Development of a Domain Specific Modeling Language for Educational Data Mining

Eronita M. L. Van Leijden and
Alexandre M. A. Maciel
University of Pernambuco - UPE
Recife, Brazil
emlvl, amam@ecomp.poli.br

Andrêza Leite de Alencar
Federal Rural University of Pernambuco
Recife, Brazil
andreza.leite@ufrpe.br

Abstract

In data mining solutions, the data selection phase plays an essential role in the success of decision-making. The tools that operate at this phase need to cater to each domain's technical and management challenges. Using a Domain-Specific Modeling Language (DSML), we found an alternative to abstract data and simplify the selection for Educational Data Mining (EDM) process. This work presents a graphic DSML to represent the problem. We used a case study methodology and implemented a CASE tool for the language evaluation. We acquired evidence that the proposed language simplifies the data selection phase for EDM because it solves the technical and management challenges addressed to this domain.

1 Introduction

Nowadays, there has been a growth in the Educational Data Mining (EDM) field of research. This area leads with the development of methods that help examine to collect data from educational platforms. This area's main objectives are to understand how students interact in their learning environments and what they learn. So it turns possible to propose decision-making actions for better educational results [1].

Among the phases of the EDM process, there is the data selection phase. In this phase, the information is identified among existing data sets and considered during the modeling process. This phase includes choosing which data to collect and ensuring that the data is coherent with the phenomenon to be analyzed [2].

The tools that operate in this phase generally need to meet the challenges of the technical aspects of data processing. These challenges are associated with a large amount of data processing from different sources and the significant data heterogeneity with structured, semi-structured, and unstructured data [3]. In addition, it is necessary to confront two management challenges: First, these solutions should enable the reuse of the models to understand already known educational phenomena, such as performance prediction; detection of behavioral patterns; evasion indicators; etc.; secondly, it is necessary to provide conditions for academic analysts, who are not experts in data processing and analysis, conduct an analytical process [2].

Model-Driven Development (MDD) is a development paradigm that uses models as the primary artifact of the development process. In MDD the implementation is (semi) automatically generated from the models [4]. In constructing an MDD tool for a specific domain, it is needed to define its modeling language initially. Domain-Specific Modeling Language (DSML) makes it possible to create rules with high-level graphic and/or textual definitions. When applied in an MDD tool, it acts as a spelling and grammar checker, with validation to avoid syntax errors or typos [5].

This work aims to develop of a Domain-Specific Modeling Language for Educational Data Mining, in which the solution considers the technical and managerial challenges of this domain. For this, it was modeled a language and developed a prototype of an experimental case tool. For validate these artefacts a case study was realized using different versions of Moodle databases to validate this work.

2 Background

This section presents essential concepts necessary for proposed solution understanding.

2.1 Domain Specific Modeling Languages (DSML)

As one of the elements used on the MDD, Domain-Specific Modeling Languages enables the creation of rules with a high-level graphic and/or textual definition to be converted into a low-level language [4]. The definition of a DSML involves at least three aspects: the domain concepts and rules (abstract syntax); the notation used to represent these concepts—let it be textual or graphical (concrete syntax); and the semantics of the language [4].

The abstract syntax of a DSML is particularized by a metamodel, which is itself a model and describes the concepts of the language, the relationships among them, and the structuring rules that constrain the model elements and their combinations in order to respect the domain rules. The concrete syntax provides a realization of its abstract syntax as a mapping between the metamodel concepts and their textual or graphical representation. The semantics of a DSML is normally given with natural language. However, although

DOI reference number: 10.18293/SEKE2022-137

users can normally deduce the meaning of most terms of a DSML, a computer cannot act on such assumptions [6].

2.2 Design Theories for Visual Notation

Design theories for visual notation provide the scientific basis for evaluating and designing visual notations. According to Moody [7], two approaches stand out: descriptive theory and prescriptive theory. The descriptive theory is used only to understand how and why the visual notations communicate (visual grammar). The prescriptive theory consists of a definition of explicit principles that deal with the design of visual notation, which handles the transformation of an unconscious process into a self-conscious process (visual vocabulary) [7].

The anatomy of a good visual notation consists of adding the definition of graphic symbols (visual vocabulary) to the rules of composition (visual grammar). To this end, Moody [7] suggests that some principles for designing and evaluating graphic symbols. The main ones are Semantic Transparency, Visual Expressiveness, Semiotic Clarity, Perceptual Discrimination and Graphic Economy [8].

2.3 Related Works

Two works were selected for this article because they deal with approaches related to structuring the input data of the EDM process.

The work of Magalhães Júnior [9] is a proposition of a data model that brings together indicators applicable in various educational phenomena. The solution lists the attributes used in the different EDM works. Under the focus of the phenomenon "student dropout," a catalog was developed based on an Entity and Relationship Diagram(DER) that served as a data integration point. After this step, the author performed new queries in this intermediate base to generate the file, which is input in the EDM process. Its objective was to reduce the efforts to select attributes and subsequent preparation of the data for the EDM.

Manhães [10] developed an architecture based on three layers according to EDM concepts: data layer, application layer, and presentation layer. Although cited, work does not explain how data collection was performed. The proposed solution describes an architecture layer destined for the "selected data", called Knowledge Repository. It stores the different student data models used as input on the EDM process, i.e., a cataloging of data sets able to be mined.

3 Proposed Language

The proposed modeling language represents the flow design to carry out the first phase of an EDM process - the data selection. It means the match between a field in a source database and a field of a target database passed through a visual notation. Next sections describe this language.

3.1 Language Rules

The DSML requirements necessary for the development of this research are presented below:

RQ1: To enable the use of data from different data sources of the educational platforms.

RQ2: To allow different composing and storing data: relational database, spreadsheets, data warehouse, log file, data stream, and web data, among others.

RQ3: To allow the cataloging of data structures by educational phenomenon (knowledge record).

RQ4: To allow the standardization of the "selected data" for the EDM process entry independent of the educational platform. It should also enable the data to continue as a file, a relational, or a multidimensional model.

3.2 Abstract syntax

The abstract syntax proposed in this work was inspired by Alencar [11]. It has similarities with the requirements RQ1, RQ2, and RQ4 listed in Section 3.1 and because of this, it was aggregated to the metamodel. The proposed metamodel is detailed in dissertation by Leijden [12] where the adjustments made are presented and the origin of new objects is explained.

3.3 Concrete syntax

The visual notation to represent the educational data selection process was developed following the main prescriptive principles proposed by Moody [7]. Here are the details of each of these principles and how they were considered in this work:

- **Semantic Transparency** – defines the visual representations used in a way that their looks suggest meaning. Area 2 of Figure 1 shows the list of defined graphic symbols and the labels for each proposed symbol.

  1. "Base Tool": represents the input data sources. It is a classic symbol for databases representation.

  2. "Base Version Tool": represents the input data sources. The "V" mark suggests it is a version of the database.

  3. "Entity Tool": is a regular blue pentagon with the letter "E" to represent the variants of the input information set. The letter "E" in the image represents an allusion to the entity in the ER model.

  4. "Mining Phenomenon": the mining cart with precious stones refers to an EDM process. The image depicts the data sources selected to perform the data mining process, polished to generate information.

  5. "Educational Phenomenon": the blue owl on a book represents the variants of the data sets selected to be mined. As the owl is a classic symbol representing

education (area of this research), the image represents the educational phenomenon as an entity.

6. ”Attribute Tool”: a red diamond with the letter “A” represents the characteristics of each entity. The letter A refers to the term attribute usually used in the ER model.

7. ”Sub Attribute Tool”: it is similar to Attribute Tool but uses different color (yellow). In addition, the ”greater than” sign was inserted to represent dependence with a specific Attribute Tool.

8. ”Association”: the blue double arrow represents relations between entities. The symbol was inspired by Bachman’s notation [13] which became known as arrow notation.

9. ”Flow”: arrows dashed in black. It represents the equivalence between the attributes of a source and those of a target entity. It is the component that shows the data flow.

- **Visual Expressiveness**-defines how should uses variables and visual capabilities and how to group strongly related elements. Area 2 of Figure 1 shows the elements distributed in 4 groups: ”Source Area”, ”Target Area”, ”Data Composition” and ”Event Composition”.

  ”Source Area”: group the elements that represent the ”entry area” concept.

  ”Target Area”: group the elements that represent the ”data to be mined area” concept.

  ”Data Composition”: group the elements related to characteristics and the relation of each entity of databases (entry and to be mined).

  ”Event Composition”: : the event symbols. The dashed arrow represents the flow that the mapped data goes between origin and educational phenomenon.

- **Semiotic clarity** - defines that there must be a 1:1 (one-to-one) correspondence between the semantic constructors (metamodel) and the graphic symbols of the language. Five metaclasses of the metamodel go mapping between the semantic constructors and the visual syntax: ”Base”, ”BaseVersion”, ”Attribute”, ”Association” and ”Mining Phenomenon”.

- **Perceptual Discriminability** - defines that different symbols must be clearly distinguished from each other. This principle is applied to the graphic symbols and the model diagramming when the graphical elements are inserted in the drawing area. It uses the container-based visual technique to demonstrate the hierarchy among the elements.

- **Graphic Economy** - defines that the number of different graphic symbols must be cognitively manageable. In this work, the ”Value” element does not have a symbol associated with its visual representation to limit the graphic and diagrammatic complexity of the model.

3.4 CASE Tool

Developed using Sirius , the CASE tool is organized into three regions (Figure 3). Region 1 consists of a drag-and-drop area for the compositions of a data selection case, Region 2 shows the components (symbols of the metamodel elements) placed in a palette tab, and Region 3 offers a properties tab for the selected objects in the drawing.

Performing data selection in a CASE tool based on metamodel and modeling language makes the tool automatically verify possible flaws in creating its routines regarding the use of symbols (visual vocabulary) and the composition rules (grammar). The CASE modeling tool can validate the diagrams, verifying that they follow the established syntax and semantics. This feature prevents users from misusing the model’s graphic symbols.

4 Analysis and Discussion

The analysis was conducted by following the methodological procedures for the study case presented by Yin [14]. It aimed to evaluate the modeling language, through a prototyped CASE tool, regarding the adequacy of its use in the circumstances of EDM projects, particularly in the first phase, which is the data selection. Table 1 shows the synthesis of the way the case study was conducted.

Table 1: Synthesis of the case study

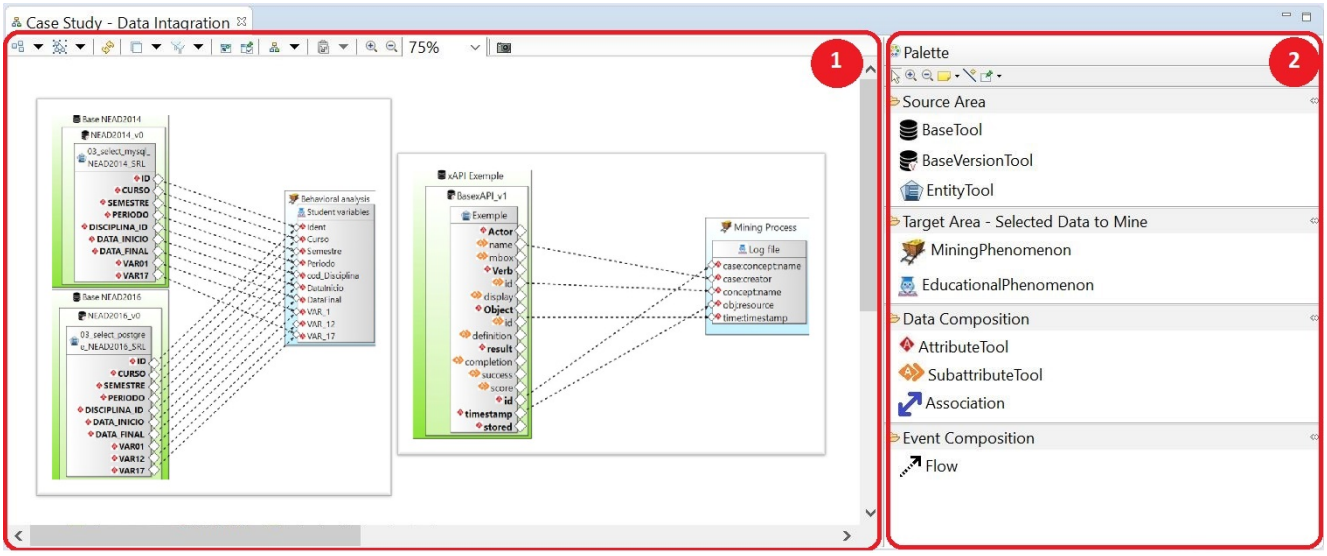
Description	Qualitative research of a descriptive character
Type Design	Single-Embedded
Study object	Language assessment using a prototype CASE tool
Study Unit	Problem situations in the data selection process
Data collect	Systematic observation
Data analysis	- Application of fragments of contentanalysis in two studies: consumption ofthe xAPI standard and unification of astructured database. - Fault simulations.

4.1 Context and Measured Variables

The observed variables are connected to the functional quality of the software. Following the model recommended by ISO/IEC 9126 (NBR13596) [15], evidence was observed regarding the variables of adequacy, accuracy, and interoperability.

The evidence about the adequacy and interoperability measures was obtained through observation when simulating the data selection process adopting the developed proto-

Figure 1: CASE Tool print screen



type and comparing the aspects observed with the requirements set out in Section 3.1. Figure 1, area 1, illustrates the diagrams created in this phase.

As for accuracy assessment, evidence was captured using the tool while executing modeling simulations that would violate the rules and restrictions created in the grammar developed in the modeling language.

4.2 Discussion

The planning and execution of the study case were carried out based on the objective of the work following research questions below:

Q1: Does the functional behavior conforms to proposed by the rule, the meta-modeling, and the language notation when using the CASE tool?

Q2: The requirements listed in section 3.1 met?

Q3: Can educational analysts, even not being data processing experts, increase the autonomy to carry out a data selection in the EDM process?

The results of these analyzes are depicted below.

Accuracy: During the modeling simulations, it was observed that the rules and restrictions placed on the visual modeling syntax (metamodel + language) were all taken into account. For instance, the tool, when correctly used, did not allow to create a flow from the input base to another input base; nor did it allow to create a data flow from the ”Mining Phenomenon” to the input base. Especially, the tool automatically made checks for flaws in the construction of the modeling that prevented the use of wrong model components in astray compositions in the process.

Adequacy: By applying the developed prototype, the result of the diagramming was evaluated, shown in Figure 3, with the requirements listed in Section 3.1. This analysis explicitly answers the research question Q2.

Requirement 1 and 2: It could be seen that the tool had achieved its goal since it was able to represent the modeling of both data from the xAPI standard, which is in JSON format, and of the Moodle database, which is a SQL structured query.

Requirement 3: For the two situations presented in the case study, the researcher chose to use a known data set structure, thus seeking to use the concept of knowledge reuse. Albeit it was possible to represent the modeling of the known data structure for analysis in EDM, this proposal does not guarantee such situation since the defined grammar only makes feasible a future development of an executable code that accesses some knowledge repository.

Requirement 4: Part of requirement 4, which deals with the issue of data representation in a semi-structured format, has not been directly validated.

Nonetheless, we can infer that this condition is valid by considering that the structure of the meta-modeling presented for the data referring to the source (which has been validated) is the same that will represent the selected data set, target base.

Interoperability: The tool can perform technical interoperability, as the solution covers two fundamental problems in information integration: data exchange and entity resolution [16]. In the tool, data exchange is promoted when the solution’s ability to represent different arrangements is demonstrated. As for entity resolution, the tool can identify and associate the information between data sources in a single destination, as depicted by the ”unification of structured databases” situation.

Given these analyzes, the specific research questions posed in Section 4.2 were considered and answered.

Regarding Q1, it was found to be true. Evidence was ac-

quired in the analysis of "accuracy" and, comprehensively, also obtained in the analysis of "adequacy". For the tool to perform the syntax's automatic validations, the metamodel must be defined according to the needs pointed out as requirements of section 3.1.

As for Q2, the answer is explicitly found in the "adequacy" analysis. The answer to Q3 is obtained while analyzing the creating process of each diagram. In neither of the two diagrams created was required the use of programming languages. Everything was done using clicks, moving graphic elements, and filling properties. This characteristic is inherent to the MDD technique. It demonstrates that non-expert users in data processing and analysis can conduct an analytical process (at least when it comes to the first phase of the EDM process).

5 Conclusions and Future Work

Throughout this work, it could be perceived, by empirical analysis, that the language created allows the diagramming of the phase of data selection to be used in EDM process, without the need for technological knowledge. In addition, the functional quality of the software was validated, as displayed in the observation on functional quality; adequacy, accuracy, and interoperability. Given what was brought and discussed, the work presented the following contributions:

Expressive metamodel - verification made when answering the Q1 of section 4.2, the functional behavior in the prototype was adherent to what was identified in the rules of language and what was proposed in meta-modeling and the language notation.

Cognitively effective notation - the work attends to the principles proposed by Moody [7].

Functionally adherent to the needs of the domain - demonstrated in detail in section 4.2, which deals with the analysis and discussion of the execution of the case study.

Simplification of the phase - verification made when answering Q3 in section 4.2.

As for future works, it is intended: i) to implement elements of transformation of the MDD. The T2M and M2T transformations in the following transformation functionalities: in the automatic data diagramming, where the data structure will be obtained from the source and automatically transformed into the model, and in the automatic generation of the source codes, so that the built model can be executed automatically at regular intervals by some task management tool like crontab, for example; ii) to expand the proposal to also carry out the pre-processing phase of EDM, promoting the solution to the ETL environment; iii) to complement the research validation, which may be an experiment, a participant observation and/or a questionnaire based on expert opinion; and iv) to develop studies of this proposal in the context of Big Data and Data Lake.

6 Acknowledgment

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] S. A. Salloum, M. Alshurideh, A. Elnagar, and K. Shaalan, "Mining in educational data: Review and future directions," in *Joint European-US Workshop on Applications of Invariance in Computer Vision*. Springer, 2020, pp. 92–102.

[2] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.

[3] E. J. Dommett, "Understanding student use of twitter and online forums in higher education," *Education and Information Technologies*, vol. 24, no. 1, pp. 325–343, 2019.

[4] M. Brambilla, J. Cabot, and M. Wimmer, "Model-driven software engineering in practice," *Synthesis lectures on software engineering*, vol. 3, no. 1, pp. 1–207, 2017.

[5] J. White, J. H. Hill, J. Gray, S. Tambe, A. S. Gokhale, and D. C. Schmidt, "Improving domain-specific language reuse with software product line techniques," *IEEE software*, vol. 26, no. 4, pp. 47–53, 2009.

[6] H. Krahn, B. Rumpe, and S. Völkel, "Integrated definition of abstract and concrete syntax for textual languages," in *International Conference on Model Driven Engineering Languages and Systems*, vol. 4735, Springer. Berlin: Springer, 2007, pp. 286–300.

[7] D. L. Moody, "The" physics" of notations: a scientific approach to designing visual notations in software engineering," in *2010 ACM/IEEE 32nd International Conference on Software Engineering*, vol. 2, IEEE. Cape Town: IEEE, 2010, pp. 485–486.

[8] H. B. M. Diniz, "Linguagem específica de domínio para abstração de solução de processamento de eventos complexos," Master's thesis, Universidade Federal de Pernambuco, 2016.

[9] P. N. Magalhães Júnior, "Um modelo de dados para apoiar a mineração de dados educacionais na investigação de evasão de estudantes," *tede.unifacs.br*, 2013.

[10] L. M. B. Manhães, "Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais," *Doutorado em Engenharia de Sistemas e Computação Instituição de Ensino: Universidade Federal do Rio de Janeiro, Rio de Janeiro. Biblioteca Depositária: BIBLIOTECA DO CT*, vol. 1, no. 1, pp. 1–157, 2015.

[11] A. L. d. Alencar, "Um meta-modelo para representação de dados biológicos moleculares e suporte ao processo de anotação de variantes genéticas," *repositorio.ufpe.br*, vol. 1, no. 1, pp. 1–152, 2018. [Online]. Available: https://repositorio.ufpe.br/handle/123456789/32659

[12] E. M. L. V. Leijden, "Desenvolvimento de uma linguagem específica de domínio para consumo de dados educacionais," 2020.

[13] C. W. Bachman, "The structuring capabilities of the molecular data model." in *ER*, 1983, pp. 55–68.

[14] R. K. Yin, *Estudo de Caso-: Planejamento e métodos*. Bookman editora, 2015.

[15] I. ISO, "Iso standard 9126: Software engineering product quality, parts 1, 2 and 3," 2001.

[16] K. Qian, "Discovering information integration specifications from data examples," Ph.D. dissertation, UC Santa Cruz, 2017.