Information Mining Projects Management Process

Sebastian Martins[1,2], Patricia Pesado[1,3], Ramón García-Martínez[2]
1. PhD Program on Computer Science, Computer Science School, National University of La Plata, Buenos Aires, Argentina
2. Information Systems Research Group, National University of Lanus, Buenos Aires, Argentina
3. III-LIDI. Computer Science School. National University of La Plata – CIC Buenos Aires, Argentina
martinssebastian@yahoo.com.ar, ppesado@lidi.info.unlp.edu.ar, rgarcia@unla.edu.ar

***Abstract*— Information Mining (also known as Knowledge Discovery Process) is a growing discipline in continuous expansion. Most of the progress accomplished, are focus on the development activities (i.e. those technical activities associated with the comprehension and adaptation of data, and the implementation of data mining algorithm). According to this conceptual framework, several process models were developed, which allow organizing and defining the set of tasks related to the development of information mining projects. These approaches omit the set of tasks oriented to the management and control of the process. In this paper, we propose a transversal management process to the development process currently in use in information mining projects. The proposed process focuses on removing existing gaps, providing an improvement on the project's maturity and quality levels.**

Keywords-component; Information Mining Engineering; Project Management of Information Mining; Knowledge Discovery Process; Data Mining.

I. INTRODUCTION

The data mining term, is strongly bound to the concept of large database and goes back to the definition of searching algorithms of knowledge patterns [1]. However, today there are research lines in such fields as text mining [2], images mining [3], patterns in information stream mining [4], web mining [5], among others. In this context, our search line considers appropriate to use the term "Information Mining" [6] as a generic reference to any of the aforementioned types of mining. Based on that software engineering has been defined in SWEBOK [7] as "the application of a systematic, disciplined and quantifiable approach to the development, operation and maintenance of software, and the study of these approaches, i.e. the application of engineering to software"; it is agreed define Information Mining Engineering (IME) [8] as the application of a systematic, disciplined and quantifiable approach to the development of Information Mining Projects and the study of these approaches, i.e. the application of engineering to Information Mining. From this perspective, information mining engineering focuses on defining procedures to guide the development of an information mining project. Its main objective is to identify interesting patterns and relevant pieces of knowledge to the organization, ensuring their correct understanding, and providing reliable support for the decision-making process. To achieve this goal it is necessary to establish a set of activities that provide the overall project structure, supporting the development process. These guidelines should not only define the development aspects, but also it must identify those management activities associated with the project progress.

It is pointed out that the IME definition focuses on determining processes, and not in the specific technical characteristics of its implementation, such as defining the information mining processes [9] and the procedure to derive it from the business domain [10], identifying prospective families of data mining algorithms to implement and their combinations to obtain certain pieces of knowledge, rather than focusing on implementing a specific algorithm. Defining an IME is necessary to move from an artisanal development to a systematic, measurable, reproducible and disciplined development. The importance of these characteristics, are not only in the possibility of documenting the process, maintaining the traceability thereof, allowing to learn from it and reuse that knowledge to other projects, but also provides a set of tools that enables the mature development of the discipline, such as estimate time and cost, anticipating potential risks, needs and requirements that the project involves, among others.

In this context, the objective of this paper is to propose a new process model for IME projects. First, several approaches defining process model for IME are presented (section II). Section III presents the proposed solution, identifying the main structure of the management process. Section IV describes the existence gaps in the models previously mentioned in section II, identifying the solutions defined by the proposed model (Section IV). Finally, section V concludes the contributions of the proposed process model with respect to the previous approaches.

II. STATE OF THE ART

The first three sections briefly describe the structure of the most applied process models for information mining projects, which are focused on technical activities more related to data mining. These models are: KDD, SEMMA and CRISP-DM. Thereupon, a process model for software development projects oriented to the IME that is based on the best practices of software engineering projects is presented.

A. Knowledge Discovery in Databases (KDD)

The forerunner process model KDD [11], aims to provide a set of tools to systematize the data mining process and the artisanal process of hypothesis selection. The underlying objective to the concept of KDD is to differentiate data mining, understood as the activity of applying different algorithms on data in order to obtain patterns, than the process needed to generate knowledge from data. That is, understanding data mining as a subelement that integrates the general process of obtaining knowledge patterns. The steps involved in this process are: *Learning the application domain*: includes relevant

DOI reference number: 10.18293/SEKE2016-009

prior knowledge and the goals of the application; *Defining a target data set*: where discovery is to be performed; *Cleaning and Preprocessing*: includes basic operations such as removing noise or outliers, strategies for handling missing data fields, etc.; *Data reduction and projection*: involves finding useful features to represent the data depending on the goal of the task; *Choosing the function of data mining*: includes deciding the purpose of the model derived by the data mining algorithm; *Choosing the data mining algorithm(s)*: matches a particular data mining method with the overall criteria of the KDD process; *Data mining*: includes searching for patterns of interest through applying the data mining algorithms; *Interpretation*: includes interpreting the discovered patterns as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns, and translating the useful ones into terms understandable by users; and *Using discovered knowledge*: includes implementing or documenting the results.

B. SEMMA Methodology

The SEMMA [12] methodology was defined by the SAS Institute as a process used to reveal valuable information and complex relationships in large amounts of data. SEMMA consists of five stages: *Sample* the data by creating one or more data tables; *Explore* the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas; *Modify* the data by creating, selecting, and transforming the variables to focus on the model selection process; *Model* the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome; and *Assess* the data by evaluating the usefulness and reliability of the findings from the data mining process.

C. CRoss Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM [13] is a widely used methodology focused on the developmental tasks involved during the whole process, considering slightly some activities related to the management of the project. The phases that make up the process model are: *Business Understanding*: aims to comprehend the objectives and requirements of the project from the business perspective, as well as identifying the data mining problem(s) and perform the project planning; *Data Understanding:* starts with an initial data collection and proceeds with a data analysis, identifying data quality problems and possible interesting subsets to apply different hypothesis; *Data Preparation*: covers all activities needed to generate the final data set from the raw; *Modeling:* covers the selection, configuration and implementation of various modeling techniques; *Evaluation:* the built model is analyzed to ensure that the business objectives are accomplished; and *Deployment:* covers those activities oriented to integrate or to document the knowledge gained.

D. Process Model for Software Projects Focused on the Development of Information Mining Systems

In [14] an integrated approach is proposed, essentially composed by CRISP-DM and the standard to develop conventional systems IEEE Std. 1074 (and it also has some elements from ISO 12207 and 15504). This proposal pointed out which phases and activities from process models for software engineering project might be used for IME projects.

III. PROPOSAL OF PROCESS MODEL

After more than a decade using process models focused on development activities associated with data mining, and because of the high failure rate (greater than 60%) [15], being CRISP-DM the main methodology, several researches highlighted difficulties in the existent process model [16-18]. From the analysis performed over different existing proposals, arises the need to define a process model for IME projects, focusing on discovering relevant pieces of knowledge to support the decision-making process, which incorporates the management vision required to successfully achieve the development of the Project.

The proposed process model, called MoProPEI, consists of two sub-processes: development and management. The development process is a version based on CRISP-DM which reduce iteration necessity, modifying the order and structure of the activities, and includes a set of tasks (such as business and problem modeling, deriving the information mining process, etc.) in order to facilitate the comprehension and implementation of the project as well as quality level of the resultant products. The management process for IME projects provides a transversal layer to the development process, which offers several mechanism that improve the quality and maturity level of the resultant product(s) and process. Additionally, this sub-process incorporates different tools developed over the last years by the research group in order to cover the existence gaps. Since the beginning of the discipline, there has been an extensive effort to develop tools or techniques belonging to the development process. Taking this into account, this paper focused on describing the management process, identifying some techniques developed by the research group (section A), and explaining in detail its structure in section B.

A. Management Tools for IME

Because of the differences among IME and other projects, emerge the need to define a set of *ad hoc* tools / techniques that cover the existent conceptual gaps. During the last years, the research group has been working in a set of tools to resolve such necessity, among them:

Britos et Al. [19] describe a framework which defines tasks required for elicitation of requirements, and propose a reference model for identification and documentation of relevant concepts to carry out an IME project, along with the correlations between these concepts. This process consists of five steps which cover different aspects of the project, which are implemented throughout the process: in the phases of business understanding and modeling belonging to development process and in the initiation phase from management process.

Pytel et Al. [20;22] define cost and effort estimation method for information mining projects. This process describes the characteristics, and analyzes the possible values over which the effort and cost required to complete the project are obtained. These features are divided into three categories according to the origin of data to be analyzed: factors related to type of project, data and resources. This tool can be applied in

estimation and responsibilities activity, belonging to the planning phase from the management sub-process.

In [21;22] a method oriented to determine the project feasibility is described. This procedure identifies the set of characteristics to evaluate, which are classified in three categories: plausibility, adequacy and success. From these criteria it is determined whether the project is feasible, if IME is the best solution to the problem and whether the results can be useful for the organization. This tool can be applied in the assess situation activity (initiation phase).

Basso et Al. [23] define a set of metrics applicable to IME projects, grouped by three categories: data, models and project. In each category identifies tasks and characteristics to measure and their associated metrics. This tool can be applied in the estimation and responsibilities activity.

B. Management Process

The management process occurs as a supporting element to development process, in which the execution of its tasks is not linear, but is made based on project progress. This sub-process, structured by three additional levels (phases, activities and tasks), provides a higher level of granularity. It consists of five phases, each one composed by several activities (also referred as unity in this paper), identifying a set of tasks with a common goal. Figure 1 shows the phases that comprise it (at the left of the bracket in the picture) and for each phase the activities associated and their existing dependencies (in curved rectangles and lines respectively). In the top and bottom of the figure, there are activities from the development process, and inside each one is identified the phase and activity name which belongs (upper and lower section, respectively). Arrows represent interaction between two activities (dependent documents). There are also two types of lines ending in point: solid and dot, both represent an input document to an activity, which is produced by the team or the client respectively.

The first phase, named *Initiation*, contains four activities: "Communication Protocol" aims to determine formal channels of interaction, according to the actual possibilities among staff and experts in the business domain. Its inputs are customer speech (external), politics of contracting organization (external), politics of organization (internal) and project leader speech (internal), and produce as output the Communication Protocol Report (CPR); "Exploration of Initial Concepts" where the team makes a first approximation to the client characteristics and their needs. As result of the activity, initial exploration report (IER) is made, using as input the customer speech and the CPR; "Assess Situation" analyses the project characteristics, client requirements, existent risks and possible contingencies to determine the feasibility of the project. Its inputs are: existent resources (external), previous experiences of projects (internal) and IER. It generates five elements as results: guidelines projects report (GPR), internal and external resources reports, and project risks and feasibility report; and "Life Cycle Definition" determining the structure and iterations of the project, in order to encourage the development and success of the project. The requirements document (RD), generated in the activity business problem comprehension, and GPR are inputs over which the life cycle model is selected.

Planning is the second phase, composed by three activities: "activities planning", whose inputs are: the life cycle selected, the RD and GPR. Its aims are to select and to determine the time required to finish each task of the project, producing as outputs the activity calendar and map, and the list of metrics; "resources planning" whose inputs are activity calendar, business problems (identified in the activity business problem comprehension), RD, internal resources report and GPR. Its main purpose is to determine what resources are required and when (including outsourcing possibility), generating as outputs: the outsourcing plan, material resources plan, human resources plan and resources required report (RRR); "estimation and responsibilities" where the time and cost of the project, and the scope and obligations of the parties are determined from the RD, business objectives (generated in the business analysis activity belonging to the development process), report of external resources, project risks and feasibility reports, GPR, business problems, activity map, outsourcing plan and RRR.

Support composed by three activities: "Life Cycle Management" whose objectives are: to formalize the scope of each project iterations, and to establish the reached achievements, readjusting the scope of the next iteration if necessary. Their inputs are: the report of external resources, life cycle selected, material resources, outsourcing and HR plans and activity calendar and their outputs are: formalization of cycle start and formalization of cycle end; "Implementation Management" covers those activities related with the development of the product, such as defining the objectives to be performed in each step of the project, integrating internal products, communicating project progress, etc. Its inputs are: staff list (internal), outsourced product (external), project quality report, occurred risks report, activity progress report (these three last elements produced in the activities control unity), control of outsourced tasks (from Resources Control), business problem risks related (from Business Problem Understanding, development process), contingency plan, activity map, outsourcing plan and RRR, and produce as outputs: outsourced work Integration report (OWIR), corrective actions implementation report (CAIR), staff responsibilities report (SRP), human and material resources contracts, list of acquired resources (HR and materials), outsourced tasks report and outsourced tasks contract; and "Configuration Management" which aims to document the relevant information produced throughout the project, using as input: OWIR, CAIR, and formalization of cycle opening and ending.

The *Controlling and Quality* phase involves four activities oriented to improve the maturity of the project and its products, and keep tracking of all the elements required to complete successfully the project: "Resources Control" whose aim is verifying the accomplishment of the planning made, regarding the incorporation of resources (materials, products and human). Its inputs are: outsourced product, business problems and objectives, SRP, human and material resources contracts, list of acquired resources (HR and materials), outsourced tasks report, outsourced tasks contract, Outsourcing Plan and RRR; "Measurements" quantifies elements of interest for the project analysis, using as input the metrics list and the cost of the project activities (internal), generating the reports of metrics and costs.

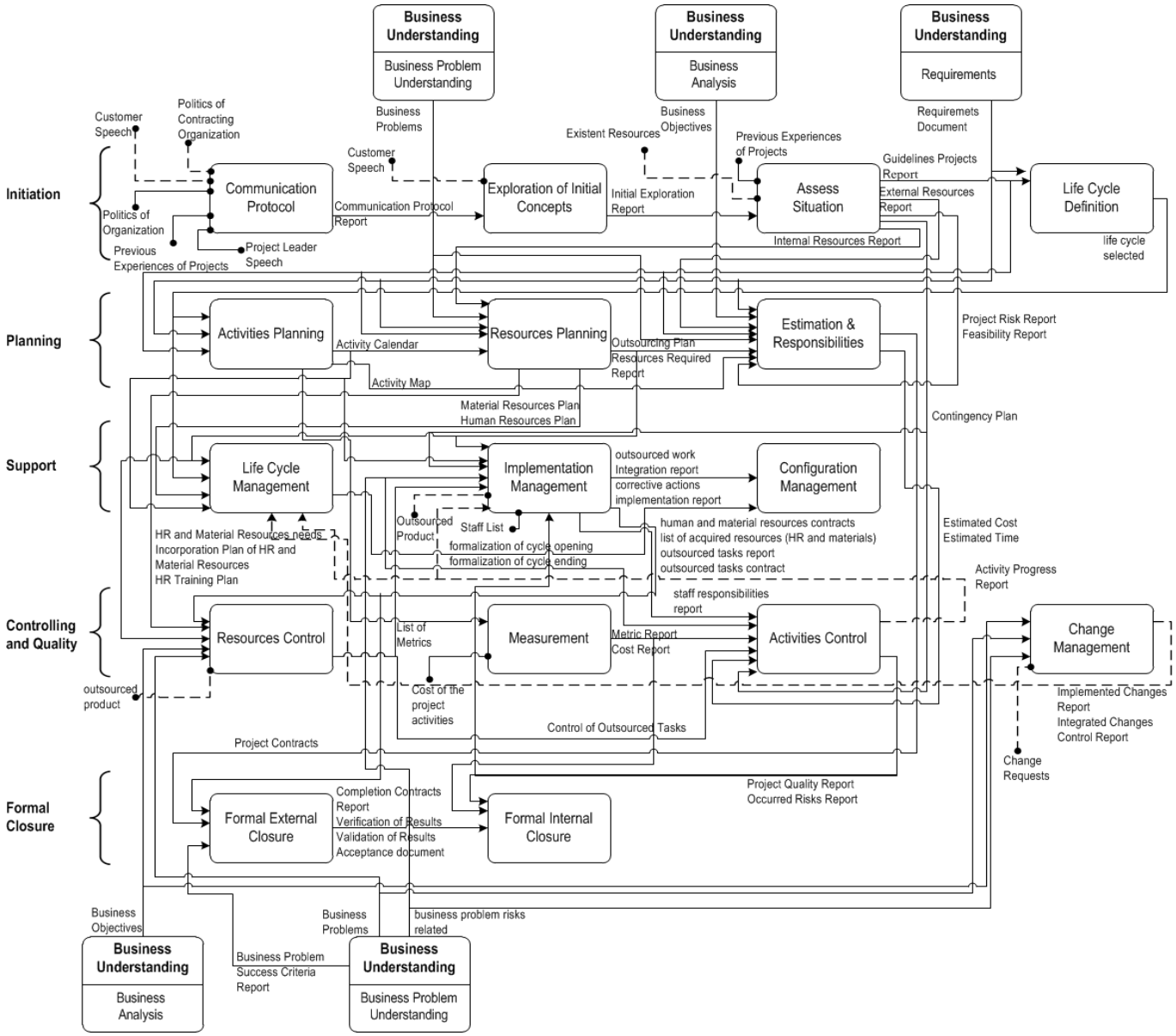

Figure 1. General Structure of the Management Process

"Activities Control" evaluates several project characteristics (progress of activities, time, cost, possible risks, quality, etc.) in order to identify possible deviations that jeopardize the successful development of the project. Its inputs are: SRP, business problem risks related, reports of metrics and cost, control of outsourced tasks, cost and time estimations and contingencies plan, producing as outputs: the reports of project quality, project risk and activity progress; and "Changes Management" which covers all activities aimed at evaluating, implementing and monitoring the changes required for the project, using the following inputs: change requests, business problems, business problem risks related and business objectives, and generates implemented changes and integrated changes control report.

The last phase *Formal Closure* is composed by two activities: "Formal External Closure" aims to ensure the achievement of the objectives required by the customer, ending with the commitments between the parties involved in the project. Its inputs are: business problem success criteria report (from the business understanding problem activity which belongs to the development process), project contracts, human and material resources contracts, list of acquired resources, outsourced tasks report and outsourced tasks contract, and generates as outputs: completion contracts report, Verification and Validation of Results, and acceptance document; and "Formal Internal Closure" whose objective is to obtain useful knowledge from the characteristics analyzed throughout project, in order to be reused in future ones. Its inputs are: the outputs from the previous activity, metric and cost report, project quality report and occurred risks report.

IV. DISCUSSION

The proposed models, mainly focused on developmental activities, have more than fifteen years in the market and have

been applied in large quantities and varieties of projects. In recent years many issues and deficiencies have been identified. Neither KDD, nor SEMMA nor CRISP-DM define a process model focused on the management and control of the project, from this point of view our proposal attempt to cover this vacancy. Furthermore, both KDD, as SEMMA do not consider activities targeted to project management, only focusing on the developmental activities associated with business and data understanding, data selection and processing, implementation of data mining algorithms and analysis of results as pieces of knowledge. Although CRISP-DM identifies a set of activities related to project management, including: risks and contingencies, produce project plan, cost and benefit, Plan Deployment, Plan Monitoring and Maintenance and Review Project, it lacks of others extremely important features associated with the maturity level and success rate of projects.

Marban et Al. [14] define a scheme that identifies from the missing activities of their predecessors, which Software Engineering activities could be applied, however they do not clearly defined their objectives, scopes, inputs/outputs and possible techniques according to the objective of IME. Our proposal differs from this approach in that it is conceived considering as the general objective of the project: extracting reliable and relevant pieces of knowledge to support the decision-making process [8; 11; 24], and it attempts to deepen those aspects that are not defined in Marban's proposal (describing in detail processes, phases, activities, tasks, inputs/outputs, techniques and their goals). Table I summarize the gaps in the previous processes, identifying whether the set of activities require for a IME project, grouped by the phases proposed in MoProPEI, are full, partially or not covered.

Table I. Comparison among processes models: management perspective

References: ■ Full ▒ Partially □ No		PROCESS				
		KDD	SEMMA	CRISP-DM	MARBAN	MoProPEI
ACTIVITIES	Initiation	□	□	▒	▒	■
	Planning	□	□	□	▒	■
	Support	□	□	□	■	■
	Controlling and Quality	□	□	□	▒	■
	Formal Closure	□	□	▒	▒	■

Due to CRISP-DM is *de facto* standard in industry [14], holding for the past 10 years its presence as the main process model used by companies and data analysts [25] and the most robust approach, in the next paragraphs we describe the novelties with respect to CRISP-DM, identifying what activities from MoProPEI pretend to cover each of the aforementioned gaps:

Project Feasibility Analysis: determining success of the project at an early stage, avoiding taking risks which cannot be overcome or their cost is too high. The feasibility method introduced in section III.A, provides a guide that helps to make the decision of continue or not with the project according to its characteristics. This method can be implemented in the activity *assess situation* in initiation phase.

Planning metrics: taking measurements and analyzing the results promotes the comprehension of project progress and product as well as its quality, using these results to improve the process to be performed and their resultant product(s). This deficiency can be resolved in the activity: activities planning by identifying and listing metrics applicable, implementing the tool for metric described in section III.A.

Planning and controlling project costs and time: monitoring these tasks allow to keep the fluctuation of these variables between preset values, being able to take early action if necessary. The activities: Estimation and Responsibilities, and Measurement, belonging to the phases "planning", and "Controlling and Quality" respectively, consider these problems by the techniques: plan of cost and metrics and the report of metrics and cost.

Project documentation Management: due to the dynamic and iterative features of IME projects, controlling the generated documentation is essential in order to identify current version over what each member of the team is working on as well as maintain traceability of documents produced. Additionally, it enables the data analyst to reproduce any result, being able to return to a previous state of the project, while providing a higher level of maturity and quality in the project. Finally, controlling and generating new ideas from the experiences documented are simplified. The technique integrated changes control report (belonging to the change management activity) covers those problems.

Defining the responsibilities and obligations of the parties involved in the project: make a legal commitment between the parties involved in the project provides support for both sides, and also places on record and clarifies the scope and limitations of the project. This problem is resolved in the Estimation and Responsibilities activity, where the contracts required for the project are defined.

Analysis and selection of life cycle model: engineers identify the possible alternatives and choose the best option according to the characteristics of the project and information mining team. The project structure, order of execution and interaction between stages are defined, adjusting the time and cost to the project needs. The "life cycle definition" activity takes into consideration this problem.

Analysis and selection of the activities associated with the project: the engineers define activities and tasks necessary to successfully perform the project, according to the characteristics of the business domain, client requirements and information mining team. The "Planning activity", which belongs to the planning phase, helps to define the structure of the project (by activities map technique), providing better support for planning and development thereof, reducing its time and costs.

Identify and manage the outsourcing possibility: according to the characteristics of the project and the team, some parts or the whole project can be outsourced. Outsourcing involves performing a set of tasks and controls to ensure the project success. These tasks are considered throughout the sub-process

by the activities: assess the situation, where the outsourcing possibility is determined by generating the report possibility of outsourcing; Estimation and Responsibilities, where the scope and obligations of the parties under contracts are defined; and resources control and configuration management, where the team tracks the progress of outsourced item and performs the integration of the products obtained.

Planning and Control the resources distribution: anticipate human and material resources needs in different sections of the process, improves the development of the project. This problem can be solved applying the techniques: report of required resources and report of resource acquisition, implemented in the activities resource planning and resource control respectively.

V. CONCLUSIONS

The contributions provided by the proposed process against previous approaches are:

[i]. We proposed a process model composed by two sub-processes (development and management). The management sub-process is conceived as a transversal layer to the development sub-process, on which the discipline has been focused on throughout its history, and its aim is to cover the current gaps in the areas of control and project management.

[ii]. The process integrates a set of tools created *ad hoc* [9;10;19-23], oriented to reduce the risk and improve the maturity and quality level of the process and the resultant product(s).

[iii]. The structure of management sub-process is described, identifying the set of activities required for IME project, detailing each activity objective and dependencies (inputs and outputs). The activities are grouped into phases, from the similarities in their goals for the overall process.

We identify as future research work:

[i]. Extend the set of available techniques for different management activities, such as for the phase of requirements elicitation and analysis and evaluate how the proposed model can be used to reduce risks and improve productivity.

[ii]. Developing experiments to compare effectiveness of the proposed process model against CRISP-DM.

REFERENCES

[1] Maimon, O. y Rokach, L. (Eds.). 2005. Data mining and knowledge discovery handbook. Springer.

[2] Tan, A. 1999. Text mining: The state of the art and the challenges. In Proc. PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases. pp. 65-70.

[3] Hsu, W., Lee, M., Zhang, J. 2002. Image mining: Trends and developments. Journal of Intelligent Information Systems, 19(1): 7-23.

[4] Gaber, M., Zaslavsky, A. Krishnaswamy, S. 2010. Data stream mining. En Maimon, O. and Rokach, L. eds. Data mining and knowledge discovery handbook. Springer, pp. 759-787.

[5] Kosala, R., Blockeel, H. 2000. Web mining research: A survey. ACM SIGKDD Explorations Newsletter, 2(1): 1-15.

[6] Gopal, R., Marsden, J., Vanthienen, J. 2011. Information mining: Reflections on Recent Advancements and the Road Ahead in Data, Text, and Media Mining. Decision Support Systems, 51(4): 727-731.

[7] Abran, A., Moore, J. W., Bourque, P., Dupuis, R., Tripp, L. 2004. Guide to the Software Engineering Body of Knowledge (2004 version). IEEE. ISBN 0-7695-2330-7.

[8] García-Martínez, R., Britos, P., Pesado, P., Bertone, R., Pollo, F., Rodríguez, D., Pytel, P., Vanrell. J. 2011. Towards an Information Mining Engineering. In Software Engineering, Methods, Modeling & Teaching. Medellín University Press. ISBN 9789588692326. pp. 83-99.

[9] García-Martínez, R., Britos, P., Rodríguez, D. 2013. Information Mining Processes Based on Intelligent Systems. Lecture Notes on Artificial Intelligence, 7906: 402-410. ISBN 978-3-642-38576-6.

[10] Martins, S., Rodríguez, D., García-Martínez, R. 2014. Deriving Processes of Information Mining Based on Semantic Nets and Frames. LNAI, 8482: 150-159. ISBN 978-3-319-07466-5.

[11] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. 1996. From data mining to knowledge discovery in databases. AI Magazine, 17(3): 37-54.

[12] SAS Institute Inc. 1998. SAS Institute White Paper, From Data to Business Advantage: Data Mining, The SEMMA Methodology and the SAS® System, Cary, NC: SAS Institute Inc.

[13] Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 2000. CRISP-DM 1.0 Step by step BI guide.

[14] Marbán, O., Mariscal, G., Menasalvas, E., Segovia, J. 2007. An Engineering Approach to Data Mining Projects. Lecture Notes in Computer Science, 4881: 578-588. Springer.

[15] Gondar, J.E. 2005. Data Mining Methodology. Data Mining Institute. ISBN: 978-84-96272-21-7.

[16] Wirth R., Hipp J. 2000. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, pp. 29–39.

[17] Yang Q., Wu X., 2006. 10 Challenging Problems in Data Mining research, International Journal of Information Technology and Decision Making 5(4), pp. 597–604.

[18] Lavrac N., Motoda H., Fawcett T., Holte R., Langley P., Adriaans P., 2004. Lessons Learned from Data Mining Applications and Collaborative Problem Solving. Machine Learning. 57. 13–34.

[19] Britos, P., Dieste, O. and García-Martínez, R., 2008, in IFIP International Federation for Information Processing, Volume 274; Advances in Information Systems Research, Education and Practice; David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, Dewald Roode; (Boston: Springer), pp. 139–150.

[20] Pytel, P., Britos, P., García-Martínez, R. 2013. A Proposal of Effort Estimation Method for Information Mining Projects Oriented to SMEs. Lecture Notes in Business Information Processing, 139: 58-74. ISBN 978-3-642-36610-9.

[21] Pytel, P. Britos, P., García-Martínez, R. 2013. Proposal and Validation of a Feasibility Model for Information Mining Projects. Proceedings 25th International Conference on Software Engineering and Knowledge Engineering. pp. 83-88. ISBN 978-1-891706-33-2.

[22] Pytel, P., Hossian, A., Britos, P., García-Martínez, R. 2015. Feasibility and Effort Estimation Models for Medium and Small Size Information Mining Projects. Information Systems Journal, 47: 01-14. Elsevier. ISSN 0306-4379.

[23] Basso, D., Rodríguez, D., García-Martínez, R. 2013. Proposal of Metrics for Information Mining Engineering Projects (In Spanish). Workshop on Data Bases and Data Mining. Proceedings XIX Argentine Congress of Computer Science. pp. 983-992. ISBN 978-987-23963-1-2.

[24] Witten, I. H., Frank, E., Hall, M. 2011. Data mining: Practical Machine Learning Tools and Techniques. 3rd ed. ISBN 978-0-12-374856-0.

[25] Kdnuggets. 2014. What main methodology are you using for your analytics, data mining, or data science projects? Poll (Oct 2014). http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html (last access 10/02/2016).