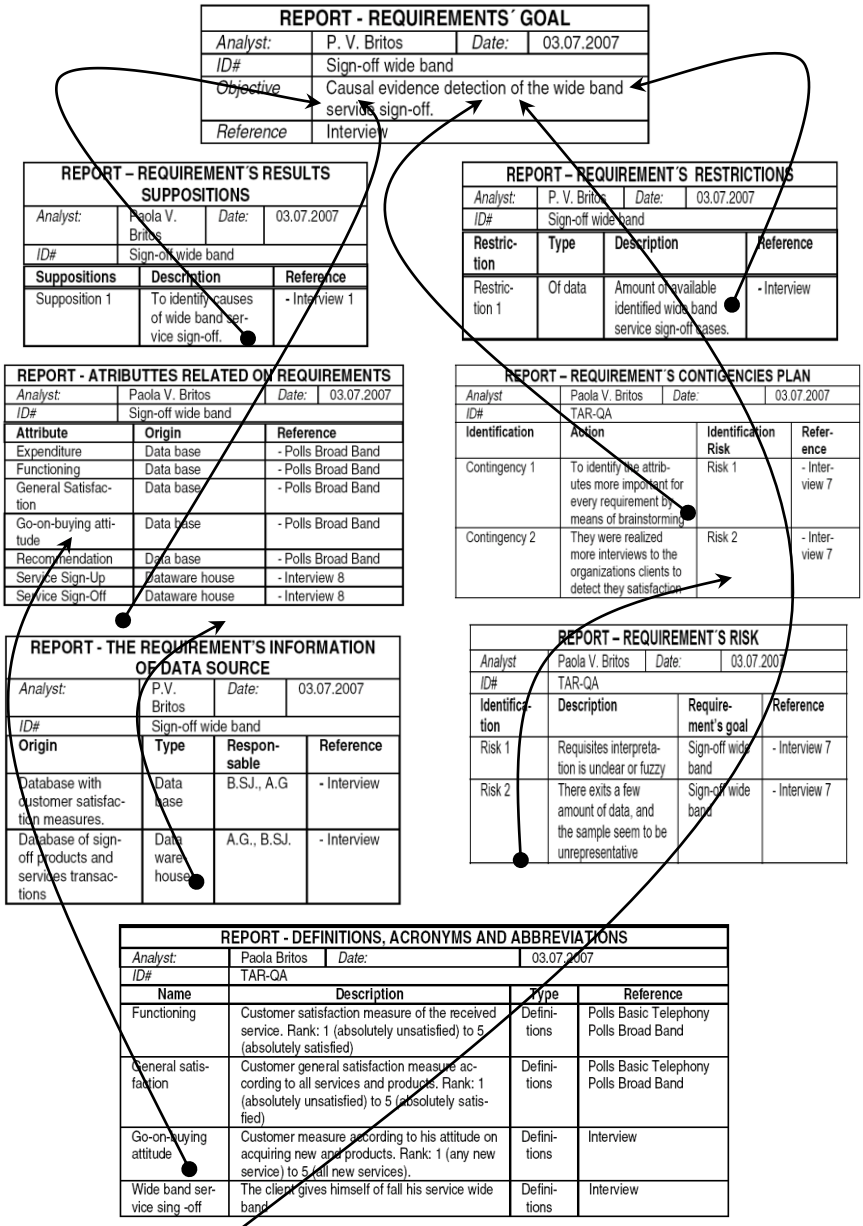
PRODUCT (concepts to be educed) / EPS	itions, acronyms and abbreviations	ct's objectives	essful criteria of the project	ct's expectations	ct's suppositions	ct's restrictions	ct's risks	ngencies plan	an resource involved	irement's goal	equirement's information or data source	irement's results suppositions	irements restrictions	utes related on requirements	irement's risks	irement's contingencies plan	lating DM-BI tools
derstand the project's domain	■								■								
ow the project's data domain	■								■	■	■	■	■	■	■	■	
derstand the project's scope	■	■	■	■	■	■	■	■	■								
ntify the human resources needed skills	■								■								
ect the correct DM-BI tool	■								■	■			■	■			■

Fig. 4. Relation among products (educed concepts) and process steps

5. Examples of Real Case Based Use of Templates

In this section we present a set of template examples based on a real DM-BI project (real case example) within the telecommunications industry. The case is centered on a company that scrutinizes the customers´ service closely, and the project objective is to show the relation between customer satisfaction (especially customers fidelity), and the company's products and qualities improvement initiatives. The DM-BI project requirements elicitation products and concept cross references are captured by the fulfillment and interaction among the different templates (i.e. for the real case example see Fig. 5 to Fig. 8).

Fig. 5 shows how the requirements objective: "causal evidence detection of the wide band service sign-off" (see Template "Report - Requirements Goal") needs supposition 1: "to identify causes of wide band service sign-off" (see Template "Report - Requirement's Results Supposition"), restriction 1: "amount of available identified wide band service sign- off cases" (see Template: "Report – Requirement's Restrictions"), attribute: "Service Sign-Off" (see Template: "Report - Attributes Related On Requirements"), contingency 1: "to identify the attributes more important for every requirement by means of brainstorming" (see Template: "Report – Requirement's Contingencies Plan"). The information origin for the attribute: "Service Sign-Off" (see Template: "Report - Attributes Related on Requirements") is in "Database of sign-off products and services transactions" (see Template "Report - The Requirement's Information of Data Source").



REPORT - REQUIREMENTS´ GOAL

Analyst:	P. V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		
Objective	Causal evidence detection of the wide band service sign-off.		
Reference	Interview		

REPORT – REQUIREMENT´S RESULTS SUPPOSITIONS

Analyst:	Paola V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		

Suppositions	Description	Reference
Supposition 1	To identify causes of wide band service sign-off.	- Interview 1

REPORT – REQUIREMENT´S RESTRICTIONS

Analyst:	P. V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		

Restriction	Type	Description	Reference
Restriction 1	Of data	Amount of available identified wide band service sign-off cases.	- Interview

REPORT - ATRIBUTTES RELATED ON REQUIREMENTS

Analyst:	Paola V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		

Attribute	Origin	Reference
Expenditure	Data base	- Polls Broad Band
Functioning	Data base	- Polls Broad Band
General Satisfaction	Data base	- Polls Broad Band
Go-on-buying attitude	Data base	- Polls Broad Band
Recommendation	Data base	- Polls Broad Band
Service Sign-Up	Dataware house	- Interview 8
Service Sign-Off	Dataware house	- Interview 8

REPORT – REQUIREMENT´S CONTIGENCIES PLAN

Analyst	Paola V. Britos	*Date:*	03.07.2007
ID#	TAR-QA		

Identification	Action	Identification Risk	Reference
Contingency 1	To identify the attributes more important for every requirement by means of brainstorming	Risk 1	- Interview 7
Contingency 2	They were realized more interviews to the organizations clients to detect they satisfaction	Risk 2	- Interview 7

REPORT - THE REQUIREMENT´S INFORMATION OF DATA SOURCE

Analyst:	P.V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		

Origin	Type	Responsable	Reference
Database with customer satisfaction measures.	Data base	B.SJ., A.G	- Interview
Database of sign-off products and services transactions	Data ware house	A.G., B.SJ.	- Interview

REPORT – REQUIREMENT´S RISK

Analyst	Paola V. Britos	*Date:*	03.07.2007
ID#	TAR-QA		

Identification	Description	Requirement's goal	Reference
Risk 1	Requisites interpretation is unclear or fuzzy	Sign-off wide band	- Interview 7
Risk 2	There exits a few amount of data, and the sample seem to be unrepresentative	Sign-off wide band	- Interview 7

REPORT - DEFINITIONS, ACRONYMS AND ABBREVIATIONS

Analyst:	Paola Britos	*Date:*	03.07.2007
ID#	TAR-QA		

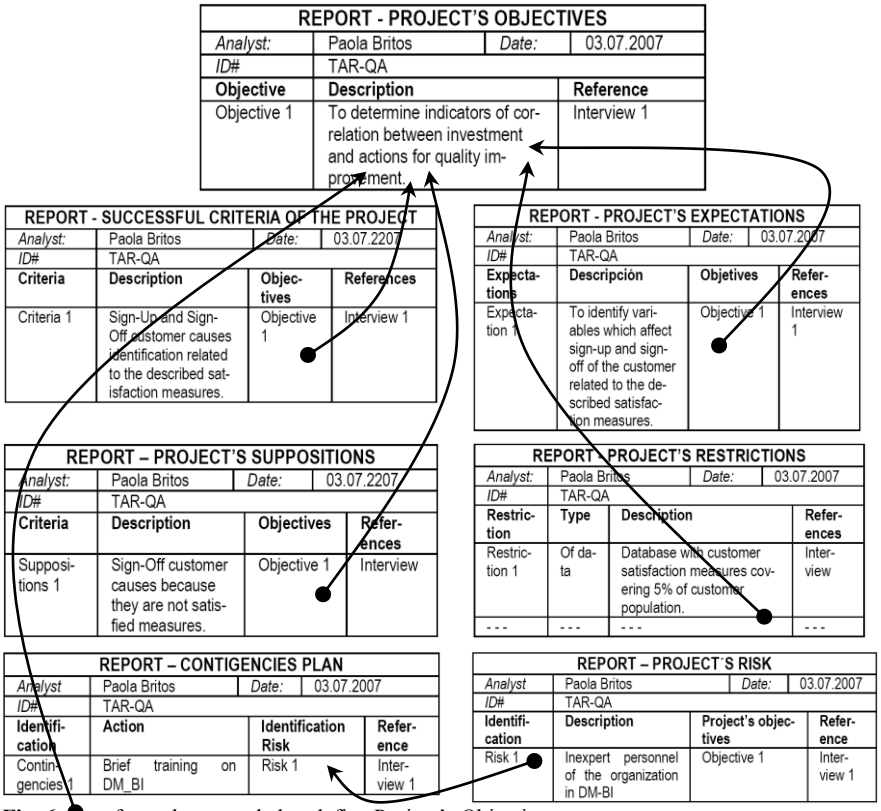
Name	Description	Type	Reference
Functioning	Customer satisfaction measure of the received service. Rank: 1 (absolutely unsatisfied) to 5 (absolutely satisfied)	Definitions	Polls Basic Telephony Polls Broad Band
General satisfaction	Customer general satisfaction measure according to all services and products. Rank: 1 (absolutely unsatisfied) to 5 (absolutely satisfied)	Definitions	Polls Basic Telephony Polls Broad Band
Go-on-buying attitude	Customer measure according to his attitude on acquiring new and products. Rank: 1 (any new service) to 5 (all new services).	Definitions	Interview
Wide band service sing -off	The client gives himself of fall his service wide band	Definitions	Interview

Fig. 5. Set of templates needed to define Requirements Goal

The information origin for the contingency 2: "they were realized more interviews to the organizations clients to detect their satisfaction" (see Template: "Report – Requirement's Contingencies Plan") is in risk 2: “there exits a few amount of data, and the sample seem to be unrepresentative” (see Tem plate: “Report – Requirement’s Risk”). The definition of the concept: “go-on-buying attitude” (see Template: “Report - Definitions, Acronyms and Abbreviations”) is used to understand the meaning of that attribute in Template: "Report - Attributes Related on Requirements". The definition of the concept: “wide band service sing -off” (see Template: “Report - Definitions, Acronyms and Abbreviations”) is used to understand the requirements objective: "causal evidence detection of the wide band service sign-off" (see Template "Report - Requirements Goal").

Fig. 6 shows that the project’s objective: "to determine indicators of correlation between investment and actions for quality improvement" (see Template "Report - Project’s Objectives") needs criteria 1: " sign-Up and Sign-Off customer causes identification related to the described satisfaction

measures." (See Template "Report – Successful Criteria of the Project"), expectation 1: “to identify variables



REPORT - PROJECT'S OBJECTIVES

Analyst:	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Objective	**Description**		**Reference**
Objective 1	To determine indicators of correlation between investment and actions for quality improvement.		Interview 1

REPORT - SUCCESSFUL CRITERIA OF THE PROJECT

Analyst:	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Criteria	**Description**	**Objectives**	**References**
Criteria 1	Sign-Up and Sign-Off customer causes identification related to the described satisfaction measures.	Objective 1	Interview 1

REPORT - PROJECT'S EXPECTATIONS

Analyst:	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Expectations	**Descripción**	**Objetives**	**References**
Expectation 1	To identify variables which affect sign-up and sign-off of the customer related to the described satisfaction measures.	Objective 1	Interview 1

REPORT – PROJECT'S SUPPOSITIONS

Analyst:	Paola Britos	Date:	03.07.2207
ID#	TAR-QA		
Criteria	**Description**	**Objectives**	**References**
Suppositions 1	Sign-Off customer causes because they are not satisfied measures.	Objective 1	Interview

REPORT – PROJECT'S RESTRICTIONS

Analyst:	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Restriction	**Type**	**Description**	**References**
Restriction 1	Of data	Database with customer satisfaction measures covering 5% of customer population.	Interview
---	---	---	---

REPORT – CONTIGENCIES PLAN

Analyst	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Identification	**Action**	**Identification Risk**	**Reference**
Contingencies 1	Brief training on DM_BI	Risk 1	Interview 1

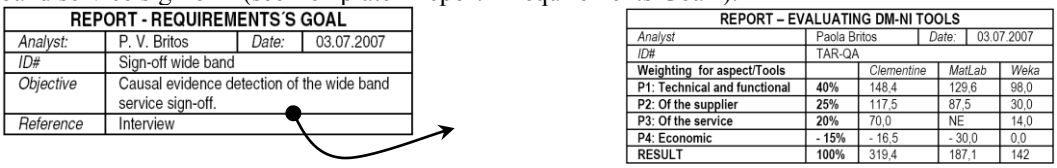
REPORT – PROJECT´S RISK

Analyst	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Identification	**Description**	**Project's objectives**	**Reference**
Risk 1	Inexpert personnel of the organization in DM-BI	Objective 1	Interview 1

Fig. 6. Set of templates needed to define Project's Objectives

which affect sign-up and sign-off of the customer related to the described satisfaction measures" (see Template: "Report – Project's Expectations"), supposition 1: " sign-Off customer causes because they are not satisfied measures " (see Template: "Report – Project's Suppositions"), restriction 1: “Database with customer satisfaction measures covering 5% of customer population” (see Template: “Report – Project's Restrictions“), contingency 1: "brief training on DM_BI" (see training Template: "Report – Contingencies Plan"). The information origin for the contingency 1: "brief on DM_BI" (see Template: "Report –Contingencies Plan") is in risk 1: “inexpert personnel of the organization in DM-BI” (see Template: “Report – Project's Risk”).

Fig. 7 shows that to evaluate DM-BI tools needs requirements objective: "causal evidence detection of the wide band service sign-off" (see Template "Report - Requirements Goal").



REPORT - REQUIREMENTS´S GOAL

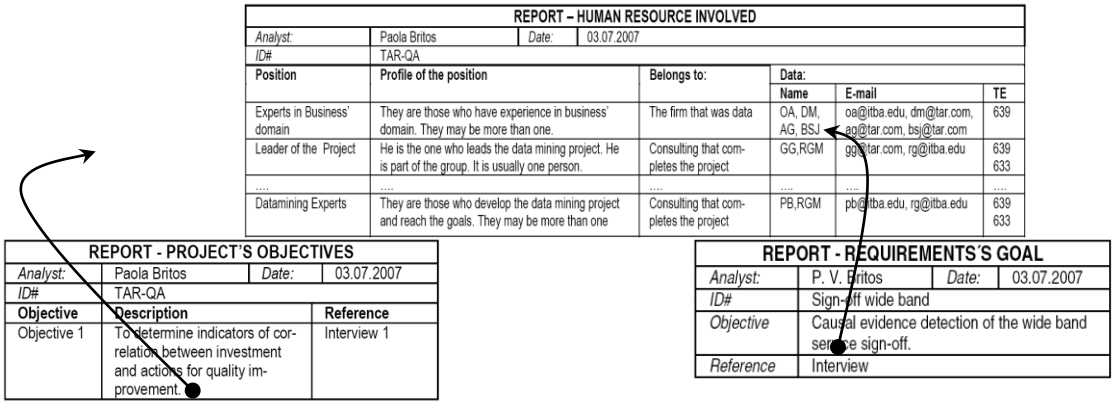
Analyst:	P. V. Britos	Date:	03.07.2007
ID#	Sign-off wide band		
Objective	Causal evidence detection of the wide band service sign-off.		
Reference	Interview		

REPORT – EVALUATING DM-NI TOOLS

Analyst	Paola Britos	Date:	03.07.2007
ID#	TAR-QA		
Weighting for aspect/Tools		*Clementine*	*MatLab*	*Weka*
P1: Technical and functional	**40%**	148,4	129,6	98,0
P2: Of the supplier	**25%**	117,5	87,5	30,0
P3: Of the service	**20%**	70,0	NE	14,0
P4: Economic	**- 15%**	- 16,5	- 30,0	0,0
RESULT	**100%**	319,4	187,1	142

Fig. 7. Set of templates needed to define Evaluating DM-BI Tools

Fig. 8 shows how the human resources involved: "experts in business' domain", “leader of the project” and “data mining experts” in other (see Template "Report - Human Resource Involved") needs objective 1 “to determine indicators of correlation between investment and actions for quality improvement" (see Template "Report - Project's Objectives"), and the requirements objective: "causal evidence detection of the wide band service sign-off" (see Template "Report - Requirements Goal") .



REPORT – HUMAN RESOURCE INVOLVED						
Analyst:	Paola Britos	*Date:*	03.07.2007			
ID#	TAR-QA					
Position	**Profile of the position**		**Belongs to:**	**Data:**		
				Name	**E-mail**	**TE**
Experts in Business' domain	They are those who have experience in business' domain. They may be more than one.		The firm that was data	OA, DM, AG, BSJ	oa@itba.edu, dm@tar.com, ag@tar.com, bsj@tar.com	639
Leader of the Project	He is the one who leads the data mining project. He is part of the group. It is usually one person.		Consulting that completes the project	GG,RGM	gg@tar.com, rg@itba.edu	639 633
....	....		....	....	....	....
Datamining Experts	They are those who develop the data mining project and reach the goals. They may be more than one		Consulting that completes the project	PB,RGM	pb@itba.edu, rg@itba.edu	639 633

REPORT - PROJECT'S OBJECTIVES			
Analyst:	Paola Britos	*Date:*	03.07.2007
ID#	TAR-QA		
Objective	**Description**		**Reference**
Objective 1	To determine indicators of correlation between investment and actions for quality improvement.		Interview 1

REPORT - REQUIREMENTS'S GOAL			
Analyst:	P. V. Britos	*Date:*	03.07.2007
ID#	Sign-off wide band		
Objective	Causal evidence detection of the wide band service sign-off.		
Reference	Interview		

Fig. 8. Set of templates needed to define Human Resource Involved

6. Discussion

Current DM-BI methodologies fail to educe all the concepts (see section 3) needed during the business understanding phase of DM-BI (shown in Fig. 9). CRISP-DM educe on set of concepts, P3TQ another and SEMMA yet a third. In general, these methodologies attend to concepts related to determining business objectives and assess situations (at least for one methodology) and concepts related to determine data mining goals and project plan production are not attended. In this context, the proposed methodology is more robust than current ones, because it educes all the necessary concepts to model the DM-BI project's requirements.

Our consulting engagements in DM-BI projects have allowed us to test our ideas in the field, but we recognize the necessity for a formal research approach. So the next step will be to carry out experiments comparing the proposed DM-BI project requirements elicitation process with existing ones. The first step in the experiment will be to build a set of testing DM-BI project cases in which each case includes a case description K_i (description of case i) and the list of the requirements of the case R_i (requirements of case i) to be educed. Then two development groups will be considered, one trained in existing DM-BI business understanding phase (control group) and the other trained in our approach (testing group); both groups will be asked to identify and document DM-BI project requirement from the set of the previously defined DM-BI project cases (the set of K_i). The list of the requirements of the case will not be shown to any groups. Referees will compare the amount of well identified requirements from the control group with those identified by the testing group, and using statistical tests, differences between the groups will be compared. We expect to validate experimentally that the amount of correct requirements educed by the testing group (the one using the proposed process) is significantly better than the amount of correct requirements educed by the control group (current methodologies).

7. Conclusion

This paper presents an approach to educe the requirements for DM-BI project that addresses identified weaknesses in current data mining methodologies. The approach is based on a list of DM-BI project requirements, needed concepts that have to be educed, a set of templates to document its elicitation and the associated process.

The proposed process and set of templates have been tuned in field cases and their effectiveness has been demonstrated. To further verify the effectiveness of the proposed approach, a formal experiment is planned for the second semester 2008 with a population of advance students at the Software Engineering Bachelor Program, University of Buenos Aires.

The focus on DM-BI project requirements documentation enables the achievement of requisite, consistent, and traceable requirements specifications over the entire project. This documentation encourages beginning the modeling activities based on a common lexis and cross-referenced concepts related to the target business domain requirements.

8. References

Chapman P, Clinton J, Keber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0 Step by step BIguide Edited by SPSS. http://www.crisp-dm.org/CRISPWP-0800.pdf Accessed 14 September 2006.

Cogliati M, Britos P, García-Martínez R (2006) Patterns in Temporal Series of Meteorological Variables Using SOM & TDIDT In: Bramer M (ed) Artificial Intelligence in Theory and Practice, Boston, Springer, IFIP Series 217:305-314

Felgaer P, Britos P, and García-Martínez R, (2006) Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques. Int. J. of Mod. Ph. C 17(3): 447-455

Grosser H, Britos P, García-Martínez R (2005) Detecting Fraud in Mobile Telephony Using Neural Networks. LNAI 3533:613-615

IEEE (1993) Standard IEEE 830-1993: Recommended Practice for Software Requirements Specifications. Institute of Electronic and Electrical Engineers Press.

IEEE (2004) Guide to the Software Engineering Body of Knowledge. IEEE Comp. Society Press

Jiang L, Eberlein A (2007) Selecting Requirements Engineering Techniques based on Project Attributes - A Case Study. 14th Annual IEEE ECBS: 269-278

Maiden N, Robertson S, Gizikis A (2004) Provoking Creativity: Imagine What Your Requirements Could be Like. IEEE Software 21(5): 68-75

Maiden N, Ncube C, Robertson S (2007) Can Requirements Be Creative? Experiences with an Enhanced Air Space Management System Proceedings 29th ICSE: 632-641

Pyle D (2003) Business Modeling and Business intelligence. Morgan Kaufmann

SAS (2008) SAS Enterprise Miner: SEMMA http://www.sas.com/technologies/analytics/ datamining/ miner/semma.html. Accessed 29 February 2008

Silva F, Freire J (2003) DWARF: An Approach for Requirements Definition and Management of Data Warehouse Systems. RE´03: 75-84

Solheim H, Lillehagen F, Petersen S, Jorgensen H, Anastasiou M (2005) Model-driven visual requirements engineering Proceedings RE´05:421-428

Valenga F, Fernández E, Merlino H, Rodríguez D, Procopio C, Britos P, García-Martínez R (2008) Minería de Datos Aplicada a la Detección de Patrones Delictivos en Argentina. VII JIISIC´08: 31-39

Winter R, Strauch B (2002) A Method for Demand-driven Information Requirements Analysis in Data Warehousing Projects. HICSS-36:231-239

Yang Q, Wu X (2006) 10 Challenging Problems in Data Mining Research. Int. J. Inf. Tech. & Decis. Mak. 5(4):597–604