Open University Learning Analytics dataset

This page introduces the anonymised Open University Learning Analytics Dataset (OULAD). It contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). Presentations of courses start in February and October - they are marked by “B” and “J” respectively. The dataset consists of tables connected using unique identifiers. All tables are stored in the csv format.

Data

You can download the latest version of the OULAD here:

Download dataset*

(https://certificates.theodi.org/en/datasets/26648/certificate)

Silver

* You can check integrity of downloaded zip file using the MD5 checksum.

OULAD testimonials

Learning Analytics & Open Data Hackathon 3.0 (https://ctlt.ubc.ca/2018/03/19/learning-analytics-hackathon-3-0-explores-how-data-can-empower-students/) at the University of British Columbia, Canada

The two-day event was held at the University of British Columbia, Canada. Over 100 participants dove into our dataset and experimented with it. Interesting projects in the area of social comparison and visualisation have been developed.

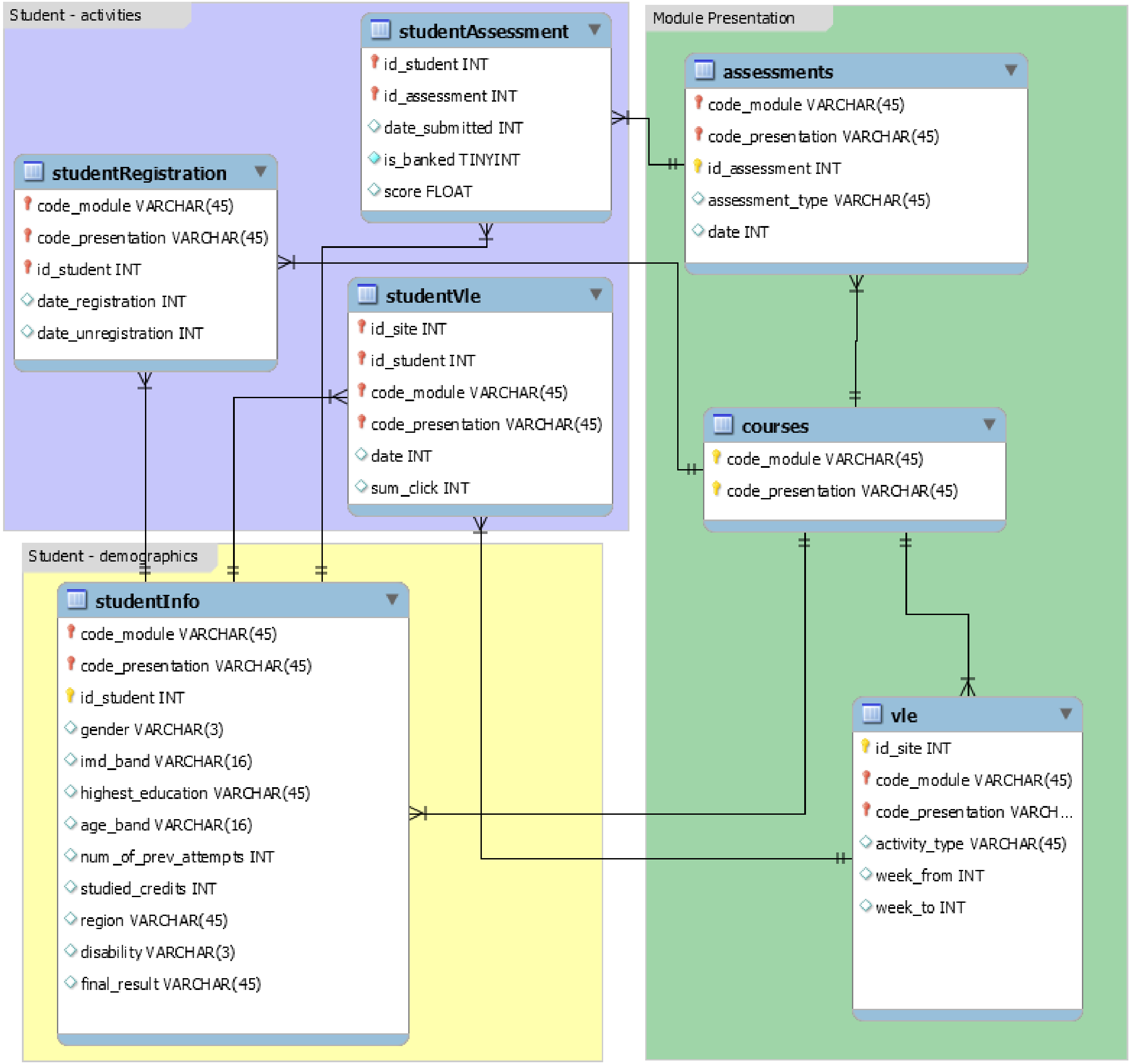
LAK18 Hackathon (https://lakhackathon.wordpress.com/) at Learning Analytics and Knowledge conference (LAK18) in Sydney, Australia

The principal aim of Hack@LAK18 was to enable multi-disciplinary thinking over key open challenges in Learning Analytics based on a problem-oriented, pragmatic approach. OULAD was one of the recommended datasets by the organisers.

More testimonials (/testimonials)

Data description

Database schema



courses.csv

File contains the list of all available modules and their presentations. The columns are:

- code_module – code name of the module, which serves as the identifier.
- code_presentation – code name of the presentation. It consists of the year and “B” for the presentation starting in February and “J” for the presentation starting in October.
- length - length of the module-presentation in days.

The structure of B and J presentations may differ and therefore it is good practice to analyse the B and J presentations separately. Nevertheless, for some presentations the corresponding previous B/J presentation do not exist and therefore the J presentation must be used to inform the B presentation or vice versa. In the dataset this is the case of CCC, EEE and GGG modules.

assessments.csv

This file contains information about assessments in module-presentations. Usually, every presentation has a number of assessments followed by the final exam. CSV contains columns:

- code_module – identification code of the module, to which the assessment belongs.
- code_presentation - identification code of the presentation, to which the assessment belongs.
- id_assessment – identification number of the assessment.
- assessment_type – type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA) and Final Exam (Exam).
- date – information about the final submission date of the assessment calculated as the number of days since the start of the module-presentation. The starting date of the presentation has number 0 (zero).
- weight - weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%; the sum of all other assessments is 100%.

If the information about the final exam date is missing, it is at the end of the last presentation week.

vle.csv

The csv file contains information about the available materials in the VLE. Typically these are html pages, pdf files, etc. Students have access to these materials online and their interactions with the materials are recorded. The vle.csv file contains the following columns:

- id_site – an identification number of the material.
- code_module – an identification code for module.
- code_presentation - the identification code of presentation.
- activity_type – the role associated with the module material.
- week_from – the week from which the material is planned to be used.
- week_to – week until which the material is planned to be used.

studentInfo.csv

This file contains demographic information about the students together with their results. File contains the following columns:

- code_module – an identification code for a module on which the student is registered.
- code_presentation - the identification code of the presentation during which the student is registered on the module.
- id_student – a unique identification number for the student.
- gender – the student's gender.
- region – identifies the geographic region, where the student lived while taking the module-presentation.
- highest_education – highest student education level on entry to the module presentation.
- imd_band – specifies the Index of Multiple Depravation (https://en.wikipedia.org/wiki/Multiple_deprivation_index) band of the place where the student lived during the module-presentation.
- age_band – band of the student's age.
- num_of_prev_attempts – the number times the student has attempted this module.
- studied_credits – the total number of credits for the modules the student is currently studying.
- disability – indicates whether the student has declared a disability.
- final_result – student's final result in the module-presentation.

studentRegistration.csv

This file contains information about the time when the student registered for the module presentation. For students who unregistered the date of unregistration is also recorded. File contains five columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the presentation.
- id_student – a unique identification number for the student.
- date_registration – the date of student's registration on the module presentation, this is the number of days measured relative to the start of the module-presentation (e.g. the negative value -30 means that the student registered to module presentation 30 days before it started).

- date_unregistration – date of student unregistration from the module presentation, this is the number of days measured relative to the start of the module-presentation. Students, who completed the course have this field empty. Students who unregistered have Withdrawal as the value of the final_result column in the studentInfo.csv file.

studentAssessment.csv

This file contains the results of students' assessments. If the student does not submit the assessment, no result is recorded. The final exam submissions is missing, if the result of the assessments is not stored in the system. This file contains the following columns:

- id_assessment – the identification number of the assessment.
- id_student – a unique identification number for the student.
- date_submitted – the date of student submission, measured as the number of days since the start of the module presentation.
- is_banked – a status flag indicating that the assessment result has been transferred from a previous presentation.
- score – the student's score in this assessment. The range is from 0 to 100. The score lower than 40 is interpreted as Fail. The marks are in the range from 0 to 100.

studentVle.csv

The studentVle.csv file contains information about each student's interactions with the materials in the VLE. This file contains the following columns:

- code_module – an identification code for a module.
- code_presentation - the identification code of the module presentation.
- id_student – a unique identification number for the student.
- id_site - an identification number for the VLE material.
- date – the date of student's interaction with the material measured as the number of days since the start of the module-presentation.
- sum_click – the number of times a student interacts with the material in that day.

Examples

Open Data Mashup 2015

Example usage of dataset demonstrated on a small subset of data created for the **Open Data Mashup 2015 (https://elevator.jisc.ac.uk/e/open-data-mashup-challenge/about)** in London:

- **Document describing a Mashup example (resources/documents/mashupExample.pdf)** (in pdf format)
- **Example subset of OULAD (resources/documents/mashupData.RData)** (in .Rdata format)

Rights statement and License

License

This dataset is released under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0/) license.

Citing the dataset

When citing the dataset please use the following reference:

Kuzilek J., Hlosta M., Zdrahal Z. Open University Learning Analytics dataset (https://www.nature.com/articles/sdata2017171) Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017).

Contact

Dataset administrators:

- Jakub Kuzilek ✉ (mailto:Jakub.Kuzilek@open.ac.uk)
- Martin Hlosta ✉ (mailto:Martin.Hlosta@open.ac.uk)
- Zdenek Zdrahal ✉ (mailto:Zdenek.Zdrahal@open.ac.uk)

Knowledge Media Institute ,
The Open University,
Milton Keynes,
MK7 6AA,
United Kingdom.