POLITECNICO DI TORINO

Master Degree Course in Computer Engineering

Master Degree Thesis

Integration of open e-learning datasets

Learning analytics on benchmark data

Supervisors
prof. Luca Cagliero
prof. Laura Farinetti

Candidate
Sabrina Camurati

December 2019

Summary

The scarcity and the heterogeneity of publicly available e-learning datasets are some of the main concerns for Learning Analytics research. For this reason, a common framework is needed to provide researchers with a richer information and a general schema, in which new datasets can be collected.

The educational context can vary depending on educational level, national policies and cultural characteristics, so the integration of datasets is challenging and requires a dynamic approach.

The thesis work focuses on the design of UNIFORM, an integrated open relational database for education, which guarantees modularity and flexibility: its structure can be dynamically edited, to allow the integration of new datasets.

The schema contains attributes and tables representing the data of learning datasets; it is designed to be general enough in order to include information from different learning institutions.

Acknowledgements

I would like to express my gratitude and appreciation for my thesis supervisor, Professor Luca Cagliero, for the proposal of this thesis; his availability and valuable feedback contributed a lot in the success of this work.

I gratefully acknowledge Professor Laura Farinetti, my thesis co-supervisor, for her recommendations and support throughout this study.

I thank Lorenzo Canale for his guidance and collaboration during the process of this thesis.

My sincere gratitude goes to my mother Susanna Tartara, to my father Albino Camurati and to my grandmother Annamaria De Ponti. It was their unconditional love and their support at every step of my life that enabled me to achieve my goals. Their believe and trust in me is the main cause of my success.

I always remember my grandfather Gianpietro Tartara, who passed away before the completion of my education. I know he would be proud and I will forever be grateful for the knowledge and values he instilled in me.

I would like to thank Professor Jorge Cordovez for his wise advice and his constant encouragement. His dedication and passion in teaching are admirable, together with his kindness and availability.

I am grateful to my colleagues and true friends Alessandra Chen , Cristiano Cavo and François-Xavier Renna for the wonderful times we shared. Their friendship and support are really important for me.

Contents

List of Tables IX

List of Figures X

1 Introduction 1
- 1.1 Motivation and context . . . 1
- 1.2 Problem statement . . . 2
- 1.3 Approach and methodology . . . 2
- 1.4 Contributions . . . 3
- 1.5 Document organization . . . 3

2 State of the art 5
- 2.1 Data integration . . . 5
- 2.2 Features adopted in Learning Analytics . . . 7

3 Datasets analysis 9
- 3.1 Datasets information . . . 11
- 3.2 Previous studies on the benchmark datasets . . . 17
  - 3.2.1 OULAD . . . 18
  - 3.2.2 COURSERA Forums . . . 18
  - 3.2.3 HarvardX and MITx . . . 19
  - 3.2.4 Portuguese Schools . . . 19
  - 3.2.5 xAPI . . . 19
  - 3.2.6 EPM . . . 19
  - 3.2.7 EDSA and ISTM . . . 19
  - 3.2.8 UoJ . . . 20
  - 3.2.9 POLITO . . . 20

4 UNIFORM schema 21
4.1 INSTITUTEs . . . 23
4.2 USERs . . . 24
4.3 USERs-INSTITUTEs . . . 26
4.4 USERs-COURSEs . . . 27
4.5 USERs-PRESENTATIONs . . . 28
4.6 COURSEs . . . 29
4.7 PRESENTATIONs . . . 29
4.8 ASSESSMENTs . . . 30
4.9 USERs-ASSESSMENTs . . . 31
4.10 EXERCISEs . . . 31
4.11 USERs-EXERCISEs . . . 31
4.12 LECTUREs . . . 32
4.13 USERs-LECTUREs . . . 32
4.14 VIDEOLECTUREs . . . 33
4.15 FORUMs . . . 33
4.16 THREADs . . . 34
4.17 POSTs . . . 34
4.18 FILEs . . . 35
4.19 ACTIVITYs . . . 35

5 Manual alignment 39
5.1 Methodology . . . 39
5.2 Extended UNIFORM schema . . . 40
5.2.1 MODULEs . . . 42
5.2.2 MODULEs-PRESENTATIONs . . . 43
5.2.3 USERs-MODULEs . . . 43

6 Evaluation 45
6.1 Objectives . . . 45
6.2 Hardware . . . 45
6.3 Manual alignment evaluation . . . 46
6.3.1 UNIFORM tables matched by each dataset . . . 46
6.3.2 Number of datasets matching a UNIFORM table . . . 47
6.3.3 Percentage of features matched per dataset . . . 48
6.4 Results . . . 51

7 Conclusions and future work 53
7.1 Conclusion 53
7.2 Future work 54

Bibliography 55

List of Tables

3.1 Datasets features. . . . 10
3.2 Datasets comparison. . . . 11
3.3 Datasets tasks. . . . 18
4.1 INSTITUTEs 24
4.2 USERs 25
4.3 USERs-INSTITUTEs 26
4.4 USERs-COURSEs 27
4.5 USERs-PRESENTATIONs 28
4.6 COURSEs 29
4.7 PRESENTATIONs 29
4.8 ASSESSMENTs 30
4.9 USERs-ASSESSMENTs 31
4.10 EXERCISEs 31
4.11 USERs-EXERCISEs 32
4.12 LECTUREs 32
4.13 USERs-LECTUREs 32
4.14 VIDEOLECTUREs 33
4.15 FORUMs 33
4.16 THREADs 34
4.17 POSTs 34
4.18 FILEs 35
4.19 ACTIVITYs 35
5.1 MODULEs 42
5.2 MODULEs-PRESENTATIONs 43
5.3 USERs-MODULEs 43
6.1 Percentage of matched attributes per UNIFORM table. . . . 49
6.2 Partial table related to correspondences between UNIFORM features and the considered datasets. . . . 50

List of Figures

3.1 General dataset information. 12
4.1 Datasets adopted for the modeling of UNIFORM schema. . . . 21
4.2 UNIFORM tables. 23
5.1 Alignment of UNIFORM and Test datasets 39
5.2 Extended UNIFORM tables 42
6.1 Number of UNIFORM tables matching a dataset 47
6.2 Number of datasets matching a UNIFORM table 48

Chapter 1

Introduction

1.1 Motivation and context

In recent years, the widespread use of Learning Management Systems in universities and schools has provided learning institutions with a huge amount of data related to students. Each institute can store information regarding demographics, students' outcomes, registration to course presentations, interactions with LMS[1], forum posts, answers to surveys,...

Learner-generated data can be processed through appropriate methods and tools to extract useful knowledge and supply interesting insights for the enhancement of the learning experience at every level of learning institution.

"Learning Analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs." [2]

E-learning datasets are appropriately transformed and used in Learning Analytics in order to fulfill a number of tasks: prediction of school dropout, prediction of students' outcomes, prediction of students at-risk of failure, students' engagement analysis, learning process patterns insights, early intervention [1].

[1]LMS: Learning Management Systems

[2]LA has been defined in the call for papers of the 1st International Conference on Learning Analytics and Knowledge (LAK 2011)

1.2 Problem statement

A large number of studies related to Learning Analytics adopt datasets which are not publicly available, as stated in [2]: institutions do not share their data, due to privacy issues and internal policies. For this reason, the results of these studies are not reproducible.

On the other hand, publicly available datasets are quite heterogeneous, as they may contain data related to: demographics, students' interaction with teaching material, students' outcomes, peer-to-peer interaction. The size and the schema of the datasets are largely variable. As mentioned in [3], when the datasets are considered individually, they may not contain enough data to provide significant results for the various research tasks.

Due to the aforementioned issues, it is arduous to find real benchmark data for the testing of learning analytics methods and algorithms.

1.3 Approach and methodology

The thesis work focuses on the design of a common framework to provide researchers with a richer information and a general schema, in which new datasets can be collected.

The first step is to collect publicly available datasets (see Chapter 3), which are then accurately analysed. Politecnico di Torino dataset, which is not publicly available, is included in this work. In fact, this dataset contains a variety of information that covers a considerable amount of data categories: student personal information, career data, assessments, courses information, interactions with LMS, online resource access, videolectures streaming and download. Videolectures information is not available in other considered data sources, so this dataset gives an important contribution to the design of the schema.

Ten publicly available datasets related to education are selected and analysed, together with POLITO dataset. The complete list of the features belonging to the data sources is then written in detail, including their description.

The general schema, UNIFORM (see Chapter 4), is generated using the features from a subset of the considered datasets. The alignment process is initially focused on the attributes that are present in most of the datasets (e.g.: demographics, course information), then the features that are peculiar of specific datasets are added to the schema (e.g.: posts from COURSERA Forums).

After the definition of the general schema, it is necessary to test its generality and modularity. To manually align (see Chapter 5) new datasets with the UNIFORM schema, for each attribute in the source dataset it is necessary to look for an approximated match with UNIFORM. If a match is not found, then a new attribute (and eventually table) is created in the extended dataset version to represent the corresponding information.

1.4 Contributions

The design of UNIFORM, an integrated open relational database for education, guarantees modularity and flexibility: its structure can be dynamically edited, to allow including new datasets.

UNIFORM contains attributes and tables representing the e-learning data of publicly available datasets. It is designed to be general enough to include data from different learning institutions (independently from educational level, national policies, cultural characteristics).

The thesis work focuses on the design of a common framework to provide researchers with a richer information and a general schema, in which new datasets can be collected.

1.5 Document organization

The document is divided in 5 chapters:

Chapter 2 describes the state of the art, with respect to Data integration (2.1) and Features adopted in Learning Analytics (2.2).

Chapter 3 describes the datasets which have been collected and analysed: dataset information (3.1) and related work (3.2).

Chapter 4 outlines the process for the creation of the UNIFORM schema and provides complete information regarding the tables and attributes.

Chapter 5 analyses the Manual alignment process, showing the method that can be used for the addition of new attributes and tables.

Chapter 6 describes the evaluation of the proposed UNIFORM schema.

Chapter 2

State of the art

Data integration is necessary in order to appropriately manage huge amounts of heterogeneous data. The adoption of common schema and format allows to gather information from different data sources in a unique dataset and to provide a more efficient use of data. While economy and industry have invested a on this research field, learning institutions have not yet put a lot of effort into this task. Section 2.1 describes the *State of the art* with respect to data integration in the learning context.

When dealing with e-learning datasets, it is also important to understand which kind of data is being studied by Learning Analytics. It is interesting to analyse the features adopted in papers, because it gives an overview on the information that can be used for the fulfillment of research tasks. Section 2.2 describes the *State of the art* with respect to the features adopted in Learning Analytics.

2.1 Data integration

In recent years, the scope of data integration has expanded: it is no longer an area for mere intellectual curiosity, but a real necessity. Data integration has been widely adopted by large enterprises, that own a huge amount of data sources. The interest in data integration has grown also in the learning environment, due to the adoption of LMSs in learning institutions and the introduction of MOOCs: it is important to aggregate heterogeneous data, in order to provide a unique framework that gathers all the information.

While economy and industry have invested on data integration, learning

institutions have not yet put a lot of effort into this task. Only a few research papers address data integration in the learning context.

[4] reviews various publications, describing the current state of Learning Analytics with respect to data integration. Collecting and merging data from different data sources allows more complete and accurate analysis, since Learning Analytics research tasks need huge amounts of data related to various aspects of the learning environment [5].

Combining data from different data sources is a time-consuming operation, so the vast majority of research papers analyse data separately without integrating datasets.

Only the authors of [6] specify that they adopt manual integration, while others do not explicitly describe the integration method (based on the type of data used, it is supposed they choose a manual approach).

Just a few of the considered studies use automatic tools for data integration: they include tools developed ad hoc for a specific research project [7], Business Intelligence software [8], SQL [9] and R scripts [10].

However, there are some limitations: in [9] data sources are related to the same institution, in [11] data integration is managed using only two e-learning platforms, in [12] the operation is restricted to EdX MOOC data with other EdX data (for Coursera MOOC data this is not addressed).

Generally, the integration of different datasets is performed using only two different data sources. Only some research papers adopt at least three different datasets. When integrating various data sources, researchers tend to prefer data that is already available in a common format. Only six of the twenty analysed papers combine data from different datasets, without having the same format.

Notice that it is a challenge to collect and merge multiple datasets due to their different format and structure; the reproducibility of previous experiments is usually tough, due to the lack of descriptions about the process of data integration.

2.2 Features adopted in Learning Analytics

A particular attention has been paid to investigate the relevance of different types of features in Learning Analytics.

In [13], the prediction of at-risk students is performed using a dataset from higher education and the other from a K-12 online school. The authors identify common predictors between the two datasets and underline the importance of relative engagement variables, which allow richer insights comparing student's interactions and results with respect to all the students registered to the same course presentation.

[14] uses features related to students' interactions with the online resources for the prediction of students' performance. A transfer learning model is applied to the current course presentation, after the pre-training with data of the former course presentation.

As students' dropout is one of the main concerns, [15] uses a temporal multi-objective model to find the earliest time in which a reliable prediction can be obtained regarding students' risk of dropping out: the model is based on features related to students' performances, without considering demographics.

[16] presents a multi-view early warning system that allows to alert students at-risk of failure and dropout, integrating many student data repositories; it does specifically target underrepresented student populations such as adult learner, freshman, first-generation, transfer and international students, which usually are underperforming. The features used for the model include data related to ethnicity, gender, residency and other personal information (adult learner, freshman, transfer, first-generation), as well as dynamic features such as students' interactions with online resources.

Five different Moodle datasets belonging to university, school teaching and online training academies are analysed in [17] for the prediction of students at-risk of not submitting assignments on time in online courses. The features used by the one-size-fits-all network are related to students' activity, course and assignment information and peers activity.

In [18] students' dropout risk prevention is performed in real online e-learning environment through an early warning system, to prevent students from leaving the university with appropriate intervention. Demographics, academic results and interactions with online resources are combined to generate derived features (absolute, relative, aggregate,...).

Early prognosis of student performance is done in [19] using attributes divided in two views: features related to students' demographics and academic results, features automatically recorded about students' online activity in the course Learning Management System.

As a matter of fact, the type of information adopted for the various research tasks is pretty variable. Some papers focus on static variables, while others prefer dynamic features or a combination of both. The data is quite heterogeneous, so some datasets may be more adapt than others for specific analysis based on different kind of information (e.g.: demographics, outcomes, interactions,...).

Notice that the datasets used in the considered research papers are not publicly available, the authors provide only a brief description of the datasets and the adopted features. The experiments are not reproducible due to the lack of shared data, as previously mentioned.

Chapter 3

Datasets analysis

The first step is to collect datasets from the Web: this research is performed through a number of Dataverse[1] (e.g.: DataShop, Kaggle, ResearchGate, UCI Machine Repository, HarvardX Dataverse,...).

POLITO (Politecnico di Torino) dataset, which is not publicly available, is included in this work. In fact, this dataset contains a variety of information that covers a considerable amount of data categories: student personal information, career data, assessments, courses information, interactions with LMS, online resource access, videolectures streaming and download. Notice that videolectures information is not available in other considered datasets, so **POLITO** dataset gives an important contribution to the design of the schema.

Ten publicly available datasets related to education are selected and analysed, together with **POLITO dataset**. A particular attention is paid on the kind of information that can be stored in e-learning datasets: six main categories of data are identified.

The content of the datasets is synthesised with the following acronyms:

- a) **SPD** (**Student Personal Data**),
  e.g. identification number, gender, date of birth, ethnicity, place of birth, residence place;
- b) **SCD** (**Student Career Data**),

[1]Dataverse meaning: https://en.wikipedia.org/wiki/Dataverse

e.g. school degrees, entry test grades, final grade, educational modules enrollment;

- c) **EMD** (**Educational Module Data**),
  e.g. available courses, course description, course prerequisites, course duration;
- d) **SAD** (**Student Assessment Data**),
  e.g. exam grades, intermediate assessment evaluations;
- e) **ERA** (**Educational Resource Access**),
  e.g. students' activities within a LMS, online resources access, videolectures streaming and download;
- f) **IAD** (**Interaction Activity Data**),
  e.g. forum posts, peer-to-peer interactions, student-teacher interactions.

Table 3.1 gives a detailed view about the features types that are contained in each dataset, with respect to the aforementioned categories.

Table 3.1. Datasets features.

	POLI (1)	**EDSA** (2)	**EPM** (3)	**HARV** (4)	**ISTM** (5)	**MITx** (6)	**OUL** (7)	**COUR** (8)	**PORT** (9)	**XAPI** (10)	**UOJ** (11)
Student Personal Data	X			X	X	X	X		X	X	X
Student Career Data	X			X	X	X	X		X	X	X
Educational Module Data	X										
Student Assessment Data	X		X	X	X	X	X		X		
Educational Resource Access	X	X	X	X		X	X				
Interaction Activity Data				X		X		X			

In Table 3.2, the characteristics of each dataset are compared: the types of stored data, the size expressed in MB and the number of tables. The analysed publicly available datasets are quite heterogeneous, in terms of schema, focus and complexity.

Table 3.2. Datasets comparison.

	POLI (1)	**EDSA** (2)	**EPM** (3)	**HARV** (4)	**ISTM** (5)	**MITx** (6)	**OUL** (7)	**COUR** (8)	**PORT** (9)	**XAPI** (10)	**UOJ** (11)
type of data	SPD, SCD, EMD, SAD, ERA	ERA	SAD, ERA	SPD, SCD, SAD, ERA, IAD	SPD, SCD, SAD	SPD, SCD, SAD, ERA, IAD	SPD, SCD, SAD, ERA	IAD	SPD, SCD, SAD	SPD, SCD	SPD, SCD
size (MB)	122.6	7.7	19.3	70.2	0.2	12.5	464.4	70.5	0.1	0.1	5.0
num. of tables	7	1	5	1	2	1	7	3	2	1	13

3.1 Datasets information

Ten publicly available datasets related to education are seected and analysed, together with **POLITO dataset**:

- **OULAD**[2] (Open University Learning Analytics Dataset), which contains students' demographic information, assessment results and data related to the interactions with a LMS [3]. As described in [20], the data is related to 22 courses and 32,593 students, from the years 2013 and 2014 at the Open University. There are 7 *.csv* files:
  - **assessments.csv**: it contains data related to assessments and final exams in course presentations, the file consists of 206 rows;
  - **courses.csv**: it contains the list of all courses and presentations, the file consists of 22 rows;
  - **studentAssessment.csv**: it contains the results of students' assessments (usually, final exam results are missing), the file consists of 173,912 rows;
  - **studentInfo.csv**: it contains demographics and students' final result in each course they studied, the file consists of 32,593 rows;
  - **studentRegistration.csv**: it contains the time when the students registered for (and eventually unregistered from) a course presentation, it consists of 32,593 rows;
  - **studentVle.csv**: it contains students' interactions with the LMS, the file consists of 10,655,280 rows;

[2]OULAD dataset: https://bit.ly/2m4a0NF

[3]LMS: Learning Management Systems

– **vle.csv**: it contains data related to the online teaching material provided by the LMS, the file consists of 6,364 rows.

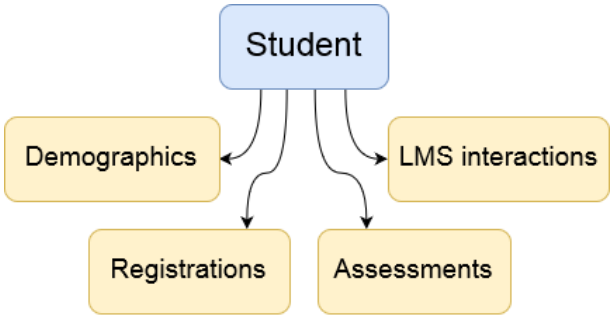

Figure 3.1. General dataset information.

- **HarvardX**[4] and **MITx**[5] Dataverse, which contains students' demographic information and data related to the activities in a edX platform course. There are 2 *.csv* files:
  - **HMXPC13_DI_v2_5-14-14.csv**: it contains data related to HarvardX and MITx, from Fall 2012, Spring 2013, and Summer 2013, the file consists of 641,138 rows;
  - **cs_MITx.csv**: it contains data related to MITx courses from Fall 2012, the file consists of 59,279 rows.
- **COURSERA Forums**[6] contain the users' posts from the discussion threads belonging to the forums of 60 Coursera MOOCs[7], in 4 different languages. There are 3 *.csv* files:
  - **course_information.csv**: it contains information of the 60 courses, the file consists of 60 rows;
  - **course_threads.csv**: it contains information about threads and the related sub-forums, the file consists of 99,629 rows;
  - **course_posts.csv**: it contains the users' posts and comments of the forums, the file consists of 739,074 rows.

[4]HarvardX dataset: https://bit.ly/2FLEz3f

[5]MITx dataset: https://bit.ly/314niIv

[6]COURSERA dataset: https://bit.ly/2mVuOas

[7]MOOCs: Massive Open Online Courses

- **Portuguese Schools** dataset[8], which contains students' performance, demographic and lifestyle data in two secondary schools from the Alentejo region of Portugal, during the academic year 2005-2006. The information was collected through school reports and students' answers to questionnaires, as described in [21]: students' alcohol consumption, Internet connection, romantic relationships, nursery attendance, health status, parents education and job. There are 2 *.csv* files:
  - **student-mat.csv**: it contains students' data related to the subject "Mathematics", the file consists of 395 rows;
  - **student-por.csv**: it contains students' data related to the subject "Portuguese language", the file consists of 649 rows.
- **xAPI**[9] (experience API) dataset, which contains students' performance, demographic and behavioral data in the University of Jordan, along with information about parents involvement in the learning process. There is 1 *.csv* file:
  - **xAPI-Edu-Data.csv**: it contains students' information, the file consists of 480 rows.
- **EPM**[10] (Educational Process Mining) dataset, which contains data referred to students' grades and behavior during interactions with a simulation environment named Deeds (Digital Electronics Education and Design Suite), adopted for e-learning in digital electronics, at the University of Genoa. There are 3 files and a folder:
  - **logs.txt**: it contains students' log data for the 6 laboratory sessions, the file consists of 115 rows;
  - **final_grades.xlsx**: it contains students' final outcomes, the file consists of 52 rows in the first sheet and 62 rows in the second sheet;
  - **intermediate_grades.xlsx**: it contains students' assignments grades per lab session, the file consists of 115 rows;
  - **Processes**: it contains students' information related to the 6 laboratory sessions, the folder consists of 6 sub-folders (containing totally 520 files);

[8]Portuguese Schools dataset: https://bit.ly/2lmoFDC

[9]xAPI dataset: https://bit.ly/2lmp2y0

[10]EPM dataset: https://bit.ly/2ltgwgU

- **EDSA**[11] (European Data Science Academy) dataset, which contains data about students' interactions with online resources in the European Data Science Academy portal. There is 1 *.csv* file:
  - **EDSAOnlineCoursesLA.csv**: it contains students' information, the file consists of 22,506 rows.
- **ISTM**[12] (International Students Time Management) dataset, which contains students' answers to survey questions about time management at Nottingham Trent International College. There are 2 *.csv* files:
  - **International students Time management data.csv**: it contains students' answers to surveys, the file consists of 125 rows;
  - **Sheet2.csv**: it contains survey questions, the file consists of 12 rows.
- **UoJ**[13] (University of Jisc) dataset, which contains information about students' performance. Data is entirely fictious, it has been created through a basic simulation (in fact the University of Jisc is not real). Anyway, this dataset is included together with the others, because the main goal is to analyse data structures and to identify a general schema. There are 13 *.json* files:
  - **course.json**: it contains information about courses, the file consists of 8 rows;
  - **courseinstance.json**: it contains information about course presentations, the file consists of 24 rows;
  - **institute.json**: it contains information about the learning institution, the file consists of 1 row;
  - **module.json**: it contains information about course modules, the file consists of 144 rows;
  - **moduleinstance.json**: it contains information about modules presentations, the file consists of 144 rows;
  - **modulevlemap.json**: it contains the mapping between the module presentation and the vle module identifier, the file consists of 144 rows;

[11]EDSA dataset: https://bit.ly/2mc0NTG

[12]ISTM dataset: https://bit.ly/2me1HyT

[13]UOJ dataset: https://bit.ly/2mxrq5L

  - **staff.json**: it contains teachers' information, the file consists of 20 rows;
  - **staffcourseinstance.json**: it contains information about teachers assigned to course presentations, the file consists of 40 rows;
  - **staffmoduleinstance.json**: it contains information about teachers assigned to module presentations, the file consists of 144 rows;
  - **student.json**: it contains students' information, the file consists of 1,000 rows;
  - **studentassessmentinstance.json**: it contains students' assessments results, the file consists of 17,458 rows;
  - **studentcourseinstance.json**: it contains students' information related to course presentations, the file consists of 1000 rows;
  - **studentmoduleinstance.json**: it contains students' information related to module presentations, the file consists of 6000 rows;

- **POLITO** (Politecnico di Torino)[14] dataset, that contains information related to first year Bachelor of Science students' performance and interactions with online learning resources, including videolectures. There are 7 files:
  - **Accessi_Download**: it contains students' downloads of videolectures, the file consists of 11,850 rows;
  - **Accessi_Materiale**: it contains students' accesses to teaching material, the file consists of 1,048,575 rows;
  - **Accessi_Streaming**: it contains students' accesses to videolectures via streaming, the file consists of 314,604 rows;
  - **Anagrafica**: it contains demographic information about the students, the file consists of 4,304 rows;
  - **Carriera**: it contains data related to students' career, the file consists of 20,565 rows;
  - **Corsi**: it contains the data related to the videotaped courses of the first year, the file consists of 10 rows;
  - **Esami**: it contains students' exam results, the file consists of 15,620 rows.

[14]Politecnico di Torino dataset: not publicly available.

In order to have the same file type for all the datasets, the *.csv* format is chosen. While most of the datasets already are in the chosen format, some of them need a little pre-processing:

- **UoJ** *.json* files are converted into *.csv*, using a Json To Csv converter [15] and then redundant data is manually removed.
- **EPM** files related to the laboratory sessions from the Process folder are merged into a unique *.csv* file . **EPM** file related to final exams is divided into two different *.csv* files.
- **POLITO** dataset is stored in *pickle* files, so they are converted into *.csv* files.

[15] Json To Csv converter: http://www.convertcsv.com/json-to-csv.htm

3.2 Previous studies on the benchmark datasets

The majority of studies related to Learning Analytics adopt datasets which are not publicly available: institutions do not share their data, due to privacy issues and internal policies. For this reason, the results of these studies are not reproducible.

However, a few scientific papers use publicly available datasets for their research works. Table 3.3 synthesizes, for each of the datasets described in section 3.1, research papers that used them for a specific task.

The research tasks analysed in these papers are:

- prediction of students' dropout,
- prediction of at-risk students,
- prediction of students' performance,
- students' engagement analysis,
- learning process insights.

Considering that the above cited datasets are used for many tasks, integrating them in a unique framework means to deal with an information which is suitable for the most important learning scopes.

The following subsections describe, for each datasets, the literature papers that adopted them for various research tasks.

Some datasets are analysed in many papers, due to their dimensions and versatility. Instead, some small and less popular datasets are not studied. One of the considered datasets does not contain real data, so it is not adopted for students' prediction.

Table 3.3 contains references to the research papers that used the considered publicly available datasets for Learning Analytics. When there is no published work related to the dataset for a certain task (at the best of our knowledge), but it is suitable for that scope, the cell contains "S".

Table 3.3. Datasets tasks.

	POLI	EDSA	EPM	HARV	ISTM	MITx	OUL	COUR	PORT	XAPI	UOJ
prediction of students' dropout				[22]		[22]	[23]	[24],[25]			
prediction of at-risk students	S				S		[26]	[27]	S	S	
prediction of students' performance	S				S		[23]	S	[21]	[28]	
students' engagement analysis		S		[29]		[29]				[30]	
learning process insights	S		[31]								S

3.2.1 OULAD

Open University Learning Analytics Dataset contains data which is related to students' interactions with the Virtual Learning Environment, so it is adopted in various papers, due to its versatility:

- In [23], **OULAD** dataset is used for *Students' outcomes prediction*.
- The authors of [26] perform early prediction of *Students at-risk of failure* using **OULAD**, in order to provide timely interventions to students.
- **OULAD** is also analysed for *Students' dropout prediction* in [23].

3.2.2 COURSERA Forums

COURSERA Forums contain textual data related to students' and teachers' posts in COURSERA MOOCs. This peculiar kind of information is adopted in some papers:

- Sentiment analysis is applied in [24] and in [25] to the text information contained in **COURSERA** users' posts for the *prediction of student attrition*.
- **COURSERA** data is used for *early prediction of students at-risk of failure* task, such as in [27].

3.2.3 HarvardX and MITx

HarvardX and **MITx** datasets have a similar structure, so they are often analysed together:

- In [29], **HarvardX** and **MITx** allow to evaluate *Students' differences in enrollment to and course completion* in STEM [16] MOOCs, based on nationality and gender.
- *Students' dropout prediction* task is performed in [22] adopting **HarvardX** and **MITx**.

3.2.4 Portuguese Schools

In [21] the authors use **Portuguese Schools** data associated to students' behavioral and lifestyle information as well as parents' education level and job for *Students' outcomes prediction*.

3.2.5 xAPI

Experience API dataset is adopted for *Students' academic performance prediction*: [28] considers features related to students' learning behavior (raised hand on class, participating in discussions groups, viewing of online resources,...), while [30] adopts features regarding students' punctuality in the class and parents' involvement in the learning process.

3.2.6 EPM

Educational Process Mining dataset, which contains information about students' grades and behavior during interactions with online resources, is analysed in [31] for the study of the *Correlation between learning path and academic performance*.

3.2.7 EDSA and ISTM

European Data Science Academy and **International Students Time Management** datasets are focused on a specific area: the former includes information about users' interaction with online resources, while the latter

[16] STEM is the acronym of Science, Technology, Engineering and Mathematics.

contains students' answers to survey questions about the time management during the course.

3.2.8 UoJ

As described in 3.1, **University of Jisc** dataset is entirely fictious, so it is not adopted for real data analysis. Anyway, this dataset is included together with the others, because the main goal is to analyse data structures and to identify a general schema.

3.2.9 POLITO

Politecnico di Torino dataset can be used for several kinds of research tasks, as it contains students' demographics, information about registration to course presentations, exam scores and interaction with online resources (which also include videolectures).

Chapter 4

UNIFORM schema

The considered learning datasets are stored in different formats and structures, so a uniform schema is needed in order to generalize all the information contained in the datasets. Storing data in a unique framework can improve the usability of data and allow researchers to easily add new datasets to the schema.
The data sources are divided in two subsets:

- the former is analysed and adopted to build a general schema;
- the latter is used for testing the capability of the schema to match features related to other datasets.

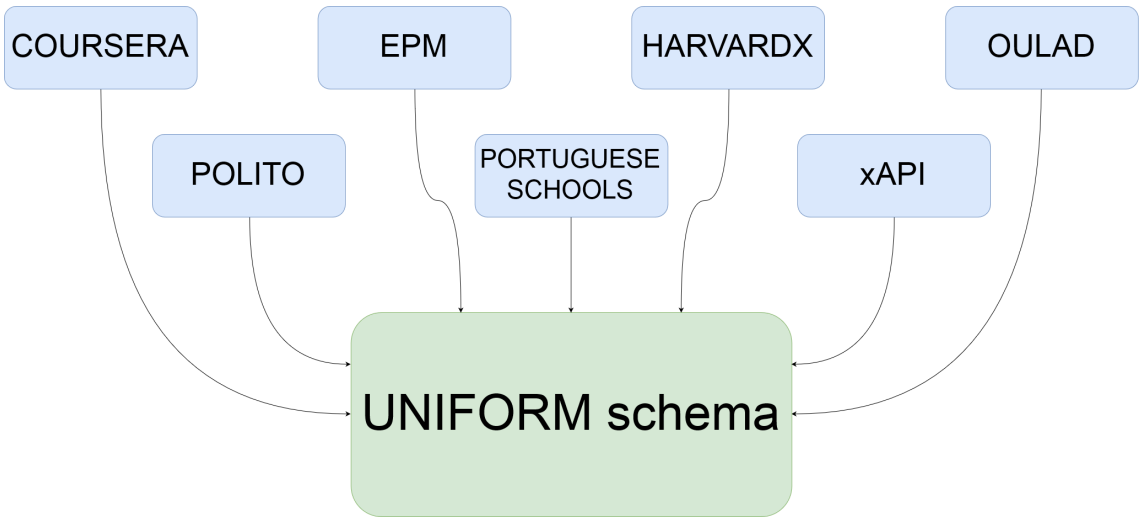

Figure 4.1. Datasets adopted for the modeling of UNIFORM schema.

As in Figure 4.1, **UNIFORM** schema is generated considering only 7 of the public datasets available: **POLITO, EPM, HarvardX, OULAD, COURSERA, Portuguese Schools** and **xAPI-Edu-Data**.

All the datasets are converted into *.csv* files, as described in 3.1, in order to have a common format. For each dataset, the complete list of the features is written in detail, including their description (when it is not provided, a sentence describing the attribute is added) and the index type (A: it is the key of the table, B: it is part of the composite key of the table, C: it is not part of a key).

The alignment process is initially focused on the features that are present in most of the datasets.

- demographic information;
- institute data;
- course and presentation information;
- students' registration to courses and presentations.

After the alignment of the common attributes, the features that are peculiar of a specific dataset are added to the schema.

- **COURSERA Forums** contain forums, threads and users' posts;
- **EPM** dataset contains specific students' exercises data;
- **OULAD** and **POLITO** datasets contain teaching material information;
- **POLITO** dataset contains data related to videolectures;
- **Portuguese Schools** dataset contains data related to students' lifestyle;
- **xAPI, POLITO** and **EPM** datasets contain information related to lectures;
- **xAPI** dataset contains information related to users behavior during lectures.

During this process the features are divided in several categories: **Institutes, Users, Courses, Presentations, Lectures, Videolectures, Assessments, Exercises, Forums, Threads, Posts, Files** and **Activities**.

After all the features and tables have been completed, the UNIFORM schema consists of 19 tables (see Figure 4.2).

For each attribute and table, a description (in English) and a Bag-of-words[1] are written in detail (both English and Italian words are included, in order to match words from both languages: id, identifier, user, student, teacher,, matricola, utente, studente, insegnante).

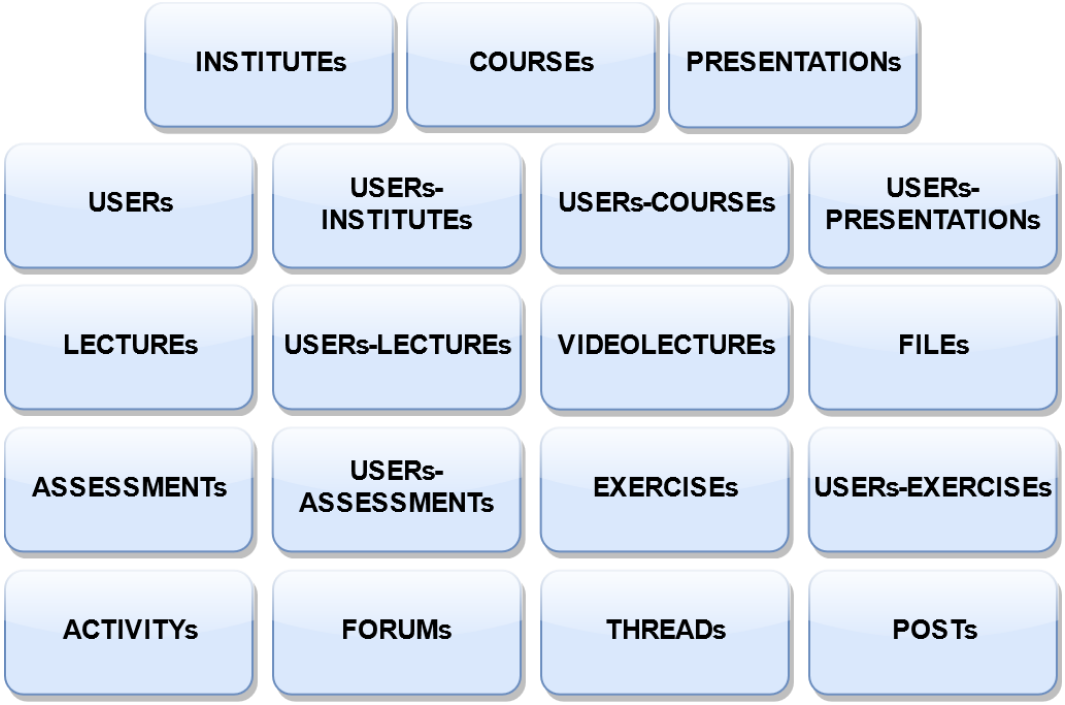

Figure 4.2. UNIFORM tables.

The completeness of the schema is then evaluated (see Chapter 5) using the other 4 as *Test datasets* : **EDSA, ISTM, MITx** and **UOJ**.

4.1 INSTITUTEs

Table 4.1 contains the information related to a learning institution.

[1]Bag-of-words: https://en.wikipedia.org/wiki/Bag-of-words_model

The attributes include the name of the institute, its location, the educational level (e.g.: Primary School, High School, College, University,...) and the type of institute.

In order to avoid redundancies, the Entry Grade Base and the Final Grade Base are inserted in this table, instead of replicating the information for each student of an institution.

Table 4.1: INSTITUTEs

attribute/feature	**description**
Institute_Id	A unique identification code of the institute
Name	Name of the institute
Place	Place in which the institute is located
EduLevel	Educational level of the institute
Type	Type of institute
FinalGradeBase	Student's final grade base
EntryGradeBase	Student's entry test grade base

4.2 USERs

Table 4.2 contains person-specific characteristics, including demographic, socio-economic, life-style and health information.

The attributes include birth date, birth and residence place, gender, educational level and user-provided data (e.g.: Internet access, Alcohol consumption, Free time, Romantic relationship, Family,...). Data related to parents' job and educational level is also inserted in this table.

Notice that users' name, detailed residence address and other sensitive information can not be included in the dataset. It is necessary to preserve users' privacy.

Instead of creating STUDENTs and TEACHERs tables, a general table is preferred: table USERs can include both students and teachers (the User type is then specified in 4.3).

The choice is driven by the following reasons:

- students and teachers do have similar personal data;
- a user may be student in an institute, while being teacher in another institute;

- a user may teach in the same institute where it has been student in the past.

Table 4.2: USERs

attribute/feature	**description**
User_Id	A unique identification code for the user
Birth_Time	User's date of birth
Gender	User's gender
Disability	It indicates whether the user has declared a disability
Birth_Place	User's place of birth
Birth_Place_Type	User's place of birth type
Residence_Place	User's residence place
Residence_Place_Type	User's residence place
ImdBand	It specifies the Index of Multiple Deprivation band of the place where the user lived during the module presentation
Education_Level	Highest level of education completed
Nationality	User's nationality
Familysize_Count	Family size
ParentStatus	Parent's cohabitation status (living together apart)
Mother_Education_Level	Mother's highest level of education completed
Father_Education_Level	Father's highest level of education completed
Mother_Job	Mother's job
Father_Job	Father's job
NurseryAttendence	User has attended nursery school
InternetHomeAccess	User has Internet access at home
RomanticStatus	User is in a romantic relationship
FamilyRelQuality	Quality of family relationships
GoingOut_Duration	Time spent by the user going out with friends
AlchoolWorkdayConsuption	Workday alcohol consumption
AlchoolWeekendConsuption	Weekend alcohol consumption
HealtStatus	User's current health status
FreeTimeQuantity	Quantity of free time after school

4.3 USERs-INSTITUTEs

Table 4.3 contains data related to a user for a specific learning institution.

User type (student, teacher) is specified in this table, because a user may be student in an institute and teacher in another.

The attributes include registration time, studied credits, entry and final grade, course of study.

Table 4.3: USERs-INSTITUTEs

attribute/feature	**description**
User_Id	An identification code for the user
Institute_Id	Identification code of the User's institute
User_Type	Type of user (student, teacher, staff)
Guardian	Student's guardian
Familysupport	Family educational support
ExtraEduSupport	Extra educational support
ChoiceReason	Reason to choose this institute (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
HToSTravel_Duration	Travel time from home to school
StudiedCredits	The total number of credits for the modules the student is currently studying
Final_Grade	User's final grade
Entry_Grade	User's entry test grade
Registration_Time	User's date of registration to the course of study
Unregistration_Time	User's date of unregistration from the course of study
Cds	User's course of study
StudentLevel	The Users are classified into intervals based on their total grade/mark (Low-Level, Middle-Level, High-Level)
Higher	It indicates whether the student wants to take higher education
ParentAnsweringSurvey	parent answered the surveys which are provided from school or not (nominal:'Yes','No')
ParentschoolSatisfaction	the Degree of parent satisfaction from school(nominal:'Yes','No')

User_Grade	Grade student belongs to (nominal: ‘G-01’, ‘G-02’, ‘G-03’, ‘G-04’, ‘G-05’, ‘G-06’, ‘G-07’, ‘G-08’, ‘G-09’, ‘G-10’, ‘G-11’, ‘G-12 ‘)

4.4 USERs-COURSEs

Table 4.4 contains data related to a user for a specific course.

The attributes include number of failures and user interactions with teaching material at course level.

Table 4.4: USERs-COURSEs

attribute/feature	**description**
User_Id	An identification code for the user
Course_Id	An identification code for a course on which the user is registered
Failures_Count	Number of past class failures
Events_Count	Number of interactions with the course, recorded in the tracking logs
InteractingDays_Count	Number of unique days user interacted with course
PlayVideo_Count	Number of play video events within the course
InteractingChapters_Count	Number of chapters with which the student interacted
ForumPosts_Count	Number of posts to the Discussion Forum
ViewedDashboard	Anyone who accessed the ‘Courseware’ tab (the home of the videos, problem sets, and exams) within the edX platform for the course
Certified	Anyone who earned a certificate: certificates are based on course grades
MandatoryPosts_Count	Minimum number of posts to receive credits for activity in the forums
ViewedCourseContent_Count	how many times the student visits a course content(numeric:0-100)
ViewedAnnouncements_Count	how many times the student checks the new announcements (numeric:0-100)
DiscussionGroups_Count	how many times the student participate on discussion groups (numeric:0-100)

4.5 USERs-PRESENTATIONs

Table 4.5 contains data related to a user for a specific course instance.

The attributes include number of absences, weekly study duration, registration time and user interactions with teaching material at course presentation level.

Table 4.5: USERs-PRESENTATIONs

attribute/feature	**description**
Presentation_Id	An identification code for a course presentation
User_Id	An identification code for the user
WeeklyStudy_Duration	Weekly study duration
ExtraPaidClasses	Extra paid classes within the course subject
ExtraCVActivitiies	Extra curricular activities
Absences_Count	The number of absence days for each student
Group	User's group for a specific course
Registration_Time	Date of course registration
Unregistration_Time	Date of course unregistration
LastInterction_Time	Date of last interaction with course
Events_Count	Number of interactions with the course, recorded in the tracking logs
InteractingDays_Count	Number of unique days student interacted with course
PlayVideo_Count	Number of play video events within the course
InteractingChapters_Count	Number of chapters with which the student interacted
ForumPosts_Count	Number of posts to the Discussion Forum
ViewedDashboard	Anyone who accessed the 'Courseware' tab (the home of the videos, problem sets, and exams) within the edX platform for the course
Explored	Anyone who accessed at least half of the chapters in the courseware
PartecipationSessions_Array	Array containing the participation to sessions
ViewedCourseContent_Count	how many times the student visits a course content(numeric:0-100)
ViewedAnnouncements_Count	how many times the student checks the new announcements (numeric:0-100)

DiscussionGroups_Count	how many times the student participate on discussion groups (numeric:0-100)

4.6 COURSEs

Table 4.6 contains the attributes related to a course.

In order to avoid redundancies, the number of credits related to the course is included in this table.

Table 4.6: COURSEs

attribute/feature	**description**
Course_Id	A unique identification code for a course
Institute_Id	Identification code of the institute
Name	Course name
Credits	Exam weight in terms of academic credits
Typology	Type of course

4.7 PRESENTATIONs

Table 4.7 contains data related to a course instance.

The attributes include course presentation duration, language, number of lectures, start and end time.

Teacher identifier for the course presentatation is also added.

Table 4.7: PRESENTATIONs

attribute/feature	**description**
Presentation_Id	An identification code for a course presentation
Course_Id	The identification code for the course to which the presentation is related
Lang	Language of the course presentation
Semester	School year semester
Duration	Duration of the module-presentation
Lectures_Count	Total number of lectures for the considered course
Start_Time	Start time of the course presentation

End_Time	End time of the course presentation
User_Id	A unique identification code for the teacher in the course presentation

4.8 ASSESSMENTs

Table 4.8 contains the information related to an assessment.

An assessment may include one or more exercises.

In order to avoid redundancies, the grade base, start time and weight are inserted in this table.

Table 4.8: ASSESSMENTs

attribute/feature	**description**
Assessment_Id	A unique identification code for the assessment
Institute_Id	Identification code of the institute
Type	Type of assessment
Course_Id	An identification code of the course (Course_Id) to which the assessment is related
Presentation_Id	An identification code of the presentation (Presentation_Id) to which the assessment is related
Lecture_Id	An identification code of the lecture (Lecture_Id) to which the assessment is related
GradeBase	Assessment base grade
Expiration_Time	The final submission date of the assessment for the module-presentation
Weight	Weight of the assessment in percentage (typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%)
Start_Time	The start date of the assessment for the module-presentation

4.9 USERs-ASSESSMENTs

Table 4.9 contains the data related to a student's assessment, including the type (final exam, test).

The attributes include student identifier, grade and submission time.

Table 4.9: USERs-ASSESSMENTs

attribute/feature	description
Assessment_Id	An identification code for the assessment
User_Id	An identification code for the user
Grade	Assessment grade obtained by the student
Submission_Time	The student's submission date of the assessment for the module-presentation
IsBanked	A status flag indicating that the assessment result has been transferred from a previous presentation

4.10 EXERCISEs

Table 4.10 contains the information related to an exercise, belonging to an assessment.

In order to avoid redundancies, the grade base is inserted in this table.

Table 4.10: EXERCISEs

attribute/feature	description
Exercise_Id	A unique identification code for the exercise
Assessment_Id	An identification code for the assessment
GradeBase	Assessment base grade

4.11 USERs-EXERCISEs

Table 4.11 contains the data related to a student's exercise.

The attributes include student identifier and grade.

Table 4.11: USERs-EXERCISEs

attribute/feature	description
Exercise_Id	An identification code for the exercise
User_Id	An identification code for the user
Grade	Exercise grade obtained by the student

4.12 LECTUREs

Table 4.12 contains the attributes related to a lecture.

The attributes include the teacher identifier for the lecture and the lecture type (lecture, laboratory).

Table 4.12: LECTUREs

attribute/feature	description
Lecture_Id	A unique identification code for a lecture
Presentation_Id	An identification code for a course presentation
User_Id	A unique identification code for the teacher
Lecture_Type	The type indicates if the lecture is laboratory or an in-class presentation
Order	The position number of the lecture, with respect to other lectures.

4.13 USERs-LECTUREs

Table 4.13 contains the data related to a user participation to a lecture.

The attributes include the student identifier and the students' interactions during the lecture.

Table 4.13: USERs-LECTUREs

attribute/feature	description
Lecture_Id	An identification code for a lecture
User_Id	An identification code for the student
Participation	It indicates whether the student participated to the lecture

RaisedHands_Count	how many times the student raises his/her hand on classroom (numeric:0-100)

4.14 VIDEOLECTUREs

Table 4.14 contains the information related to a videolecture.
The attributes include the teacher identifier and the recording time.

Table 4.14: VIDEOLECTUREs

attribute/feature	description
Videolecture_Id	A unique identification code for a videolecture
Lecture_Id	An identification code for a lecture
Presentation_Id	An identification code for a course presentation
User_Id	A unique identification code for the teacher
Recording_Time	Date in which the videolecture was recorded

4.15 FORUMs

Table 4.15 contains the attributes related to a forum.

Table 4.15: FORUMs

attribute/feature	description
Forum_Id	A unique identification code of the forum
File_Id	An identification code of the file to which the activity is related (File_Id)
Videolecture_Id	An identification code of the videolecture to which the activity is related (Videolecture_Id)
Lecture_Id	An identification code of the lecture (Lecture_Id) to which the assessment is related
Course_Id	An identification code for a course to which the forum is related
Presentation_Id	An identification code for a course presentation to which the forum is related
OgForum_Id	Identifier of the original sub-forum
Threads_Count	Number of threads in the forum

Title	Title of the forum
Og_Forum_Title	Name of the original (sub)forum
ParentForum_Id	Name of the parent (sub)forum
ParentForum_Title	Identifier of the parent (sub)forum
Forum_Chain	Complete sequence of (sub)forum names from root to current subforum
Depth	Depth number of (sub)forum in forum_chain
TitleTags_Count	Number of tags associated to the subforum title
Users_Count	Number of unique users active in the forum (anonymous users are counted as 1 unit)

4.16 THREADs

Table 4.16 contains the attributes related to a thread, belonging to a forum.

Table 4.16: THREADs

attribute/feature	**description**
Thread_Id	A unique identification code of the thread
Views_Count	Number of views
Forum_Id	Identification code of the forum in which the thread is located

4.17 POSTs

Table 4.17 contains the attributes related to a post, belonging to a thread in a forum.

Table 4.17: POSTs

attribute/feature	**description**
Post_Id	A unique identification code of the post
Post_Time	Time in which the post has been published
NormalizedPost_Time	Normalized time in which the post has been published
Votes_Count	Sum of the votes received by the post
User_Id	A unique identification code for the user

Words_Count	Number of words in the post
Order	Order of the post in the thread
Thread_Id	Identification code of the thread in which the post have been uploaded
ParentPost_Id	parent post identifier

4.18 FILEs

Table 4.18 contains the attributes related to teaching material.

The attributes include file name and format (e.g.: pdf, csv, json, html).

Table 4.18: FILEs

attribute/feature	**description**
File_Id	A unique identification code for a file
Course_Id	An identification code for a course
Lecture_Id	An identification code for a lecture
Presentation_Id	An identification code for a course presentation
User_Id	A unique identification code for the teacher that uploaded the file
Title	Name of the file
Format	Format of the file

4.19 ACTIVITYs

Table 4.19 contains a user's activity related to files, videolectures, lectures, forums, threads, posts, assessments and exercises.

The activities that are recorded include: mouse movements, number of clicks (left, right and mouse wheel click), number of keystrokes, start and end time, idle time.

Table 4.19: ACTIVITYs

attribute/feature	**description**
Activity_Id	A unique identification code for an activity
File_Id	An identification code of the file to which the activity is related (File_Id)

Videolecture_Id	An identification code of the videolecture to which the activity is related (Videolecture_Id)
Forum_Id	An identification code of the forum to which the activity is related (Forum_Id)
Thread_Id	An identification code of the thread to which the activity is related (Thread_Id)
Post_Id	An identification code of the post to which the activity is related (Post_Id)
Lecture_Id	An identification code of the lecture to which the activity is related (Lecture_Id)
Assessment_Id	An identification code of the assessment to which the activity is related (Assessment_Id)
Exercise_Id	An identification code of the exercise to which the activity is related (Exercise_Id)
ActionType	The action types are labeled based on the type of activity or on the title of web pages that are on focus / in the view of the student
User_Id	A unique identification code for the user
Activity_Time	The date of student's interaction with the material
Sum_Click	The number of times a student interacts with the material in that day
Start_Time	It shows the start date and time of a specific activity
End_Time	It shows the end date and time of a specific activity
Idle_Time	It shows the duration of idle time between the start and end time of an activity
Mouse_Wheel	It shows the amount of mouse wheel during an activity
Mouse_Wheel_Click	It shows the number of mouse wheel clicks during an activity
Mouse_Click_Left	It shows the number of mouse left clicks during an activity
Mouse_Click_Right	It shows the number of mouse right clicks during an activity

Mouse_Movement	It shows the number of mouse movements during an activity
Keystroke	It shows the number of keystrokes during an activity
Type	It describes the activity type

Chapter 5

Manual alignment

5.1 Methodology

After the definition of the UNIFORM schema, it is necessary to evaluate its generality and modularity by trying to find a match with the features belonging to the *Test datasets* (4).

Figure 5.1 shows UNIFORM and the Test datasets.

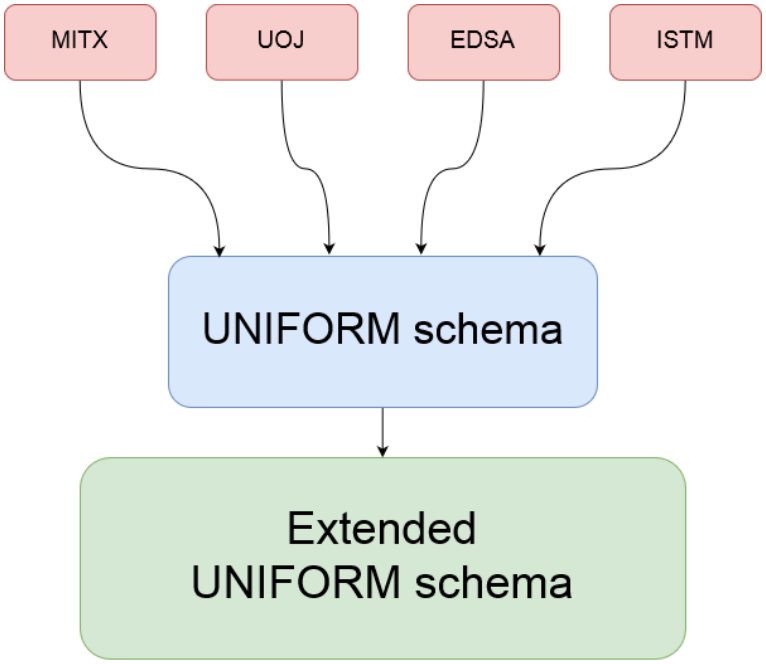

Figure 5.1. Alignment of UNIFORM and Test datasets

To manually align the original datasets with the UNIFORM schema, for each attribute in the source dataset it is necessary to look for an approximated match with UNIFORM. If a match is not found, then a new attribute (and eventually table) is created in the extended dataset version to represent the corresponding information.

- **MITx** dataset attributes match various tables of the UNIFORM schema, in fact its structure is similar to **HarvardX** dataset.

- **ISTM** dataset contains information about students' survey questions: the alignment is performed considering the questions as EXERCISEs belonging to a unique assessment, and surveys as ASSESSMENTs.

- **EDSA** dataset information is related to users' activities, so they match ACTIVITYs table of UNIFORM schema.

- **UoJ** dataset integration can not be done directly: its course presentations are divided in modules, so a further level of granularity is required.

5.2 Extended UNIFORM schema

Thanks to the modularity of the UNIFORM schema, it is possible to easily add new tables and features, if needed. The use of a dynamic schema, that can be modified in order to include datasets with different structure, allows more flexibility. As more datasets are analysed and integrated in UNIFORM, the schema becomes more accurate.

While performing manual alignment, for each attribute in the source dataset it is necessary to check if there is a match with UNIFORM schema. If no match is found, the new attribute is added to the general dataset.

University of Jisc dataset contains information related to student personal data and student career data. The peculiarity of this dataset is that course presentations are divided in modules: so it is required to add this entity.

The integration of **UoJ** dataset is solved by adding new tables to UNIFORM:

- **MODULEs**: it contains the information about modules, related to specific course presentations (a further granularity is added: COURSE -> COURSE-PRESENTATION -> MODULE);
- **MODULEs-PRESENTATIONs**: it contains the information about modules presentations, related to specific modules (a further granularity is added: COURSE -> COURSE-PRESENTATION -> MODULE -> MODULE-PRESENTATION);
- **USERs-MODULEs**: it contains the information about users, related to specific module presentations (the idea is to add a table which is similar to the existing ones: USERs-INSTITUTEs, USERs-COURSEs and USERs-PRESENTATIONs).

Notice that when adding new tables, it may be also necessary to add attributes in the other already existing tables. Table identifiers allow to represent relationships between different tables.

The attribute module presentation identifier (*Module_Instance_Id*) is also added to the tables: **ASSESSMENTs, LECTUREs, FORUMs** and **FILEs**. In fact, those entities may be also referred to module presentations.

Figure 5.2 shows the 22 tables included in Extended UNIFORM schema, which is created adding 3 tables to the original UNIFORM schema.

Due to the generality and modularity of UNIFORM, it is not necessary to drastically change the schema: in this case, a further level of granularity is required to include **UoJ** attributes related to course modules.

The new tables are colored in light-green, while the original UNIFORM tables are colored in light-blue.

In the following sections, the new tables are described:

- MODULEs, see Section 5.2.1
- MODULEs-PRESENTATIONs, see Section 5.2.2
- USERs-MODULEs, see Section 5.2.3

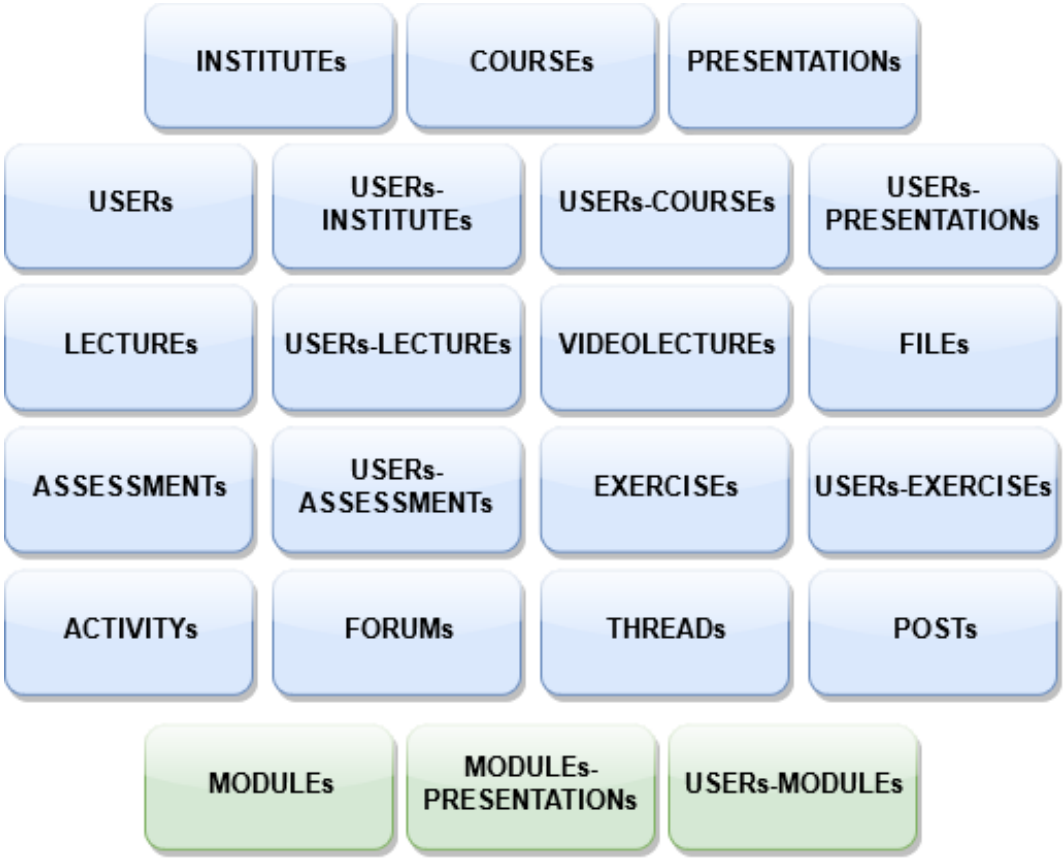

Figure 5.2. Extended UNIFORM tables

5.2.1 MODULEs

A further granularity is added to UNIFORM schema: COURSE -> COURSE-PRESENTATION -> MODULE.

Table 5.1 contains the attributes related to a course module. A module is related to a course presentation. The table attributes include the module name, the related subject and the number of credits.

Table 5.1: MODULEs

attribute/feature	**description**
Module_Id	The unique identifier standard across SRS and LMS for the module.
Presentation_Id	An identification code for a course presentation
Name	The actual name of the module

Subject	Module subject name.
Credits	Number of credits awarded for the module
Level	Level of credit points.

5.2.2 MODULEs-PRESENTATIONs

A further granularity is added to UNIFORM schema: COURSE -> COURSE-PRESENTATION -> MODULE -> MODULE-PRESENTATION

Table 5.2 contains the attributes related to a course module presentation. There are several presentation of a module. The table attributes include the year and the semester of the module presentation.

Table 5.2: MODULEs-PRESENTATIONs

attribute/feature	**description**
Module_Instance_Id	Institution's unique identifier for this module instance
Module_Id	The unique identifier standard across SRS and LMS for the module.
Semester	Period to which module instance relates (e.g. semester 1)
Online	Whether or not this module instance is delivered entirely online.
Year	Academic year that this module runs within.
Optional	Whether or not this is an optional module

5.2.3 USERs-MODULEs

The idea is to add a table which is similar to the existing ones: USERs-INSTITUTEs, USERs-COURSEs and USERs-PRESENTATIONs.

Table 5.3 contains the attributes related to a user for a specific course module presentation. The table attributes include information about start and end time, the grade obtained by a student and the credits.

Table 5.3: USERs-MODULEs

attribute/feature	**description**
User_Id	The institution's own unique identifier of the student.

Module_Instance_Id	Institution's unique identifier for this module_instance
Start_Time	Start date of this module_instance
End_Time	End date of this module_instance
Grade	The grade recorded after any moderation or confirmation processes.
Credits	The number of credits awarded for the module.

Chapter 6

Evaluation

6.1 Objectives

The goal of this analysis is to demonstrate that UNIFORM schema is general enough to allow integrating new datasets related to learning context.

Section 6.2 describes the characteristics of the hardware adopted for the evaluation.

Section 6.3 describes the manual alignment evaluation:

- Subsection 6.3.1 evaluates the generality of the schema, by checking the number of UNIFORM tables matched by each dataset.
- Subsection 6.3.2 gives an overview on the kind of information that is available in the considered datasets with respect to the schema, by checking the number of datasets matching a UNIFORM table.
- Subsection 6.3.3 shows the percentage of UNIFORM features matched per dataset and highlights the heterogeneity of the considered learning datasets.

6.2 Hardware

Characteristics of the PC which has been used:

- Intel(R) Core (TM) i7-4500U CPU @ 1.80GHz 2.40GHz
- RAM: 8.00 GB
- Operating System: Windows 8.1

6.3 Manual alignment evaluation

Some graphics are included in the following subsections, in order to provide an overview on the statistics related to the matching between UNIFORM schema and the considered learning datasets.

6.3.1 UNIFORM tables matched by each dataset

In order to evaluate the generality of the schema, it is interesting to check how many UNIFORM tables are effectively matched by every single dataset. Notice that the match with UNIFORM table is detected when at least a feature of that table is matched by the considered dataset.

Figure 6.1 shows, for each learning dataset that is considered, the number of UNIFORM tables manually matched.

None of the datasets matches all 19 UNIFORM tables: the content of the datasets is heterogeneous, as seen in Section 3.1.

The type of information contained in each dataset may include *Student Personal Data, Student Career Data, Educational Module Data, Students Assessment Data, Educational Resource Access* and *Interaction Activity Data* (Figure 3.1).

10 datasets contain information related to at least 47% of the tables, while one (**EDSA**) is focused on users' activities. This shows that the schema is general enough to contain learning data belonging to different contexts: it has been designed in a modular way, allowing to adapt it dynamically when new datasets are included. New tables can be added to the schema, if necessary.

- **POLITO** and **EPM** datasets match 13 tables;
- **OULAD** dataset matches 11 tables;
- **ISTM** and **COURSERA** datasets match 10 tables;
- **EDSA** dataset matches 5 tables;
- the **other datasets** match 9 tables.

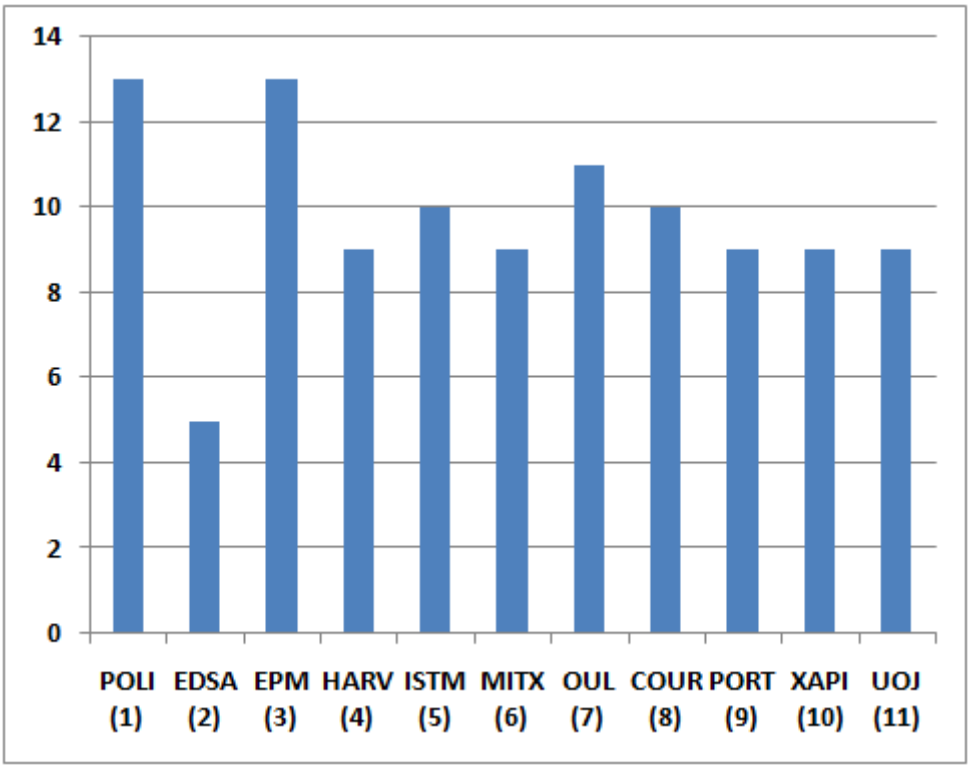

Figure 6.1. Number of UNIFORM tables matching a dataset

6.3.2 Number of datasets matching a UNIFORM table

In order to have an overview on the kind of information that is available in most of the dataset, it is necessary to compute the number of datasets that match at least one feature for the single UNIFORM table.

Figure 6.2 shows, for each UNIFORM table, the number of learning datasets that match them.

It is interesting to notice that:

- Data related to institute, user, course, presentation and assessment are present in most of the datasets.
- Activities, files, lectures and exercises are only included in few datasets.
- Information regarding post, thread, forum, user lecture and videolecture are contained in a single dataset.

– **COURSERA Forums** is the dataset which contains users' posts, threads and forums.

– **POLITO** dataset contains information about videolectures.

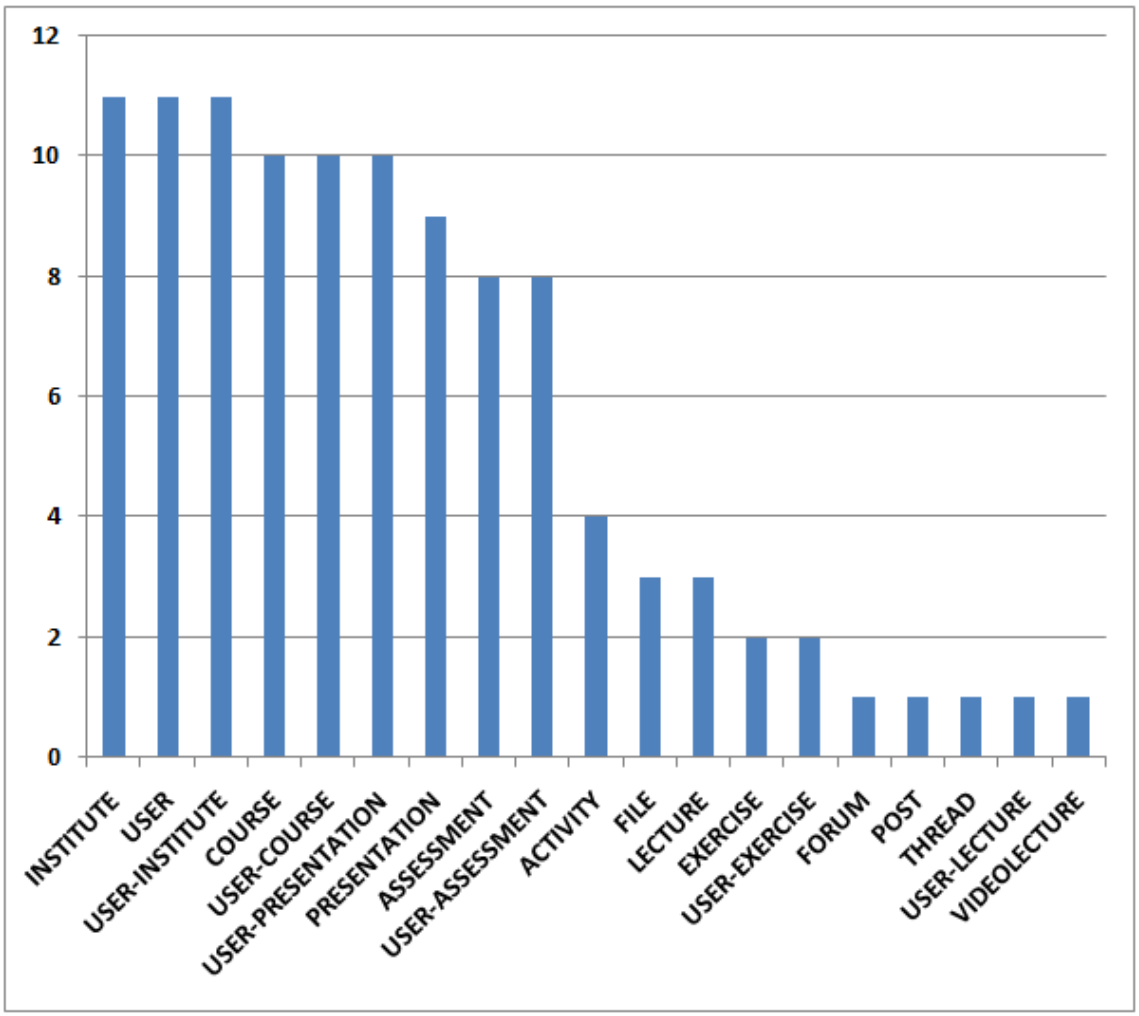

Figure 6.2. Number of datasets matching a UNIFORM table

6.3.3 Percentage of features matched per dataset

The percentage of matched attributes per UNIFORM table for each original dataset is shown in Table 6.1. It is generated by adopting the data from the complete version of Table 6.2, which contains "0s" and "1s": the columns represent the considered datasets, while the rows represent all the features

belonging to UNIFORM tables. The cell corresponding to a certain feature is marked 1 when it has been identified in a specific dataset.

Table 6.1. Percentage of matched attributes per UNIFORM table.

	POLI (1)	EDSA (2)	EPM (3)	HARV (4)	ISTM (5)	MITx (6)	OUL (7)	COUR (8)	PORT (9)	XAPI (10)	UOJ (11)
LECTURE	60.0%	0.0%	40.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	40.0%	0.0%
PRESENTATION	55.6%	0.0%	22.2%	22.2%	0.0%	33.3%	33.3%	66.7%	22.2%	33.3%	55.6%
USER-EXERCISE	0.0%	0.0%	100.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
POST	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
ASSESSMENT	50.0%	0.0%	40.0%	40.0%	30.0%	40.0%	70.0%	0.0%	40.0%	0.0%	60.0%
EXERCISE	0.0%	0.0%	66.7%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
THREAD	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
USER-ASSESSM.	60.0%	0.0%	60.0%	60.0%	40.0%	60.0%	100.0%	0.0%	60.0%	0.0%	60.0%
USER-LECTURE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	75.0%	0.0%
ACTIVITY	21.7%	26.1%	65.2%	0.0%	0.0%	0.0%	21.7%	0.0%	0.0%	0.0%	0.0%
COURSE	80.0%	0.0%	40.0%	40.0%	60.0%	40.0%	40.0%	80.0%	40.0%	40.0%	60.0%
VIDEOLECTURE	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
USER	19.2%	3.8%	3.8%	19.2%	15.4%	19.2%	26.9%	3.8%	73.1%	19.2%	34.6%
FORUM	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	75.0%	0.0%	0.0%	0.0%
USER_INSTITUTE	31.6%	15.8%	10.5%	10.5%	15.8%	10.5%	21.1%	15.8%	42.1%	36.8%	10.5%
INSTITUTE	42.9%	14.3%	14.3%	14.3%	14.3%	14.3%	14.3%	14.3%	28.6%	14.3%	28.6%
USER-COURSE	14.3%	0.0%	14.3%	21.4%	14.3%	14.3%	21.4%	21.4%	21.4%	28.6%	14.3%
FILE	42.9%	28.6%	0.0%	0.0%	0.0%	0.0%	57.1%	0.0%	0.0%	0.0%	0.0%
USER-PRESENT.	19.0%	0.0%	14.3%	52.4%	9.5%	52.4%	19.0%	14.3%	28.6%	33.3%	9.5%

Table 6.1, indicates for each dataset the percentage of matched attributes per UNIFORM table, where the self-explanatory table names indicated in the left hand-side column describe the facet of the related attributes. The results show that UNIFORM integrates most of the original data attributes, but the percentage of matching of each facet is relatively low due to the high heterogeneity of the input data.

e.g.:

- **Portuguese Schools** dataset matches 19 out of 26 features of USERs table, so the percentage is 73.1% in Table 6.1;
- **POLITO** dataset matches 5 out of 26 features of USERs table, so the percentage is 19.2% in Table 6.1;
- **EPM** matches 15 out of 23 features of ACTIVITYs table, so the percentage is 65.2% in Table 6.1.

Notice that 100% in Table 6.1 is usually reached when the UNIFORM table is generated entirely using the attributes of a specific dataset:

- VIDEOLECTUREs for POLITO
- POSTs and THREADs for COURSERA

- USER-EXERCISE for EPM
- USER-ASSESSMENT for OULAD

Table 6.2 contains a partial version of the correspondences table. As previously mentioned, it has been manually written: it contains "0s" and "1s". The columns represent the eleven considered datasets, while the rows represent all the features belonging to UNIFORM tables. There are 194 features, belonging to the 19 UNIFORM tables. A cell is marked with 1 when an attribute of a specific dataset matches the UNIFORM feature.

This table gives a detailed information, with respect to the overall coverage of the UNIFORM tables. Of course, the matrix is sparse: this proves that the kind of data that is provided by the considered data sources is pretty heterogeneous. However, there are several correspondences of features that are included in more than one dataset. This indicates that the schema is general enough to represent data from different datasets, because each UNIFORM feature can be matched by many learning datasets.

Table 6.2. Partial table related to correspondences between UNIFORM features and the considered datasets.

table	feature	COUR	EDSA	EPM	HARV	ISTM	MITX	OUL	POLI	PORT	XAPI	UOJ
USERs	User_Id	1	1	1	1	1	1	1	1	1	1	1
USERs	Birth_Time	0	0	0	1	1	1	1	1	1	0	1
USERs	Gender	0	0	0	1	1	1	1	1	1	1	1
USERs	Disability	0	0	0	0	0	0	1	0	0	0	1
USERs	Birth_Place	0	0	0	0	0	0	0	1	0	1	0
USERs	Residence_Pl	0	0	0	1	0	1	1	1	0	0	1
USERs	ImdBand	0	0	0	0	0	0	1	0	0	0	0
USERs	Education_L	0	0	0	1	0	1	1	0	0	1	0
USERs	Nationality	0	0	0	0	1	0	0	0	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
COURSEs	Course_Id	1	0	1	1	1	1	1	1	1	1	1
COURSEs	Institute_Id	1	0	1	1	1	1	1	1	1	1	1
COURSEs	Name	1	0	0	0	1	0	0	1	0	0	1
COURSEs	Credits	0	0	0	0	0	0	0	1	0	0	0
COURSEs	Typology	1	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
FILEs	File_Id	0	1	0	0	0	0	1	1	0	0	0
FILEs	Course_Id	0	0	0	0	0	0	1	1	0	0	0
FILEs	Present._Id	0	0	0	0	0	0	1	1	0	0	0
FILEs	Format	0	1	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...
ACTIVITYs	Activity_Id	0	1	1	0	0	0	1	1	0	0	0
ACTIVITYs	File_Id	0	1	0	0	0	0	1	1	0	0	0
ACTIVITYs	Videol._Id	0	0	0	0	0	0	0	1	0	0	0
ACTIVITYs	Lecture_Id	0	0	1	0	0	0	0	0	0	0	0
ACTIVITYs	Assess._Id	0	0	1	0	0	0	0	0	0	0	0
ACTIVITYs	Exercise_Id	0	0	1	0	0	0	0	0	0	0	0
ACTIVITYs	ActionType	0	1	1	0	0	0	0	0	0	0	0
ACTIVITYs	User_Id	0	1	1	0	0	0	1	1	0	0	0
ACTIVITYs	Act._Time	0	1	0	0	0	0	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...

6.4 Results

The evaluation demonstrates that UNIFORM is general enough to allow the integration of new datasets related to learning context in the general schema: it has been possible to integrate all the considered datasets, despite their heterogeneity.

The analysis highlights the importance of the modular structure chosen for the schema. UNIFORM schema modularity facilitates the challenge of inserting tables and features that are not yet present in UNIFORM schema. The availability of different kinds of data that can be collected in this common framework provides a huge potential: it can be adopted for several research tasks.

Chapter 7

Conclusions and future work

7.1 Conclusion

The thesis work focuses on the design of a common framework to integrate publicly available learning datasets.

UNIFORM is an integrated open relational database for education, which guarantees modularity and flexibility: its structure can be dynamically edited, to allow including new datasets. This schema contains attributes and tables representing the e-learning data of learning datasets.

It is designed to include learning data from different learning institutions (independently from educational level, national policies, cultural characteristics).

This study demonstrates that it is possible to model a common schema which can be adopted to collect in a unique dataset a variety of e-learning datasets. In fact, UNIFORM is general enough to allow the integration of new datasets related to learning context in the general schema: it has been possible to integrate all the considered datasets, despite their heterogeneity.

The choice of a modular structure facilitates the challenge of inserting tables and features that are not yet present in the schema.

Ideally, integrating in UNIFORM the largest number of datasets from different learning institutions would lead to an Expanded UNIFORM schema which includes all the possible attributes and tables representing learning environment entities.

Collecting different kinds of data in this common framework provides a huge potential: this unique dataset can be adopted for several research tasks related to Learning Analytics.

7.2 Future work

In future, the current study can be extended in the following ways:

- Expand the UNIFORM schema by integrating new tables and attributes: the goal is to make the schema more and more general, so that it finally contains the largest number of attributes related to the learning environment.
- Including multimodal datasets (e.g.: gesture, eye-tracking, biosensors) in the UNIFORM schema would provide a variety of information: it can be analysed to obtain useful insights about students' engagement. In fact, as described in [32], researchers are currently analysing this new kind of data to extract useful knowledge related to the learning process, because it is based on voice, gesture, several biological and mental processes.
- Since the alignment has been limited to English-only: it would be interesting to address the integration of multilingual learning datasets into a unified database model.
- Finally, making the integrated database publicly available to the research community would be useful to support future Learning Analytics projects.

Bibliography

[1] George Siemens and Ryan S. J. d. Baker. Learning analytics and educational data mining: Towards communication and collaboration. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 252–254, New York, NY, USA, 2012. ACM.

[2] Misato Oi, Masanori Yamada, Fumiya Okubo, Atsushi Shimada, and Hiroaki Ogata. Reproducibility of findings from educational big data: A preliminary study. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 536–537. ACM, 2017.

[3] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat. Predicting student performance from lms data: A comparison of 17 blended courses using moodle lms. *IEEE Transactions on Learning Technologies*, 10(1):17–29, Jan 2017.

[4] Jeanette Samuelsen, Weiqin Chen, and Barbara Wasson. Integrating multiple data sources for learning analytics—review of literature. *Research and Practice in Technology Enhanced Learning*, 14(1):11, 2019.

[5] A Cooper and T Hoel. Data sharing requirements and roadmap. *Public Deliverable D*, 7, 2015.

[6] Zhiru Sun, Kui Xie, and Lynley H Anderman. The role of self-regulated learning in students' success in flipped undergraduate math courses. *The Internet and Higher Education*, 36:41–53, 2018.

[7] Daniele Di Mitri, Maren Scheffel, Hendrik Drachsler, Dirk Börner, Stefaan Ternier, and Marcus Specht. Learning pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 188–197, New York, NY, USA, 2017. ACM.

[8] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R

Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[9] C. E. López Guarín, E. L. Guzmán, and F. A. González. A model to predict low academic performance at a specific enrollment using data mining. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 10(3):119–125, Aug 2015.

[10] Michail N Giannakos, Kshitij Sharma, Ilias O Pappas, Vassilis Kostakos, and Eduardo Velloso. Multimodal data as a means to understand the learning experience. *International Journal of Information Management*, 48:108–119, 2019.

[11] Katerina Mangaroska, Boban Vesin, and Michail Giannakos. Cross-platform analytics: A step towards personalization and adaptation in education. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, pages 71–75, New York, NY, USA, 2019. ACM.

[12] Zachary A. Pardos and Kevin Kao. moocrp: An open-source analytics platform. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15, pages 103–110, New York, NY, USA, 2015. ACM.

[13] Jui-Long Hung, Brett E Shelton, Juan Yang, and Xu Du. Improving predictive modeling for at-risk student identification: A multi-stage approach. *IEEE Transactions on Learning Technologies*, 2019.

[14] Han Wan, Kangxu Liu, Qiaoye Yu, and Xiaopeng Gao. Pedagogical intervention practices: Improving learning engagement based on early prediction. *IEEE Transactions on Learning Technologies*, 2019.

[15] Fernando Jimenez, Alessia Paoletti, Gracia Sanchez, and Guido Sciavicco. Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Transactions on Learning Technologies*, 2019.

[16] Alberto Cano and John Leonard. Interpretable multi-view early warning system adapted to underrepresented student populations. *IEEE Transactions on Learning Technologies*, 2019.

[17] David Monllaó Olivé, Du Huynh, Mark Reynolds, Martin Dougiamas, and Damyon Wiese. A quest for a one-size-fits-all neural network: Early prediction of students at risk in online courses. *IEEE Transactions on Learning Technologies*, 2019.

[18] Alvaro Ortigosa, Rosa M Carro, Javier Bravo-Agapito, David Lizcano, Juan J Alcolea, and Oscar Blanco. From lab to production: Lessons

learnt and real-life challenges of an early student-dropout prevention system. *IEEE Transactions on Learning Technologies*, 2019.

[19] Georgios Kostopoulos, Stamatis Karlos, and SB Kotsiantis. Multi-view learning for early prognosis of academic performance: A case study. *IEEE Transactions on Learning Technologies*, 2019.

[20] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. Open university learning analytics dataset. *Scientific data*, 4:170171, 2017.

[21] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. *EUROSIS*, 01 2008.

[22] Jeff KT Tang, Haoran Xie, and Tak-Lam Wong. A big data framework for early identification of dropout students in mooc. In *International Conference on Technology in Education*, pages 127–132. Springer, 2015.

[23] Nikhil Indrashekhar Jha, Ioana Ghergulescu, and Arghir-Nicolae Moldovan. Oulad mooc dropout and result prediction using ensemble, deep learning and regression techniques. In *CSEDU*, 2019.

[24] Devendra Singh Chaplot, Eunhee Rhim, and Jihie Kim. Predicting student attrition in moocs using sentiment analysis and neural networks. In *AIED Workshops*, volume 53, pages 54–57, 2015.

[25] Safwan Shatnawi, Mohamed Medhat Gaber, and Mihaela Cocea. Automatic content related feedback for moocs based on course domain ontology. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 27–35. Springer, 2014.

[26] Martin Hlosta, Zdenek Zdrahal, and Jaroslav Zendulka. Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 6–15. ACM, 2017.

[27] Jiazhen He, James Bailey, Benjamin IP Rubinstein, and Rui Zhang. Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[28] Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. Preprocessing and analyzing educational data set using x-api for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, pages 1–5. IEEE, 2015.

[29] Suhang Jiang, Katerina Schenke, Jacquelynne Sue Eccles, Di Xu, and Mark Warschauer. Cross-national comparison of gender differences in the enrollment in and completion of science, technology, engineering, and mathematics massive open online courses. *PloS one*, 13(9):e0202463, 2018.

[30] Isma Farrah Siddiqui, Qasim Ali Arain, et al. Analyzing students'academic performance through educational data mining. *3C Tecnologia*, 2019.

[31] Mehrnoosh Vahdat, Luca Oneto, Davide Anguita, Mathias Funk, and Matthias Rauterberg. A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. In *Design for teaching and learning in a networked world*, pages 352–366. Springer, 2015.

[32] Agathe Merceron, Paulo Blikstein, and George Siemens. Learning analytics: from big data to meaningful data. *Journal of Learning Analytics*, 2(3):4–8, 2015.