Development of a Domain Specific Modeling Language for Educational Data Mining

AUTHORS:
Eronita Mari Luizines Van Leijden,
Andrêza Leite de Alencar and
Alexandre Magno Andrade Maciel

SEKE 2022

KSIR Virtual Conference Center, Pittsburgh, USA

July 1 - July 10, 2022

SCHEDULE

- *Introduction*
- *Background*
- *Proposed Language*
- *Analysis and Discussion*
- *Conclusions and Future Work*

Introduction

Contextualization and Motivation

Growth in the Educational Data Mining (EDM) field of research

→

Among the phases of the EDM process, there is the data selection phase.

→

This phase includes

- Choosing which data to collect
- Ensuring that the data is coherent with the phenomenon to be analyzed

Challenges in the tools that operate in this phase

Technical challenges :

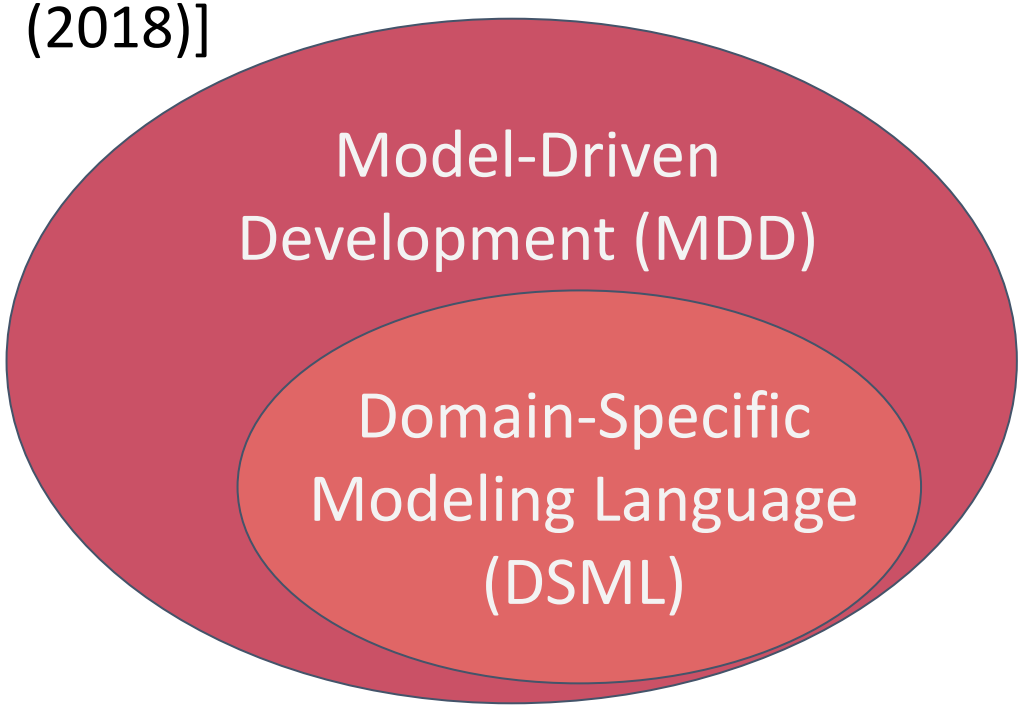
- Sources
- formats
- Structure

Management challenges

- to enable the reuse of the models to understand already known educational phenomena
- Use by non-experts

Contextualization and Motivation

MDD is presented as a **promising alternative**
[Brambilla *et. al* (2017)] [Henrie *at. al* (2018)]



➔ 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.

Background

Topics Studied

- *Domain Specific Modeling Languages (DSML)*
- *Design Theories for Visual Notation*
- *Related Works*

Domain Specific Modeling Languages (DSML)

Enables the **creation of rules with a high-level graphic and/or textual definition** to be **converted into a low-level language,** so that there are no syntax errors or typos.

Abstract syntax - the domain concepts and rules (metamodelo)

Concrete syntax - the notation used to represent these concepts—let it be textual or graphical

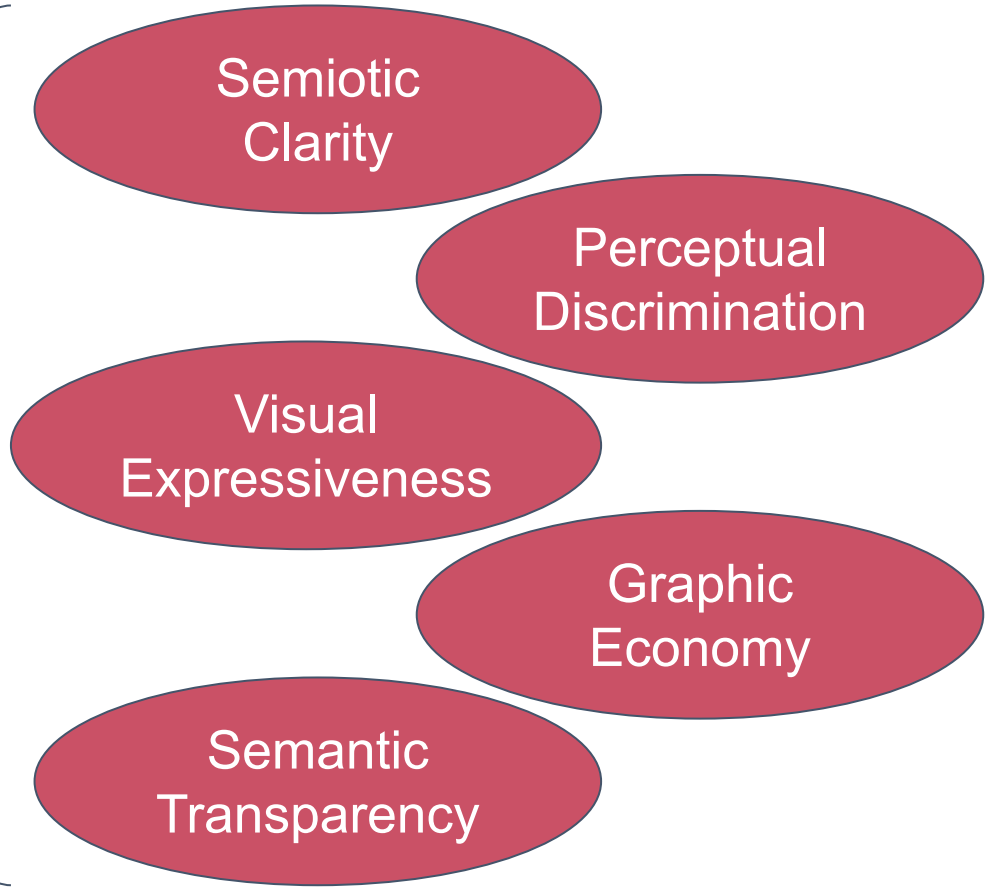
Semantics - explain the meanings

Design Theories for Visual Notation

"The anatomy of a good visual notation consists of adding the definition of graphic symbols (visual vocabulary) to the rules of composition (visual grammar)"

Moody (2010)

*Main **prescriptive principles** presented by Moody (2010)*



Related Works

- Unidentified no work that uses DSML (Domain Specific Modeling Languages) in EDM (Educational Data Mining)
- So, the research looked for works that approached the structuring of the data input in the Education Data Mining process.

Fonte: Fayyad et al. (1996)

Structuring of the data input in the EDM

Magalhaes Junior (2013)

Focus of the phenomenon "student dropout,"

Product: catalog based on an Entity and Relationship Diagram (DER)

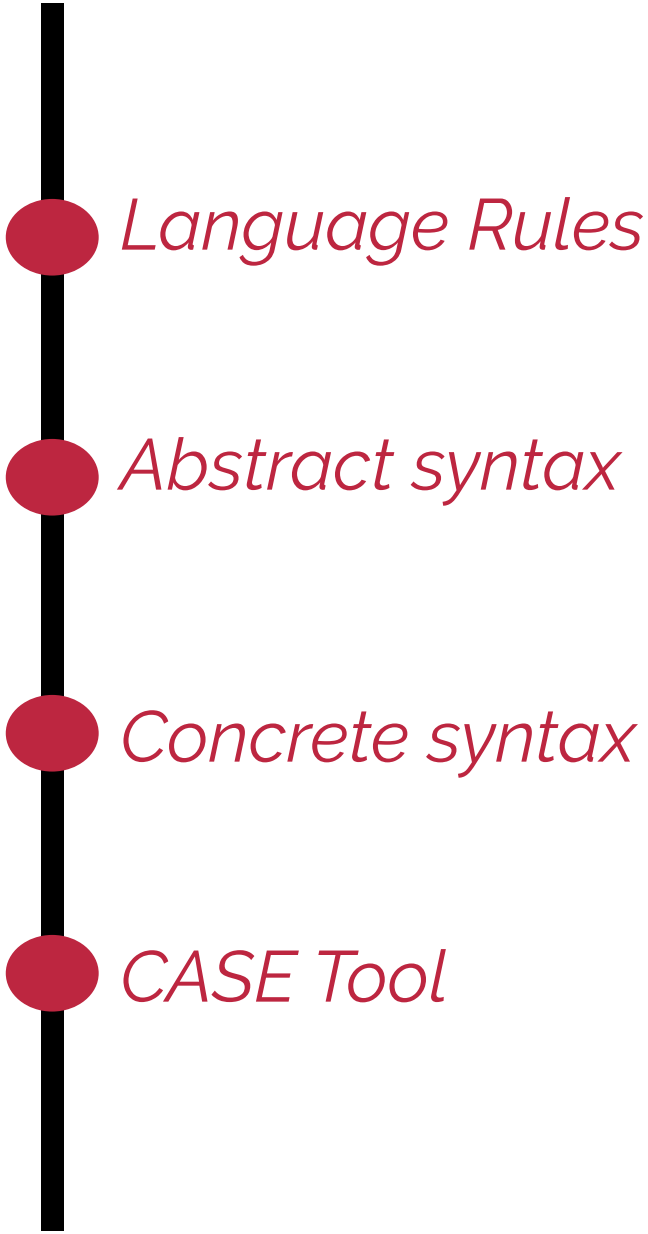
Manhães (2015)

Developed an architecture based on three layers according to EDM concepts

Product: describes an architecture layer destined for the "selected data", called Knowledge Repository

Proposed Language

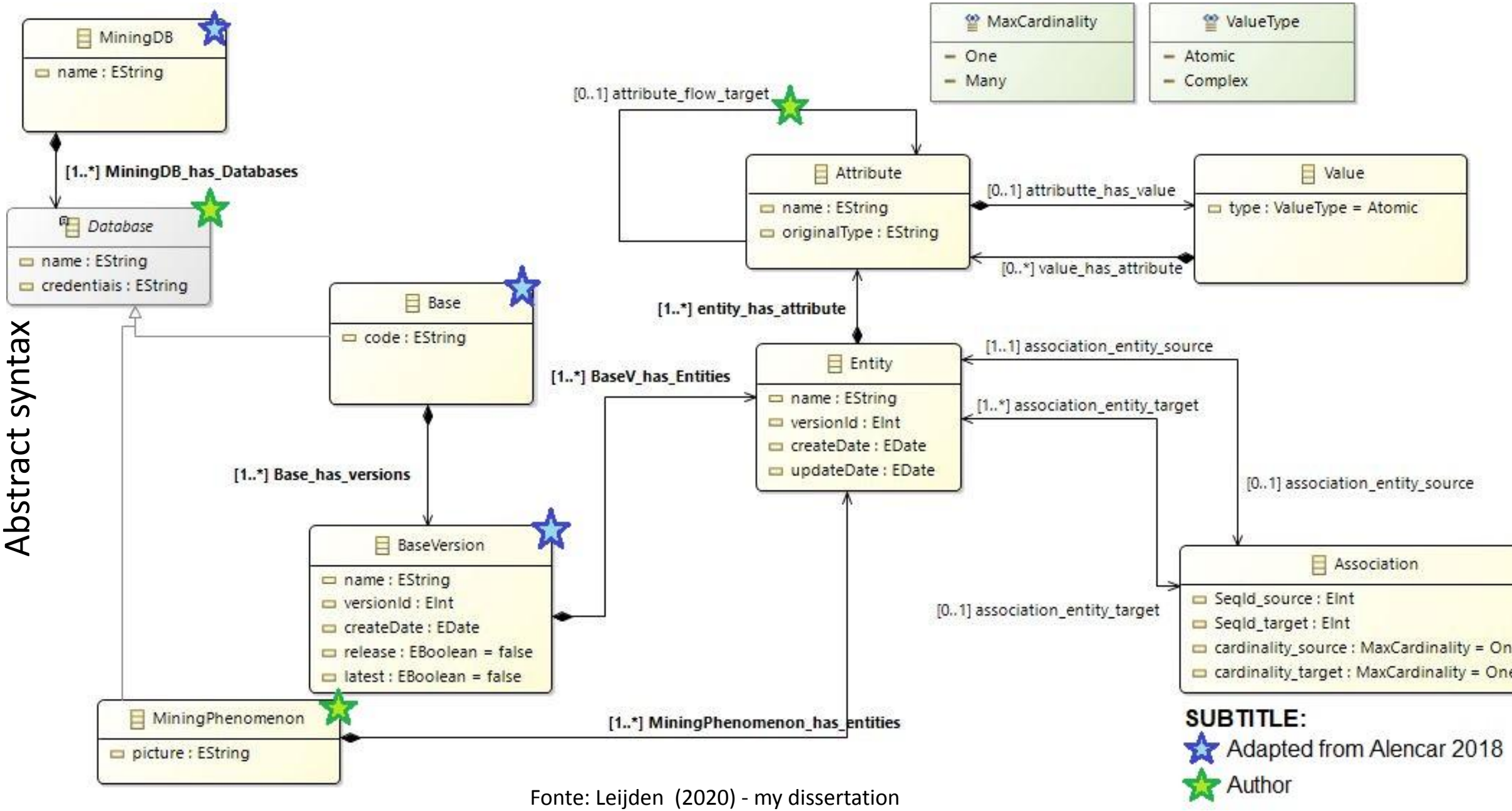
The proposed modeling language represents the flow design to carry out the first phase of an EDM process - the data selection. It means the match between a field in a source database and a field of a target database passed through a visual notation.



Language Rules

- 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.

METAMODEL
Abstract syntax

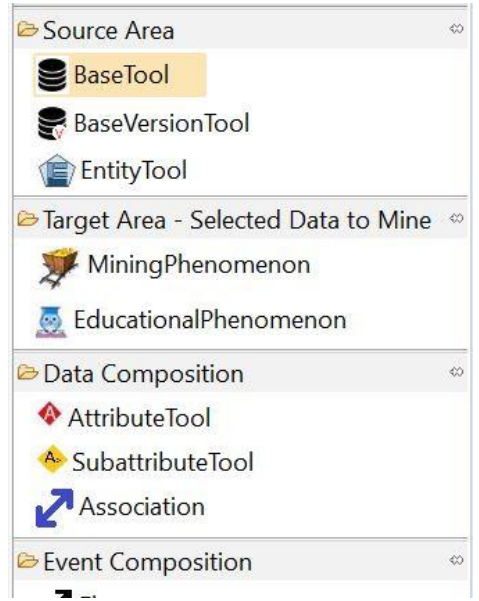


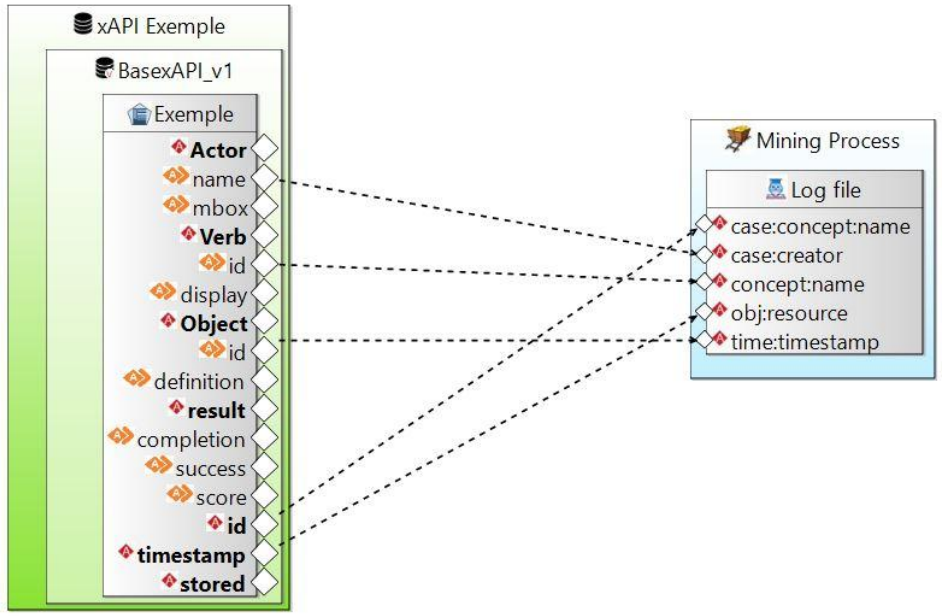
Fonte: Leijden (2020) - my dissertation

Concrete syntax

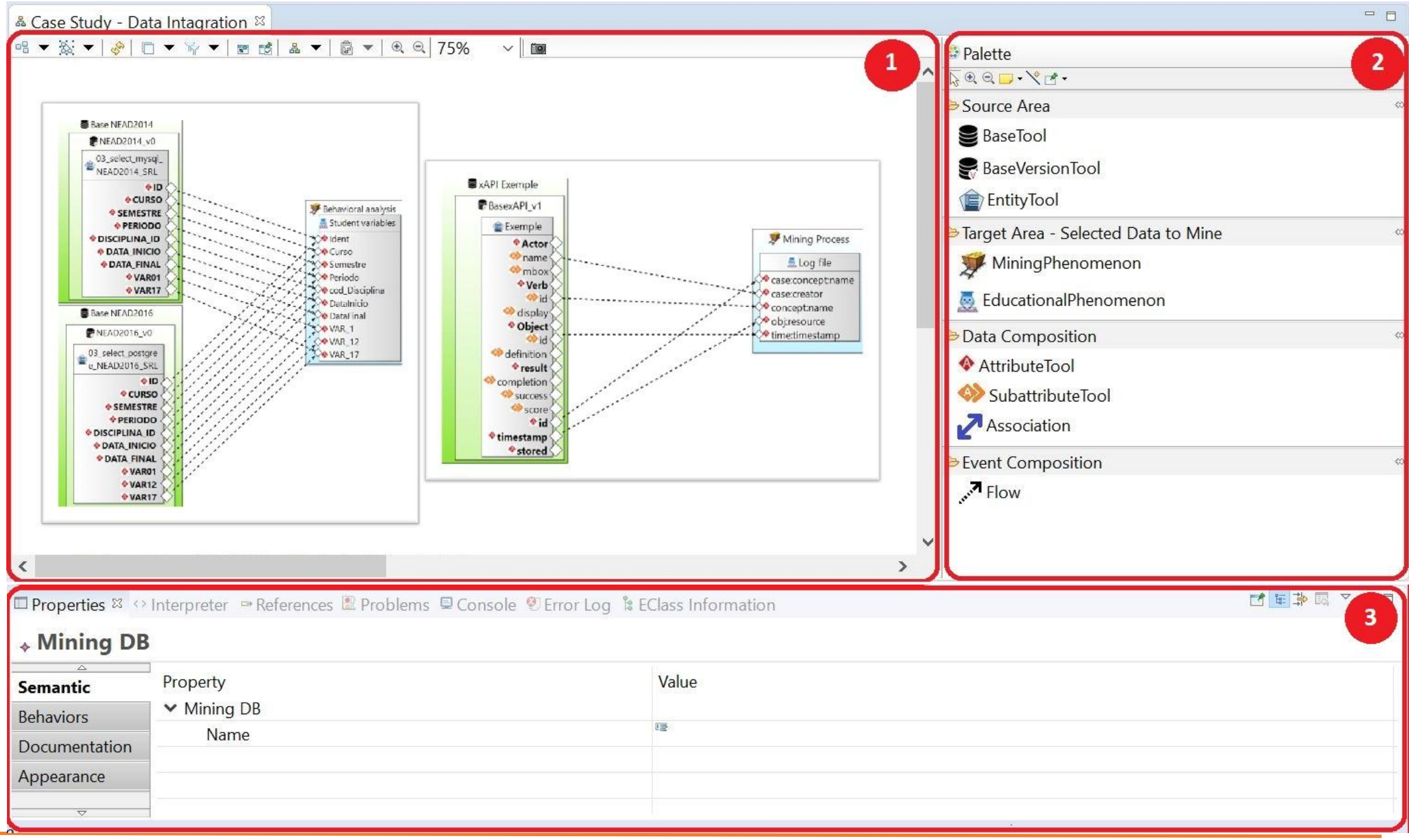
*Main **prescriptive principles** presented by Moody (2010)*

- Semiotic Clarity
- Perceptual Discrimination
- Visual Expressiveness
- Graphic Economy
- Semantic Transparency





CASE Tool


Case Study - Data Intagration
75%
1
Base NEAD2014
NEAD2014_v0
03_select_mysql_NEAD2014_SRL
ID
CURSO
SEMESTRE
PERIODO
DISCIPLINA_ID
DATA_INICIO
DATA_FINAL
VAR01
VAR17
Base NEAD2016
NEAD2016_v0
03_select_postgre_NEAD2016_SRL
ID
CURSO
SEMESTRE
PERIODO
DISCIPLINA_ID
DATA_INICIO
DATA_FINAL
VAR01
VAR12
VAR17
Behavioral analysis
Student variables
Ident
Curso
Semestre
Periodo
cod_Disciplina
DataInicio
DataFinal
VAR_1
VAR_12
VAR_17
xAPI Exemple
BasexAPI_v1
Exemple
Actor
name
mbox
Verb
id
display
Object
id
definition
result
completion
success
score
id
timestamp
stored
Mining Process
Log file
case:concept:name
case:creator
concept:name
obj:resource
time:timestamp
Palette
2
Source Area
BaseTool
BaseVersionTool
EntityTool
Target Area - Selected Data to Mine
MiningPhenomenon
EducationalPhenomenon
Data Composition
AttributeTool
SubattributeTool
Association
Event Composition
Flow
Properties
Interpreter
References
Problems
Console
Error Log
EClass Information
3
Mining DB
Semantic
Behaviors
Documentation
Appearance
Property
Value
Mining DB
Name

Analysis and Discussion

Synthesis of the case study

- The analysis was conducted by following the methodological procedures for the study case presented by Yin (2015).

- 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.

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.

Context and Measured Variables

The observed variables are connected to the **functional quality** of the software.

Model recommended by ISO/IEC 9126 (NBR13596)

- Adequacy
- Accuracy
- Interoperability

Discussion - search questions

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 slide “Language Rules” met?

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

Discussion - search questions

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

Results found in the "accuracy" analysis and, in a comprehensive way, also found in the "adequacy" analysis.

Q2: The requirements listed in slide “Language Rules” met?

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

Discussion - search questions

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

Results found in the "accuracy" analysis and, in a comprehensive way, also found in the "adequacy" analysis.

Q2: The requirements listed in slide “Language Rules” met?

Positive result found in the “adequacy” analysis.

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

Discussion - search questions

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

Results found in the "accuracy" analysis and, in a comprehensive way, also found in the "adequacy" analysis.

Q2: The requirements listed in slide “Language Rules” met?

Positive result found in the “adequacy” analysis.

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

Result considered positive because the diagrams generated in both cases did not need to use the programming language.

Conclusions and Future Work

Conclusions

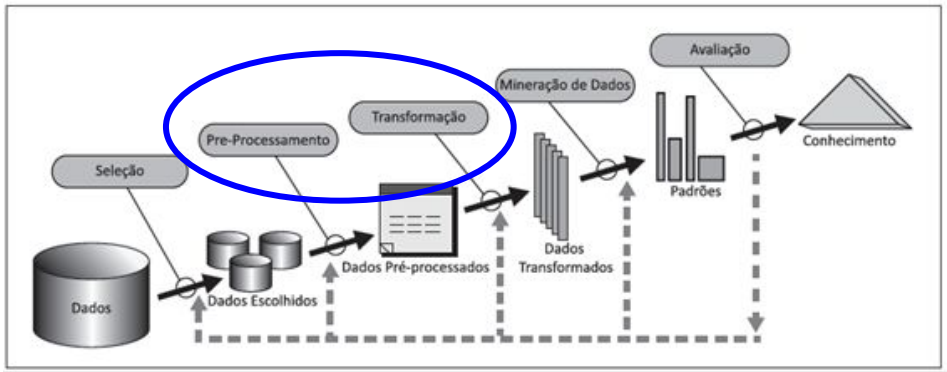
Throughout this work, it could be perceived, by empirical analysis, that the language created allows the diagramming of the phase of data selection to be used in EDM process, without the need for technological knowledge. In addition, the functional quality of the software was validated, as displayed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

Given what was brought and discussed, the work presented the following contributions:

- Expressive metamodel
- Cognitively effective notation
- Functionally adherent to the needs of the domain
- Simplification of the phase “selected data”

Future Work

- **to implement elements of transformation of the MDD:** Text to Model and Model to Text transformations in the following transformation functionalities.
- **to expand** the proposal to also carry out the **pre-processing phase of EDM**;
- **to complement the research validation**
- **to develop studies** of this proposal in the context of **Big Data and Data Lake.**



References

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.

C. Romero and S. Ventura, “Data mining in education,” Wiley In terdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 12–27, 2013.

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.

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.

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.

H. Krahn, B. Rumpe, and S. V¨olkel, “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.

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.

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.

P. N. Magalhães Júnior, “Um modelo de dados para apoiar a mineração de dados educacionais na investigac¸ ˜ao de evasão de estudantes,” tede.unifacs.br, 2013.

L. M. B. Manh˜aes, “Predic¸ ˜ao do desempenho acadˆemico de graduandos utilizando mineração de dados educacionais,” Doutorado em Engenharia de Sistemas e Computac¸ ˜ao 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.

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

E. M. L. V. Leijden, “Desenvolvimento de uma linguagem espec´ific de dom´inio para consumo de dados educacionais,” 2020.

C. W. Bachman, “The structuring capabilities of the molecular data model.” in ER, 1983, pp. 55–68.

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

I. ISO, “Iso standard 9126: Software engineering product quality, parts 1, 2 and 3,” 2001.

K. Qian, “Discovering information integration specifications from data examples,” Ph.D. dissertation, UC Santa Cruz, 2017.

Thank You!

Any Questions?
emlvl@ecomp.poli.br

AUTHORS:
Eronita Mari Luizines Van Leijden,
Andrêza Leite de Alencar and
Alexandre Magno Andrade Maciel

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