ISSN 2007-9737

A Representation Based on Essence for the CRISP-DM Methodology

Claudia Elena Durango Vanegas[1], Juan Camilo Giraldo Mejía[2], Fabio Alberto Vargas Agudelo[2], Dario Enrique Soto Duran[2]

[1] Universidad de San Buenaventura, Facultad de Ingeniería, Medellín, Colombia

[2] Tecnológico de Antioquía, Facultad de Ingeniería, Colombia

claudia.durango@usbmed.edu.co, {jcgiraldo1, fvargas, dsoto}@tdea.edu.co

Abstract. CRoss Industry Standard Process for Data Mining (CRISP-DM) is a data mining project development methodology that establishes tasks and levels of abstraction, hierarchically structured to facilitate its implementation through a set of actions that help in making decisions. Essence is a theory that helps identify best practices and essential, common, and universal elements to all endeavor in the software development cycle. In the literature, there are different models of representation of the CRISP-DM methodology, such as verbal model, conceptual model, process understanding model, and ontology. However, it considered that these representation models lack the incorporation of some elements, such as, activities, work products, and roles of the CRISP-DM methodology. In this paper we propose a representation based on Essence of the CRISP-DM methodology, incorporating the essential elements that we believe are missing from existing representations. With the representation in Essence that is proposed, the aim is to improve the understanding of best practices and the essential, common, and universal elements of the CRISP-DM methodology for future implementations in data mining projects. In addition, it seeks to validate that Essence can be used in different of data mining projects.

Keywords. CRISP-DM methodology, data mining, representation model, essence.

1 Introduction

CRoss Industry Standard Process for Data Mining (CRISP-DM) is a methodology that is grouped into activities that are described in four levels of abstraction, ranging from general to specific: phases, generic tasks, specific tasks, and processes instantiated in life cycle of data mining projects.

CRISP-DM contains phases, tasks, and relationships between tasks, where phases are neither rigid nor consecutive, but are flexible, discontinuous, and repetitive throughout the project's life cycle. Similarly, CRISP-DM defines tasks, products of these tasks, terminologies, and it characterizes types of data mining problems.

CRISP-DM has six phases: business understanding, data understanding, data preparation, modeling, evaluation, and development. These phases contain a set of actions that help users to make decisions in data mining projects [1–3].

Essence is a theory that finding a kernel widely used and accepted elements, and redefine a solid theoretical base supported by proven principles and best practices of different methods of software development.

Therefore, Essence is considered to generalize software engineering because it identifies universal and common actions and elements of the software life cycle.

These elements help describe different methods of software development for using simple and universal language. Furthermore, Essence allows evaluating, comparing, and measuring the efforts and essential, common, and universal

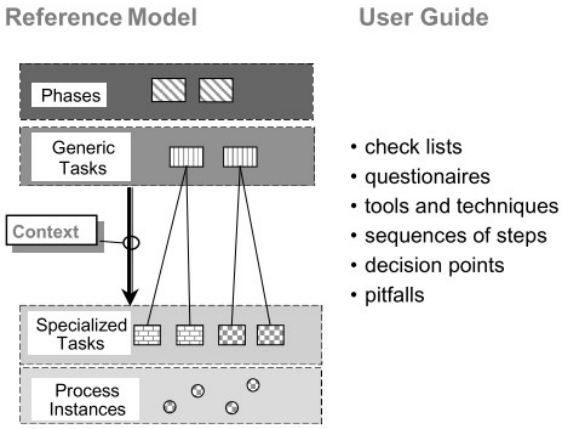


Fig. 1. Representation of four levels of the CRISP-DM methodology [3]

elements of the best practices identified in the software life cycle [4–6].

In the literature there are different representation models of the CRISP-DM, such as: verbal model [7], conceptual model [8], process understanding models [1–3], and ontologies [9], among others.

However, we are observed that these representation models lack incorporating some elements, such as: activities, work products, and roles of the CRISP-DM methodology.

For this reason, in this paper we propose a representation of the CRISP-DM methodology based on Essence, incorporating essential elements considered necessary in the existing representations.

Representations proposed seek to improve the understanding of best practices and essential, common, and universal elements of the CRISP-DM methodology for future implementations in data mining projects life cycle.

In addition, representations seek to validate that Essence can be used in different data mining projects.

The paper is organized as follows: firstly, theoretical framework is presented, where a conceptualization of CRISP-DM and Essence. Next, background related to representations of CRIPS-DM found in similar works is presented. Then, proposal for the representation of CRISP-DM in Essence. Finally, discuss conclusions and future work.

2 Theoretical Frameworks

This Section introduces key concepts CRISP-DM methodology and Essence.

2.1 CRISP-DM Methodology

CRISP-DM is a methodology approved of oriented data mining projects. It is considered a methodology because it includes description of phases, tasks, and relationships between tasks, and it is considered a process model because it offers a summary of the data mining projects life cycle. Therefore, CRISP-DM is considered to provide a standardized description of the data analysis projects life cycle. Where, phases of the CRISP-DM methodology are flexible and customizable, because its allows interactivity between phases and personalization of the result according to the selection of the phase or task that will be carried out later [2, 3, 10].

CRISP-DM methodology is described in terms of a hierarchical process model, made up of four levels of abstraction ranging from general to specific.

These levels are phase, generic task, specialized task, and instantiated process (see **Fig. 1**). In the first level is phase, each phase consists of several secondary levels of generic tasks. In addition, it is identified that data mining process is composed of a small number of phases. At the second level are generic tasks that it can be called complete or established according to the possibility of being performed.

At the third level is specialized task that it describes the actions that the generic tasks must perform according to the types of situations. At the fourth level is process instance that represents a record of actions, decisions, and results of a real data mining engagement. A process instance is organized according to the tasks defined at the higher levels, it represents what happens in a particular commitment, instead of what happens in a general commitment [1, 3, 11].

2.2 Essence

Essence is a theory that arises from a call to action by Jacobson, Meyer, and Soley in 2009. The call

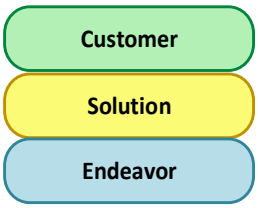


Fig. 2. Essence: Areas of Interest [5]

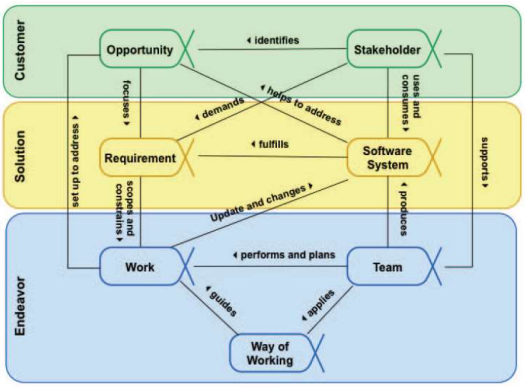


Fig. 3. Alpha: Things we always work with [5]

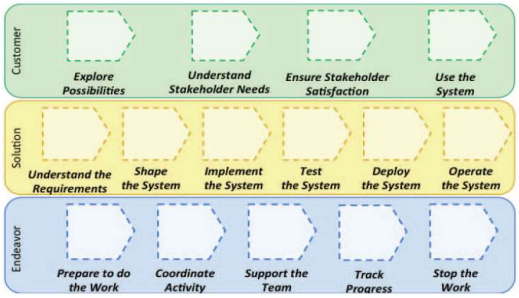


Fig. 4. Activity space: Things we always do [5]

originates from the identification of problems in software engineering due to the lack of a solid and widely used theoretical base, the existence of many methods and variants of software development methods, and the separation of academic and research practices from the industrial or production sector. Due to the above, Essence is defined as a solid theory, with proven principles and best practices that support the process of redefining software engineering.

In addition, it is identified that Essence has a simple language that we help to represent different methods and best practices in different areas of knowledge [5]. Essence is organized into three areas of interest (see Fig. 2) that focus on a specific dimension of the software life cycle [6, 12]:

– Customer: involves the customer of the software system, allowing to know their perspectives to ensure the development of an appropriate solution.
– Solution: involves everything related to the specifications and development of the software system, seeking to solve the problem.
– Endeavor: it involves everything related to the development team and way the work is done.

Essence includes a group of elements that represent “things we always work with” called Alpha (see Fig. 3) and “things we are always do” called Activity Spaces (see Fig. 4) in the software systems life cycle [5].

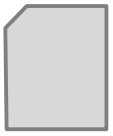
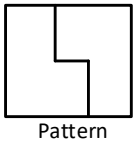
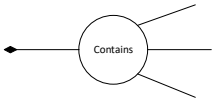
In Table 1, we present main elements of the Essence that we are used in the representation of the CRISP-DM methodology.

3 Related Work

Zapata and Gil (2011) propose to incorporation of two software engineering diagrams in the first phase of the CRISP-DM methodology: preconceptual schemes, and goal diagram. Authors propose incorporating four additional activities in the first phase: Develop preconceptual schema (see Fig. 5), for including the achievement verbs within preconceptual schema relationships, develop objective diagram (see Fig. 6) and choose the variables that participate in the model.

Authors therefore seek to improve the semi-formal validation of the information required in the project, increasing stakeholder participation, and making a less subjective choice of the variables *involved in the project. Therefore, the authors* consider that incorporation of the two work products manages to identify organizational needs, variables of interest of the model required for the data mining process, and with the use of semi-automatic techniques contributes to the standardization of the data mining process of the organization. As a result, authors obtain a preconceptual schema that incorporates the analysis of the verbal model of the organization and with goal diagram we can visualize hierarchy and assignment of the importance of the objectives in the model that is made [7].

Table 1. Principal elements of Semat Essence kernel [13]

Description	Symbol
Alpha is element that represent the things that need to be tracked in terms of progress and health to guide the effort to successful completion.	
Activity space is an element that complement the alphas. In addition, they represent the "things we always do", providing insight into software engineering activities.	
Work product is an artifact of value and relevance to the software engineering effort. A work product can be a document or part of the software.	 Work product
Activity is used to define one or more types of work items and orient how to do them.	
Pattern is a description of a structure in a practice. Phases and roles can be represented with the symbol.	 Pattern
Association pattern is used to connect the pattern with its associate elements.	
Practice is a repeatable approach to achieving a specific goal.	

This representation identifies important variables of interest and standardizes the organization's data mining process, reducing the participation of data mining experts in the first phase of the CRISP-DM methodology.

However, the proposed work products lack other important elements of the CRISP-DM methodology to improve their implementation and monitoring, such as the activities to be implemented and the work products resulting from the application of the CRISP-DM methodology.

Sharma and Mansotra (2016) propose a conceptual model called the Data Mining Model to Manage the Public Health System (DM-PHCS). Conceptual model focuses on identifying state of data mining and decision-making, roles of different process officials (administrator, data mining expert, and end users), and data mining methods (association, segmentation, and classification) where selection depends on the requirements of the public health system.

Therefore, conceptual model is considered for allowing too easy application by the administrator with only a basic level of understanding of data mining concepts, which allows them to enable an understanding of data mining concepts for correct interpretation of the model.

DM-PHCS consists of eight phases: domain understanding, data sources, data understanding, data preparation, data mining techniques, model

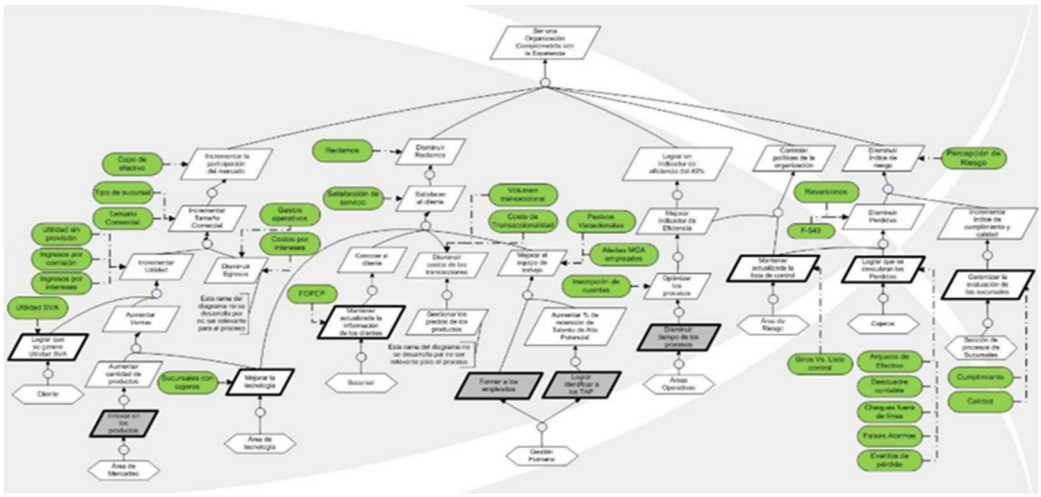


Fig. 5. Objective diagram resulting from the analysis of the preconceptual schema [7]

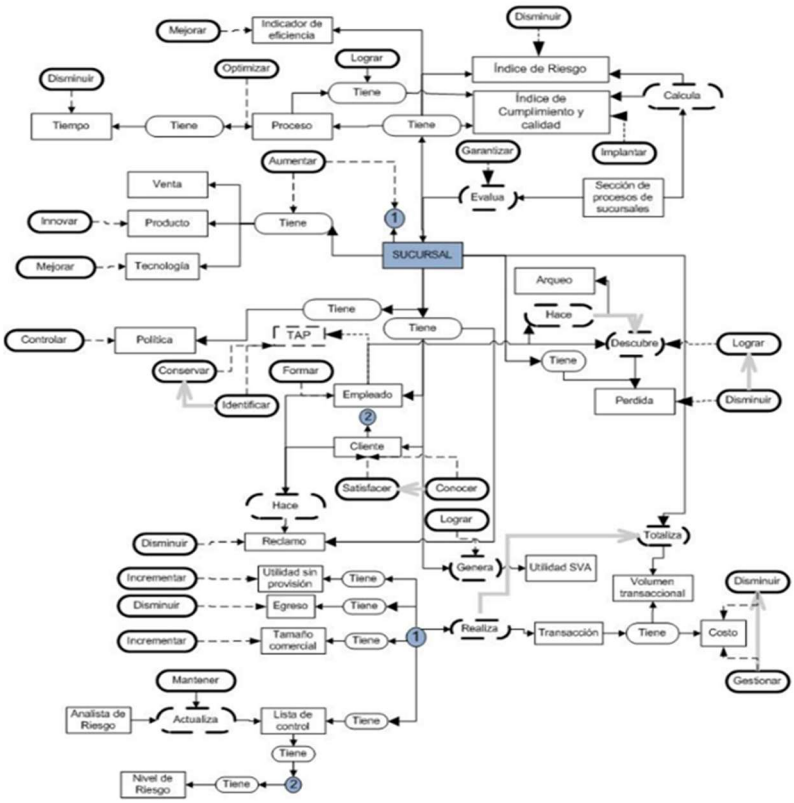


Fig. 6. Preconceptual scheme of the analysis of the verbal model of the organization [8]

construction, knowledge base, and model evaluation (see Fig. 6).

However, although authors define roles, patterns, and relationships are considered in the

decision-making process, these elements are not clearly identified in the phases or conceptual model.

Panov *et al*. (2008) present an ontology called OntoDM, seeking a formal representation in the field of data mining. OntoDM is based on the general framework and contains basic definitions of CRISP-DM data mining, such as: data type (primitive and structural), dataset, data mining tasks (predictive model, pattern discovery, probability distribution grouping, and estimation), generalizations (predictive model, patterns, clustering, and probabilistic distribution), data mining algorithms, data mining algorithm components (distance, core, and feature functions), and constraints (evaluation and language).

OntoDM representation (see Fig. 8) presents the <dataset> class as an extension of maximum level of the <aggregate> class and has two part of <data_example> and properties has_information <dataset_structure>.

The class <dataset_structure> presents information about dataset characteristics, such as: has number <number_of_data_examples> and has number <number_of_attributes>. Attributes are quality elements in the set of information related to the class <dataset> and it expresses with the property has_quality <attribute>.

Each attribute has a specific value that expresses with relation has_information <attribute_valued>, also, attributes have different roles in the set of information as what defines the class <attribute_role>.

The class <attribute> expresses the type of fact of the attribute and it is defined that a property has_information with the class <datatype>. Because a <datatype> can have a complex structure, ontology allows to handle information structured with the class <structured_datatype> connected with <datatype_constructor>. Of equal way, the <datatype> can have a primitive structure [9].

This representation is based on identifying classes, relationships, attributes, roles, structure, and type of data constructor, failing to represent phases, activities, and work products that must be performed to implement CRISP-DM methodology.

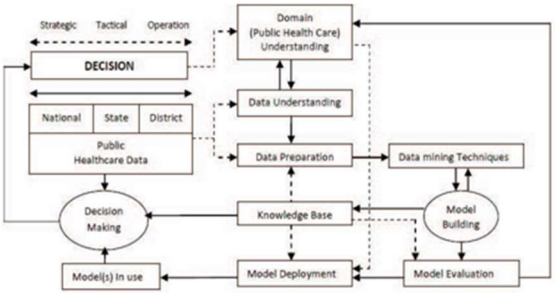


Fig. 7. Conceptual model diagram: DM-PHCS [8]

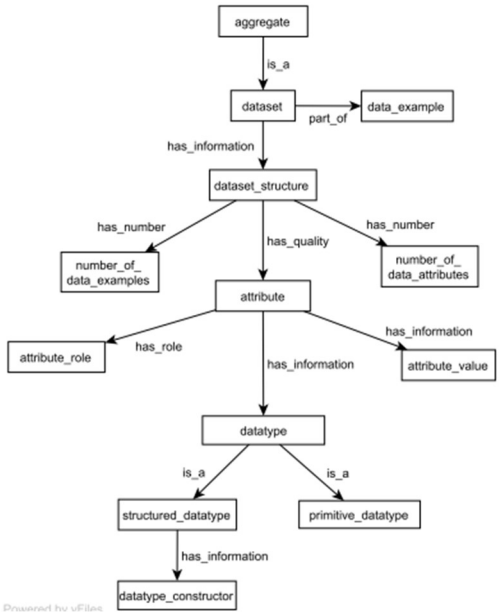


Fig. 8. Ontology: OntoDM [9]

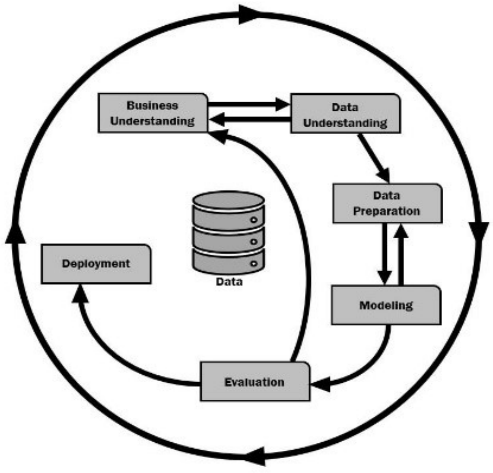


Fig. 9. Phases CRISP-DM Methodology [7]

4 Semat Kernel Representation of the CRISP-DM Methodology

The model of reference of CRISP-DM methodology for data mining originates from the review of data mining projects, identifying that it is composed of six phases: business understanding, understanding data, data preparation, modeling, evaluation, and development (see Fig. 9). Generally, these phases are developed sequentially and in constant iterations by the data analyst [14]. These phases are described below with their main elements [3, 7, 10, 11].

– Business understanding: this phase focuses on understanding objectives and requirements of the project from a business perspective. The activities that are performed are set the business objectives, evaluate the problem, set the objectives of mining data, and generate the project plan. As work products it is obtained description of the business with the targets of the business, context of the business, targets of the informational advancing, criteria of evaluation of the informational advancing, plan of the project and initial evaluation of hardware and skills of informational advancing.

– Understanding data: this phase focuses on understanding the collection and processes of data, through activities that seek to become familiar with data, identify data quality issues, build a first idea of data, or detect interesting subsets of data to form hidden data hypotheses. The activities include collecting initial data, describing data, exploring data, and verifying data quality. The work products obtained are initial data reporting, data description reporting, data exploration reporting, and data quality reporting.

– Data preparation: this phase focuses on preparing data for the construction activities of the final dataset to be incorporated into the modeling tools. Similarly, data preparation tasks are identified as repetitive and unordered. The activities that are performed are select the data, clean up the data, build the data, integrate the data, and format the data. The work products obtained are list of data inclusion and exclusion rules, clean data

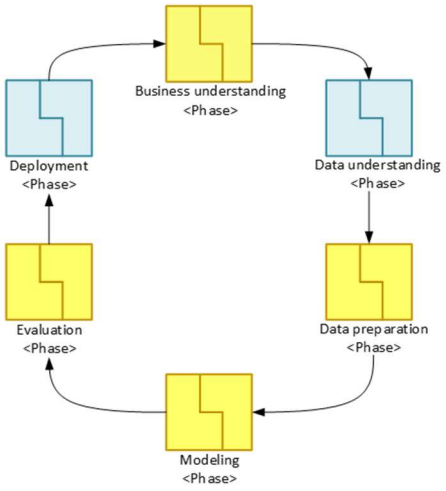


Fig. 10. Representation in Essence of phases CRIPS-DM methodology

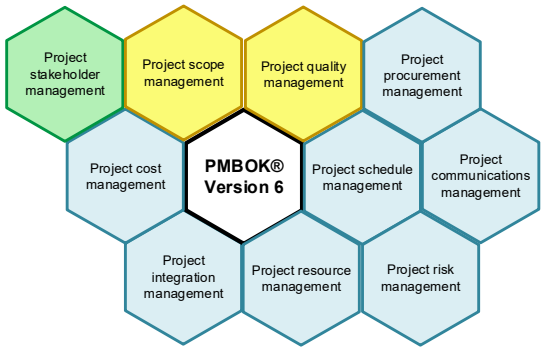


Fig. 11. Best practices PMBOK® represented in Essence [13]

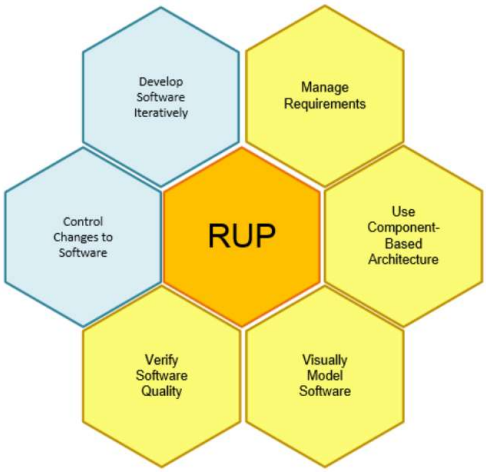


Fig. 12. Six RUP best practices in Essence [17]

Table 2. Identification of activities and work products for each phase of the CRISP-DM methodology

Phase	Activity	Work product
Business understanding	Determine business goals	Business description
	Assess situation	Business context
	Determine data mining goals	Data mining goals
		Data mining evaluation criteria
	Produce Project plan	Project plan
		Initial evaluation of datamining tools and techniques
Data understanding	Collect initial data	Initial data report
	Describe data	Description data report
	Explore data	Data exploration report
	Verify data quality	Data quality report
Data preparation	Select data	List of data inclusion and exclusion rules
	Clear data	Clean data report
	Construct data	Attributes and derived records
	Integrate data	Integration algorithms
	Format data	Formatted data report
Modeling	Select modeling techniques	Modeling techniques and assumptions
	Construct model	Parameter setting
		Model and model description
	Determine design test	Design test
	Evaluate model	Model evaluation with adjustments
Evaluation	Evaluate results	Model results report
	Review process	Approved model
	Determine next steps	Process review report according to business objectives
		List of actions to follow
Deployment	Develop implementation plan	Implementation plan
		Deployment plan
	Monitor and maintain plan	Monitor and maintain plan
	Produce final report	Final report
	Review project	Documentation of the experience

reporting, attributes and derived records, integration algorithm, and reformatted data.

– Modeling: this phase focuses on selecting and applying data modeling techniques. Similarly, parameters are calibrated to obtain the optimal values of the data. The activities that are performed are select the modeling technique, generate design tests, build the model, and evaluate the model. The work products obtained are modeling techniques and assumptions, test design, parameter configuration, model and model description, and model evaluation report with settings.
– Development: this phase focuses on organizing and presenting the acquired knowledge so that the client can use it. This phase identifies that deployment complexity can be simple, such as report generation, or complex, such as implementing a repeatable data mining process. The activities that are carried out are develop the implementation and deployment plan, monitoring and maintenance plan of the previous plans, final report, and documentation of the experience. The work products obtained are development plan, monitoring and maintenance plan, final production report, and project documentation.

The representations presented in the previous Section seek to identify organizational needs, variables of interest of the data mining method, main uses of semi-automatic techniques, state of data mining and decision-making, of abstraction levels, and understanding of the understanding basic definitions of the CRISP-DM methodology.

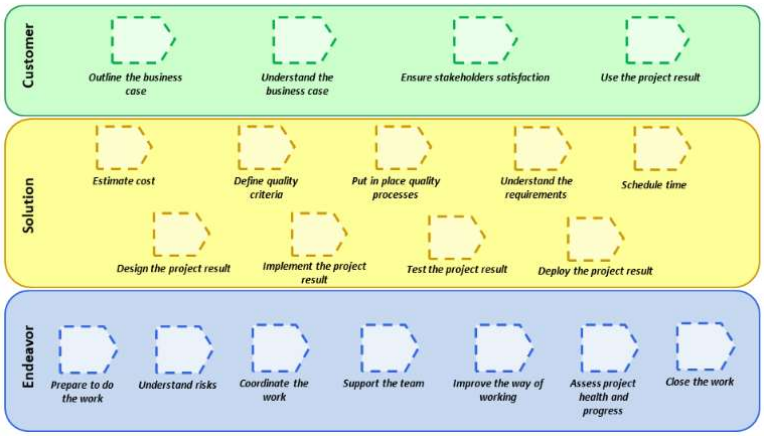


Fig. 13. Project management quintessence kernel activity spaces [16]

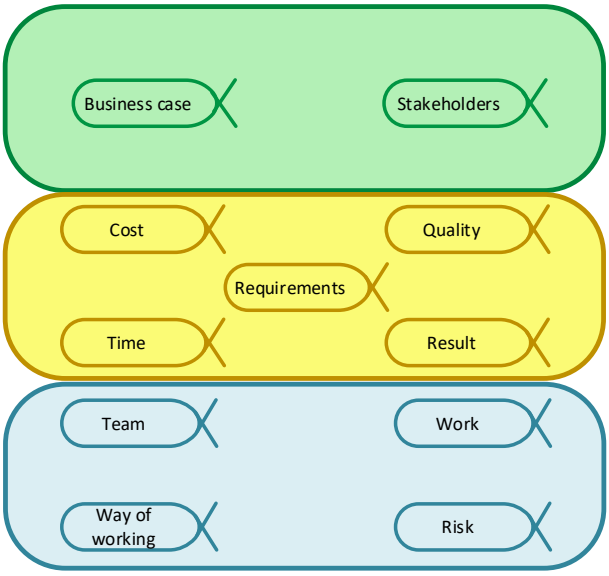


Fig. 14. Project management quintessence kernel Alpha [16]

Table 3. Best practices, activity spaces, and alphas associated to each phase CRISP-DM methodology

Phase	Best Practice	Activity space	Alpha
Business understanding	Project scope management	Understand the requirements	Requirements
Data understanding	Project procurement management	Prepare to do the work	Work
	Project quality management	Define quality criteria	Quality
Data preparation	Component-based development	Implement the project result	Result
Modeling	Component-based development	Design the project result	Result
		Implement the project result	
	Project quality management	Test the project result	
Evaluation	Project quality management	Test the project result	Result
Deployment	Project integration management	Coordinate the work	Work

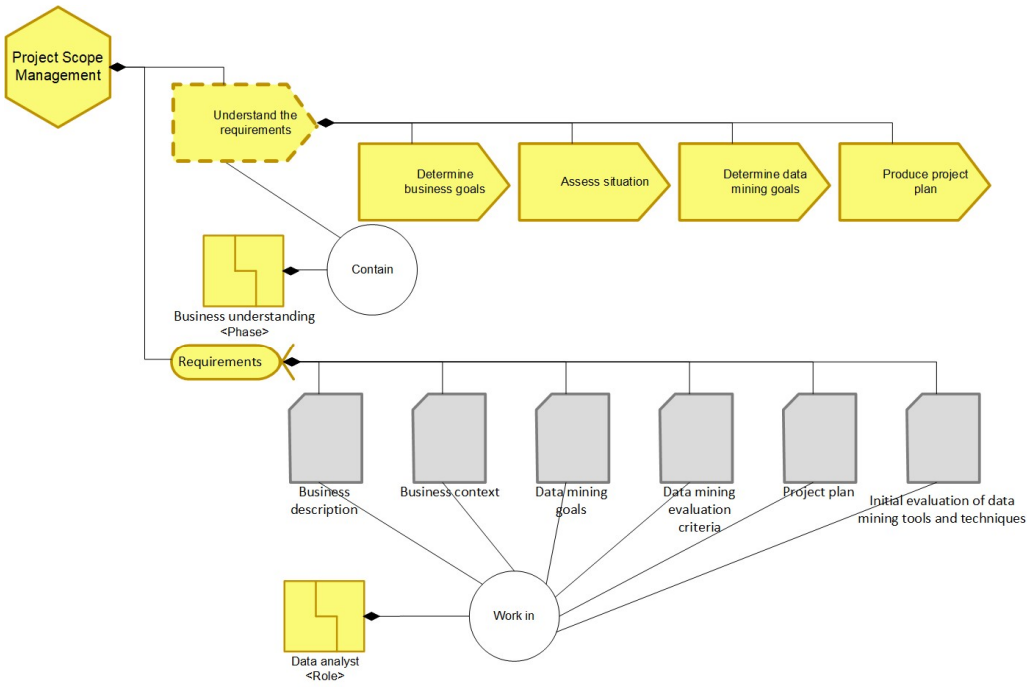


Fig. 15. Representation Essence: phase business understanding

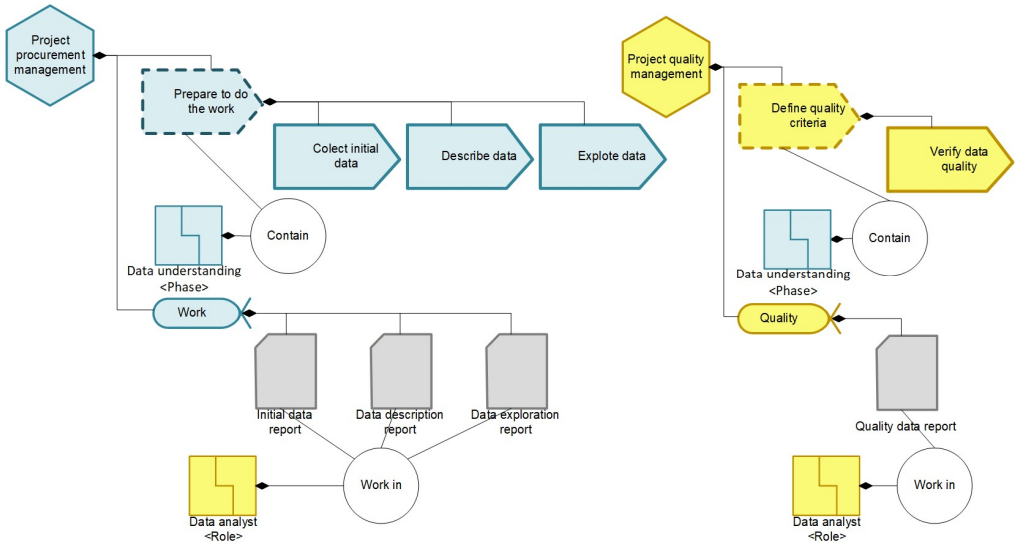


Fig. 16. Representation Essence: phase data understanding

Nevertheless, the depth lack is demonstrated in the identification of the essential elements of the methodology CRISP-DM like the categorization of the phases with activities and products of work to be executed.

For it, in this Section we propose a representation of the methodology CRISP-DM in Essence, identifying the “things we always work with” and the “things we always do”. Fig. 10 presents the proposal for representation of the

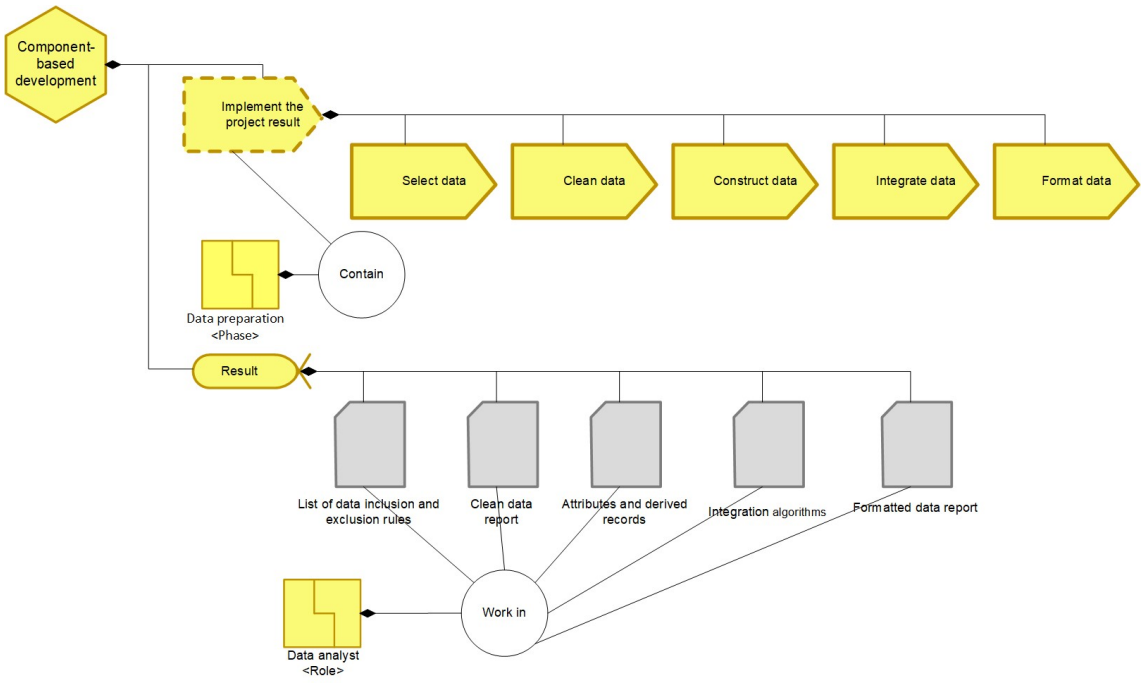


Fig. 17. Representation essence: phase data preparation

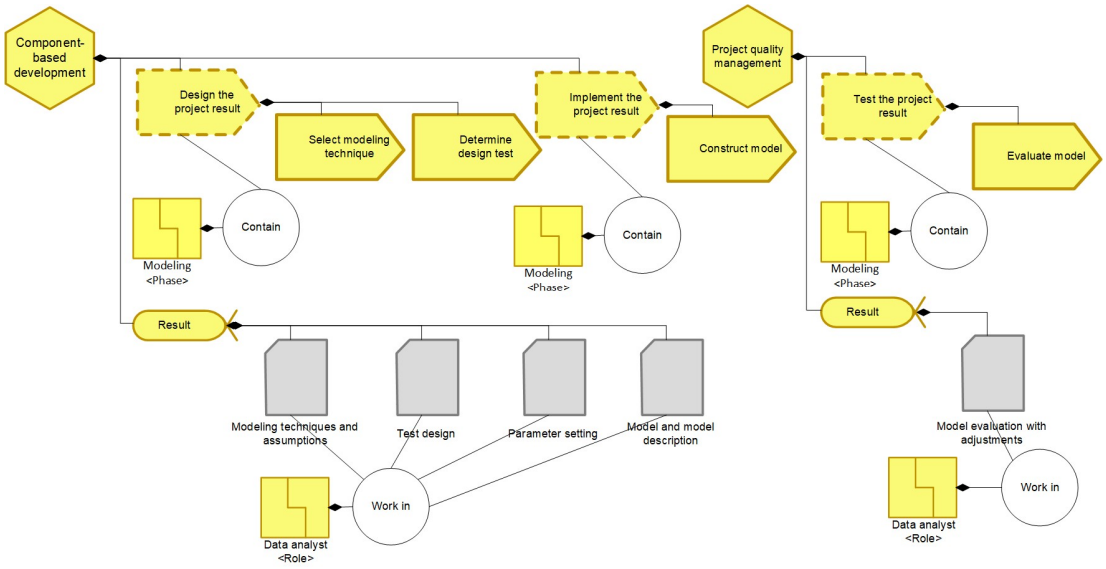


Fig. 18. Representation essence: phase modeling

methodology phases: business understanding, understanding data, data preparation, modeling, evaluation, and development. Fig. 10 presents the phases deployment and data understanding are area of interest "Endeavor", while the other phases are area of interest "Solution". To perform the representation of the Essence of the CRISP-DM methodology in each phase, a summary of the activities and work products associated with each phase is carried out (see Table 2).

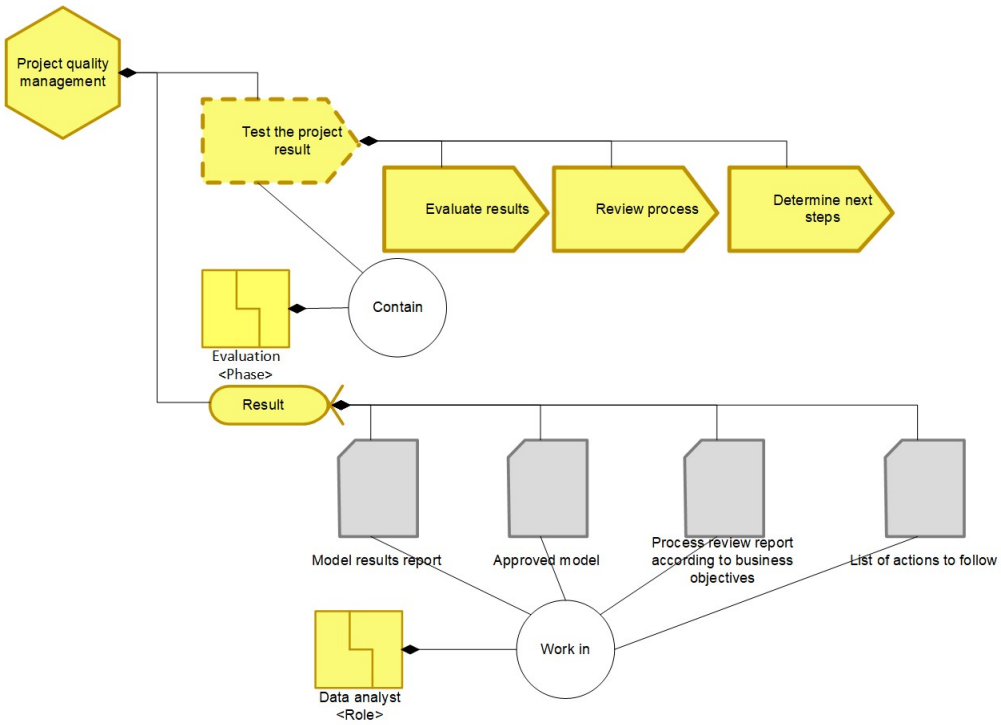


Fig. 19. Representation essence: phase evaluation

Next, we proceed to identify best practices of Essence of Project Management Body of Knowledge (PMBOK®) and Rational Unified process (RUP®).

– Durango (2019) presents ten areas of PMBOK knowledge® version 6 represented in Essence (see Fig. 11) [13].
– Gonzalez *et al*. (2013) represent best practices RUP® in Essence (see Fig. 12) [15].

Next, we proceed to identify other elements that activity spaces and alpha. Henao (2018) proposes in some builders who help define a multidisciplinary theory for project management. To achieve this, it proposes additional elements to those existing in Essence for activity spaces (see Fig. 12) and alpha (see Fig. 13) [16] for project management.

Next, we associated best practices of RUP®, PMBOK® and main elements of Fifth Essence, according to the elements of the CRISP-DM methodology (see Table 3).

Next, we present to representations in Essence of six phases of the methodology CRIPS-DM.

5 Validation of the Proposed Essence Representation

To validate the understanding of this representation, it is applied in two research projects. A project with an undergraduate student and with a research group from Tecnológico de Antioquia. As a method of validation of the proposal, we applied surveys with open questions to identify the improvements identified in the application of the CRISP-DM methodology in the projects using representation, with the following results:

– In the first phase of the validation, we identify that it is necessary to generate the work products required in each phase of the CRISP-DM methodology. Therefore, we define six work products according to the needs of CRISP-DM.
– Work team review progress and health of the data mining project
– Work team can add best practices, activities, and work products identified in other software

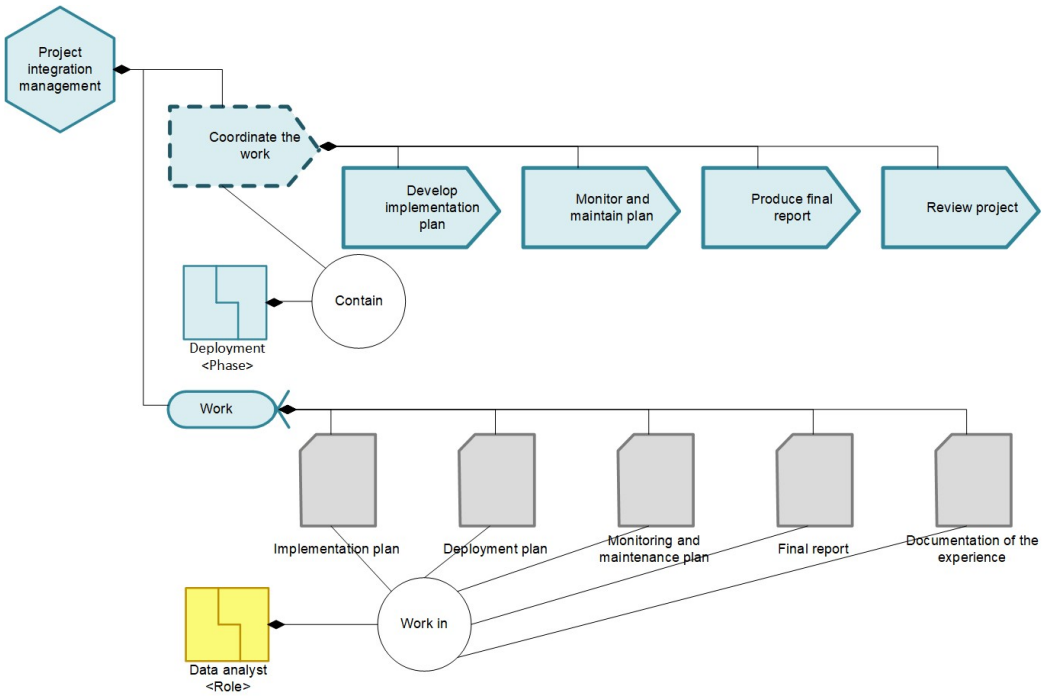


Fig. 20. Representation essence: phase deployment

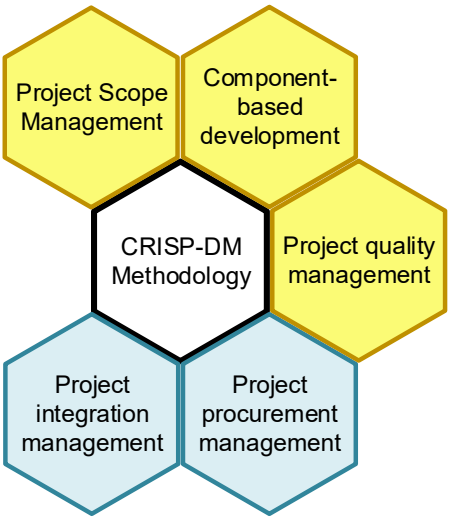


Fig. 21. Best practices of the CRISP-DM methodology

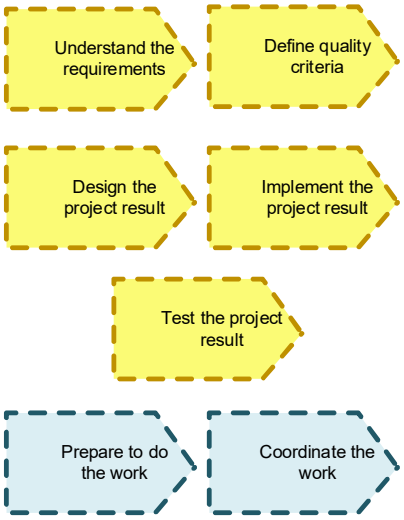


Fig. 22. Activity space of the CRISP-DM methodology

development methodologies, such as: user stories, use cases, and software architectures, among others.

– The proposed representation is a support tool for the work team, helping to identify the activities that must be carried out in the data mining project.
– Representation helps evaluate current practices and improve the way the development team works

6 Conclusion

According to Essence representations of elements that make up the six phases of the CRIPS-DM methodology, we obtained the following conclusions:

– Fig. 23 presents the best practices associated with the CRISP-DM methodology based on PMBOK® and RUP®, evidencing that it is necessary to define the best practices related to the area of interest “Customer”.
– Fig. 25 presents activity spaces associated with the CRISP-DM methodology, it is necessary to associate elements of the area of interest “Customer” and to improve customer satisfaction and participation in the project life cycle.
– Essence can be used in representations of different projects and methodology data mining.

7 Future Work

We propose that according to Essence representations, it is necessary to define the roles of a work team specify for CRISP-DM methodology, associating roles and responsibilities of data mining. In the representations proposed, we used only the role of “Data analyst” without functions and responsibilities in the project life cycle.

References

1. **Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000).** CRISP-DM 1.0: Step-by-step data mining guide. SPSS, Vol. 9, No. 13, pp. 1–73.
2. **Shearer, C. (2000).** The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehouse, Vol. 5, No. 4, pp. 13–22.
3. **Wirth, R., Hipp, J. (2000).** CRISP-DM: Towards a standard process model for data mining. Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, pp. 29–39.
4. **Jacobson, I., Ng, P. W., McMahon, P. E., Spence, I., Lidman, S. (2014).** La esencia de la ingeniería de software: El núcleo de Semat. Revista Latinoamericana de Ingeniería de Software, Vol. 1, No. 3, pp. 71–78. DOI:10.18294/relais.2013.71-78.
5. **Zapata-Jaramillo, C. M., Gil, N. (2011).** Incorporation of both pre-conceptual schemas and goal diagrams in CRISP-DM. 6th Colombian Computing Congress, pp. 1–6. DOI: 10.1109/COLOMCC.2011.5936284.
6. **Sharma, A., Mansotra, V. (2016).** Data mining based decision making: A conceptual model for public healthcare system. 3rd International Conference on Computing for Sustainable Global Development, pp. 1226–1230.
7. **Panov, P., Džeroski, S., Soldatova, L. (2008).** OntoDM: An ontology of data mining. International Conference on Data Mining *Workshops*, pp. 752–760. DOI: 10.1109/ICD MW.2008.62.
8. **Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Daimlerchrysler, C., Shearer, R., Wirth, R. (1999).** CRISP-DM 1.0: Step-by-step data mining guide. pp. 1–78.
9. **Anand, S., Grobelnik, M., Herrmann, F., Hornick, M., Lingenfelder, C., Rooney, N., Wettschereck, D. (2007).** Knowledge discovery standards. Artificial Intelligence Review, Vol. 27, No. 1, pp. 21–56. DOI: 10.1007/s10462-008-9067-4.
10. **Object Management Group (2018).** Kernel and language for software engineering methods. https://www.omg.org/spec/Essence /1.2.
11. **Durango-Vanegas, C. (2019).** Definición de buenas prácticas de desarrollo de sistemas de información geográfica utilizando el núcleo de Semat (Tesis de Doctorado). Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión, Universidad Nacional de Colombia, Medellín, Colombia.
12. **Bošnjak, Z., Grljević, O., Bošnjak, S. (2009).** CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises’ data. 5th International Symposium on Applied Computational Intelligence and

Informatics, 114, pp. 509–514. DOI: 10.1109/SACI.2009.5136302.

13. **Ortiz-Pabon, H. J., Zapata-Jaramillo, C. M., González-Calderón, G. (2014).** La gestión de programas académicos desde la perspectiva de la gestión del conocimiento apoyada con esquemas preconceptuales. Revista Ingenierias Universidad de Medellín, Vol. 13, No. 25, pp. 191–205. DOI:10.22395/rium.v13n25a12.
14. **Henao-Roqueme, A. J. (2018).** Towards a theory for defining a project management multidisciplinary kernel: An approach based on abstract level progress health attributes (Magister thesis). Universidad Nacional de Colombia, Medellín.
15. **González-Pérez, M., Zapata-Jaramillo, C. M., González-Palacio, L. (2013)** Toward a standardized representation of RUP best practices of project management in the SEMAT kernel. Software engineering: methods, modeling, and teaching, Vol. 3, Chapter 7, pp. 47–52.

Article received on 26/06/2020; accepted on 16/09/2021.
Corresponding author is Claudia Elena Durango Vanegas.