AFEL

Analytics for Everyday Learning

EC Project 687916

D2.2 - Large-Scale Dataset for (Social) Learning Analytics, Release 1

Deliverable Coordinator: Mathieu d’Aquin (OU)

Contributors: Alessandro Adamou (OU), Besnik Fetahu (LUH)

Reviewers: Ran Yu (LUH), Ilire Hasani-Mavriqi (KNOW)

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

Change Log

Version	Date	Amended by	Comment
0.1	18/11/2017	B. Fetahu	Added ToC
0.2	11/01/2017	M. d'Aquin	Revised skeleton + executive summary
0.3	06/03/2017	A. Adamou	Intro and overviews
0.4	09/03/2017	M. d'Aquin	Further edits and details
0.5	10/03/2017	M. d'Aquin	Integration of anonymisation, dynamic data plan sections and conclusion
1.0	10/03/2017	A. Adamou	General revision for QA submission
1.1	29/03/2017	M. d'Aquin, A. Adamou	Addressing internal reviewers' comments
1.1	31/03/2017	K. Dave	Formatting and finalising report

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

Executive Summary

This document accompanies the initial release of the AFEL dataset for (social) learning analytics. This dataset is a compilation of data resources from various origins that can be used in learning analytics research, especially focusing on online, social and informal learning. It is at the basis of research within the AFEL project, and is released here to support the research community in a similar way. As such, it includes mostly two types of datasets: metadata of online resources used for learning, and traces of learners' activities online in different environments. The dataset is released as a snapshot of those heterogeneous datasets (so as to provide a fixed based for research), integrated under the AFEL data schema and designed following the initial work carried out in WP1 to represent the pivot data vocabulary for the project. We also describe here how the dataset will evolve in future versions to include more learner activity data from sources specific to AFEL, and how we will deal with the privacy aspects of this, as well as the plan to open up the APIs from the AFEL platform (see deliverable D1.2) to also provide the research community with access to dynamic data.

Table of Contents

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

Introduction

The core objective of the AFEL project is to study methods and trajectories for leveraging everyday online activity data of users in order to analyse their learning behaviour. In return, learners will be supported in the exploitation of online resources and learning platforms, by improving their effectiveness (e.g. which activities they choose to perform to progress in their learning) and efficiency (e.g. in the way they use those resources). As most processes targeting this objective are necessarily data-intensive, a collection of structured data relating to online activities is a necessary resource for this study. These data will have to encompass the everyday usage of the Web by potential learners, as well as the materials they come across and the platforms they interact with. Because many of those data are generated from information that users will typically not consciously share with other than the service providers behind each platform, collecting them enables the AFEL Data Platform [AFdA16] as a primary source for user-centric data.

There is therefore much value in such a resource for research within the AFEL project. Likewise, it is not hard to imagine a degree of added value for other researchers in learning analytics and in the broader field of technology-enhanced learning. In this report, we therefore describe ongoing activities related to the release of datasets for (social, online) learning analytics. In particular, this report describes:

1. The first public release of static/snapshot datasets obtained from crawling repositories of resources (such as LRMI, Web of Know-How, DBLP, etc.), as well as from extracting historical data from dynamic sources at a given time point. Here we present the various datasets that have been collected, how they relate to each other, and their general purpose in the framework of the data source classification established in deliverable D1.1 [AFGYL+16].

2. The plans towards future releases of this dataset, with priority on extending it with personal data collected through AFEL data extraction tools. Here we focus on identifying what those data will be, and what approaches should be taken to ensure the right trade-off between releasing data of value to research and protecting the privacy of the learners. Second, we also discuss the opening-up of the AFEL APIs to enable third-party access to integrated, enriched and dynamic data. To this end, we will assume to rely upon the data platform architecture established in WP1 (especially D1.2 [AFdA16]).

First release of the AFEL Static Dataset Release

The AFEL Dataset for Learning Analytics (V1.0) is a collection of data dumps from various sources, integrated under the AFEL Core Data Model (aka. The AFEL data schema)[1]. It is described on its home page[2] in the following way:

> *The AFEL Dataset for Learning Analytics is in itself a collection of datasets that are useful for performing analytics in online/social learning contexts. It is distilled from the content of the AFEL Data Catalogue,[3] excluding datasets containing user-centered data and others that are not freely redistributable.*
>
> *The datasets in this collection can be downloaded individually as dumps in RDF format. This page provides the links to each of the corresponding snapshots as of March 2017. The collection aggregates over 434 million distinct RDF triples, obtained both by refactoring existing linked datasets made available by members of the AFEL project, and by reengineering third-party datasets that were originally not in RDF (e.g. Coursera, OU Analyse, Outline Maps).*
>
> *All dumps are provided as BZipped N-Triples or N-Quads (serialisation format for RDF with or without named graph indications, respectively), except where otherwise noted.*

Table 1 below gives an overview of the datasets included, their origin, size in number of RDF triples and content in terms of the taxonomy of data sources provided in [AFGYL+16]. We follow with a brief description of the AFEL Core Data Model, used to align those heterogeneous datasets, and of each of the datasets individually.

Table 1: Datasets included in the AFEL Dataset for Learning Analytics V1.0.

Dataset	**Source**	**Size**	**Content**
Coursera MOOC Discussion Thread	Metadata from posts and threads in coursera forums related to 60 MOOCs on Coursera in 2013	~5M triples	DT5 - User activity records DT5.5 - Communication
DBLP – Computer Science Bibliography	Metadata of research articles in computer science from the DBLP platform	~165M triples	DT2 - Resource data DT2.1 - Basic metadata DT2.5 - Provenance and authority

[1] AFEL Data schema, http://data.afel-project.eu/catalogue/dataset/afel-core-data-model/
[2] AFEL Dataset for Learning Analytics, V1, http://data.afel-project.eu/catalogue/learning-analytics-dataset-v1/
[3] AFEL Data Catalogue, http://data.afel-project.eu/catalogue/data-catalogue/

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org
© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

LAK Dataset	Metadata and full text content of articles from Learning analytics and Educational Data Mining conferences and journals	~91K triples	DT2 - Resource data DT2.1 - Basic metadata DT2.5 - Provenance and authority
LRMI Resource metadata	Metadata of learning resources described using LRMI on the web	~115M triples	DT2 - Resource data DT2.1 - Basic metadata
Open University courses	Description of courses on offer at the Open University in the UK	~1M triples	DT2 - Resource data DT2.1 - Basic metadata DT2.2 - Indicators of complexity, heterogeneity, controversiality, bias
OU Analyse	Anonymised learner VLE activities and results from the Open University's Learning Analytics platform for 22 courses automatically generated to simulate real courses	~55M triples	DT5 - User activity records DT5.4 - Consumption
Outline Maps (Slepé mapy)	Results of answers to geography (map-based) quizzes from ~91k users of slepemapy.cz	~71M triples	DT5 - User activity records DT5.6 - Gaming
Web of Know How	Resource metadata and human activity from WikiHow and SnapGuide	~23M triples	DT2 - Resource data DT2.1 - Basic metadata DT5 - User activity records DT5.4 - Consumption DT5.5 - Communication

The AFEL Core Data Model

As already described in D1.2 [AFdA16], the AFEL Core Data Model is an ongoing effort to support the integration of different data sources within the project, and provide a guideline regarding the type of data and the attributes of data entities we are expecting to manipulate through different data sources.

The basis of the model is the taxonomy of data sources described in D1.1 [AFGYL+16], whereas the model extends *Schema.org* as a base schema. Indeed, *Schema.org* naturally includes concepts and properties to describe user profiles and resources, and through the

LRMI[4] initiative, already has extensions to tackle learning resources specifically. Some aspects where extensions are required, however, include entities and attributes related to the user's activities, where a small vocabulary of actions exists, but is not always sufficient.

The core data model is currently available in RDF through the AFEL Data Platform's own catalogue at http://data.afel-project.eu/catalogue/dataset/afel-core-data-model/. Figure 1 below shows an overview of the current version of the model as an RDF Schema (for the sake of legibility, a link to a zoomable online version is also provided).

[4] Learning Resource Metadata Initiative, https://www.lrmi.net/

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

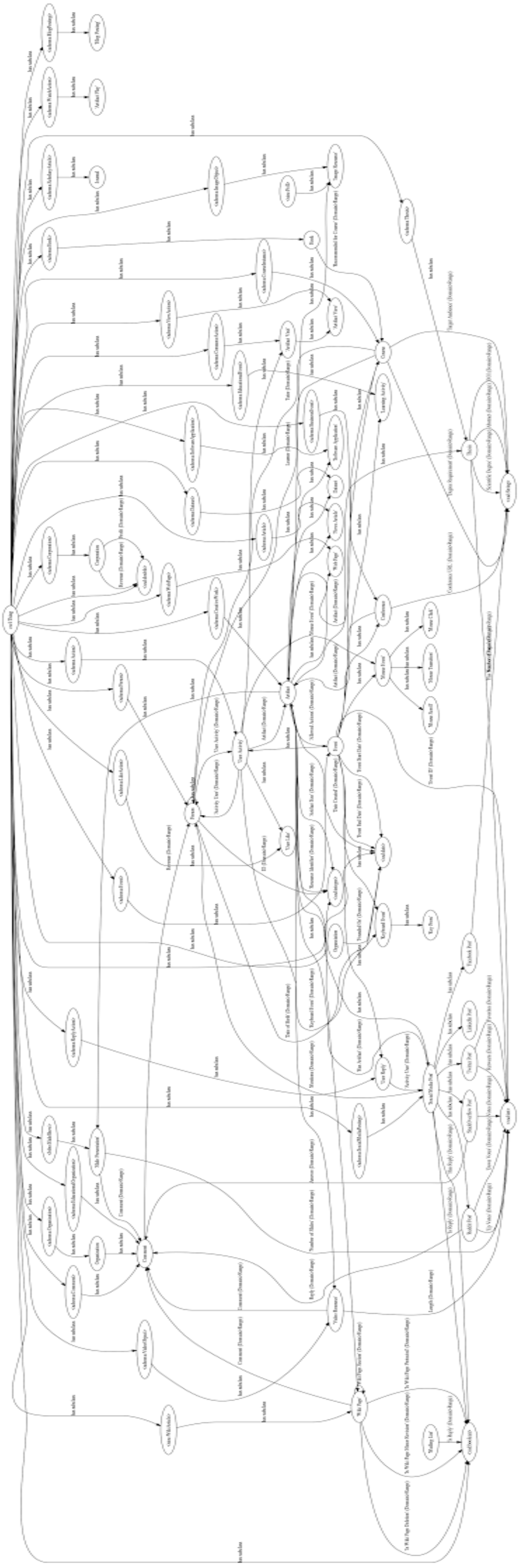

Figure 1: Overview of the AFEL Core Data Model. See http://data.afel-project.eu/catalogue/dataset/afel-core-data-model/ for a zoomable version.

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

Coursera MOOC Discussion Thread

This dataset provides anonymized versions of the discussion threads from the forums of 60 Coursera Massive Open Online Courses (MOOCs), for a total of about 100,000 threads, including about 740,000 posts by approximately 110,000 users [RG14]. The original dataset is available in a dedicated Github repository[5] and has been converted into RDF using the AFEL Core Data Model, including classes such as Course, Forum and User.

DBLP – Computer Science Bibliography

Linked Data export of open bibliographic information on major journals and proceedings in computer science from DBLP, as made available by LUH/L3S.

LAK Dataset

The LAK Dataset makes publicly available machine-readable versions of research sources from the Learning Analytics and Educational Data Mining communities.[6] It includes in particular metadata and some full texts of 697 articles published at the LAK (2011-14) and EDM (2008-14) conferences from 1,213 distinct authors, with over 11k citations present in the articles and available as individual data.

LRMI Resource metadata

This dataset is a collection of online learning resources annotated in accordance with the Learning Resource Metadata Initiative (LRMI) extracted from the Web Data Commons (WDC), respectively the 2013, 2014, 2015 releases of WDC.[7] It catalogues over 228k learning resources (respectively 28,948 from 2013, 80,775 from 2014 and 118,388 from 2015) categorised into 4,145 different low-level learning resource types as per the original classification present in the datasets. For more details about the generated LRMI corpus, refer to deliverable D2.1 [FGYDA17] and the associated paper [DTYBD17].

Open University courses

This dataset includes metadata from data.open.ac.uk[8] about online courses, material and learning opportunities provided by The Open University,[9] including relation between courses and resources, cost of courses, topics, descriptions, etc. Includes 2289 past and present courses.

[5] Coursera forum data repository on GitHub, https://github.com/elleros/courseraforums

[6] Linked Data for Learning Analytics and Educational Data Mining community, http://lak.linkededucation.org/

[7] Web Data Commons, http://webdatacommons.org/

[8] The Open University Linked Data platform, http://data.open.ac.uk

[9] The Open University, http://www.open.ac.uk

This document is part of the AFEL project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687916 - see http://afel-project.org

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

OU Analyse

This dataset provides anonymised Open University Learning Analytics Dataset (OULAD) [KHHZW15].[10] It contains data about courses, students and their interactions with the OU's Virtual Learning Environment (VLE) for seven selected courses held at The Open University. Includes more than 10M interactions from 32K students with 6.3K online resources as well as the 174K results those students obtained at 207 assessments. The original data has been converted into RDF using the AFEL Core Data Model, including classes such as Course, Artifact, Artifact Visit, Event and Person.

Outline Maps (Slepé mapy)

This dataset includes data regarding user interactions with quizzes for adaptive learning of geography as published by Slepé mapy (Outline maps in English).[11] Data are taken from a snapshot as of May 2015, including more than 10m answers to quizzes from 91k users. The original data has been converted into RDF using the AFEL Core Data Model, including classes such as Place, Event and User.

Web of Know-How

This dataset is a re-engineered dataset from WikiHow[12] and SnapGuide[13] into RDF as part of the Web of Know-How project[14]. This dataset includes resource metadata about nearly 260k tutorials and instructions for specialised and everyday-life tasks, as well as records of human activities around those resources. Instruction sets and activities are aligned with the DBpedia category taxonomy[15].

Plan for future data releases

What is described above represents the very first release of the AFEL dataset for analytics, representing a variety of different data at, already in this early stage, a large scale (more than 430 million triples in total). Through packaging those data under a common format (RDF/NTriples) and aligning them through the AFEL Core Data Model, it therefore represents a valuable resource for researchers in learning analytics and technology-enhanced learning in general. However, at this stage of the AFEL project, there are other types of data that we could not yet include, mostly for two reasons: they represent personal data that require to be protected, and they are dynamic data for which the "snapshot-based" packaging of the current dataset release is not suitable. We describe below the plans for integrating such

[10] OUAnalyse, https://analyse.kmi.open.ac.uk/open_dataset
[11] Outline Maps (Slepé mapy), http://slepemapy.cz
[12] WikiHow, http://www.wikihow.com/Main-Page
[13] SnapGuide, https://snapguide.com/
[14] The Web of Know-How, http://homepages.inf.ed.ac.uk/s1054760/prohow/index.htm
[15] DBpedia ontology documentation, http://wiki.dbpedia.org/services-resources/ontology

data in future dataset releases (especially with regard to the upcoming D2.4 “Large-Scale Dataset for (Social) Learning Analytics, Release 2” due in December 2017).

Anonymised learner-generated data from AFEL tools

While many of the datasets included in the dataset release described above were originally produced by AFEL partners and aligned to/refactored according to the AFEL Core Data Model, they do not include at this stage data that were directly generated through the AFEL data capture (WP1) and enrichment (WP2) activities of the project. One of the key reasons for this is that most of these activities have focused on generating data that are user-centric in nature (i.e., that are directly related to a user/learner) and that include personal information. Those include in particular:

- Data from the GNOSS/Didactalia platform, which include visits from users/learners to resources in Didactalia[16]. While very recently deployed on the production platform of Didactalia, this has already collected around one million activity traces and is growing steadily.

- Data from user browsing activities collected through the AFEL browser extension (see https://github.com/afel-project/browsing-history-webext), which is a single extension compatible with the latest versions of *Firefox*, *Google Chrome*, *Opera* and other minor open source Web browsers based on the Gecko and Blink engines. This tool has been developed as part of activities in WP1 and generates, for every user, very large amounts of data related to traces of visits to online resources. Because it requires review and approval from the hosts of application repositories such as the *Mozilla Add-on Marketplace*[17] and the *Chrome Web Store*[18], before being made available to the general public, the tool is yet to see adoption by large numbers of users.

- Data from the social media environments of users, collected in particular through the Facebook[19] and Twitter[20] data extractor developed in WP1. Similarly to the browsing history data, those datasets are currently hampered by the fact that they need to be adequately packaged and registered in online application repositories (in turn requiring approval from the social media sites in question) in order to enable a more widespread adoption by users at this stage.

Considering the nature of the data in those datasets, any public release would require mechanisms to be put in place to ensure that users are not directly or indirectly identifiable, to

[16] Didactalia, https://didactalia.net/

[17] AFEL Activity monitor at the Mozilla Add-on Marketplace (under review), https://addons.mozilla.org/en-GB/firefox/addon/afel-activity-monitor/

[18] AFEL Activity Monitor at the Chrome Web Store (under review), https://chrome.google.com/webstore/detail/afel-activity-monitor/llipmmlocnefdomgmljdfgmlnhaphpoi

[19] AFEL extractor for Facebook activities, source code, https://github.com/afel-project/facebook-activity-to-rdf

[20] AFEL extractor for Twitter feeds, source code, https://github.com/afel-project/twitter-search-to-rdf

comply with both Data Protection regulation, and the terms and conditions of the various systems involved (including GNOSS and the AFEL Data Platform, see Deliverable 6.3 [Tho16]). We describe below the three options that we have identified and that will be investigated as part of WP2 to integrate in the AFEL Data Platform [AFdA16] in order to provide anonymisation services suitable for a privacy-preserving release of those data for research:

- **K-anonymity** [Swe02] relies on the idea that, after de-identification (i.e. after removing direct identifiers), a dataset can be anonymised by guaranteeing that there should not be less than *k* records having the same values for attributes that can indirectly identify a person, *k* being given to reflect the desired probability to identify a person within a group of *k* individuals. A typical example is a dataset that contains, for example, the age and town location of people. The original dataset might not be even 2-anonymous since, for example, there could be only one person being 92 in a small town. In order to anonymise the dataset according to this criterion, the process is to either obfuscate values (e.g. remove the age value for the person being 92) or generalise them (replace the town by a broader region for this record). While k-anonymity is a very common approach to anonymising static datasets, it has obvious limitations for datasets that are dynamic (as discussed in the next section), when only a few people are represented in the dataset (as is the case currently for some of the datasets described above) or when attributes of the data are not easily generalisable (as is the case for activity traces).

- **Differential privacy** [Dwo08] is an approach similar in principle with k-anonymity, but which is designed for dynamic access to data, rather than for static datasets. While the details of differential privacy are complex and out of the scope of this report, the general idea is to introduce noise into the result of queries at the time of executing the query in the dynamic database, in order to minimise the chance of the records being identifiable, while keeping it as accurate as possible. Similarly to k-anonymity, differential privacy is better applied in cases where a large number of users are present in the record of the system and is still practically only applied in cases when a snapshot of the data could not be obtained or would not provide value.

- **Using generated data** [Agr00] is an approach that goes a step further where the data being released is not the original data, but is generated to reflect the characteristics of the original data for the purpose of analysis (a similar approach was applied in [KHHZW15] on the OU learning analytics dataset described above). The main issue with this is that, first, it works better with data for which clear data distributions can be extracted to make the generated data realistic. This also assumes that the data can be characterised through a set of simple distributions, while sub-populations of the dataset might have widely different behaviours, and therefore follow different distributions. Approaches relying on (co-)clustering [DMM03] have been proposed to

© Copyright OU and other members of the EC H2020 AFEL project consortium (grant agreement 687916), 2017

overcome the particular issues, but naturally require a large number of people being reflected in the data.

Our objective for future releases of the AFEL Dataset for Learning Analytics is therefore to apply those techniques on the kind of datasets described above, once those datasets have reached a scale (especially in number of people represented) for those techniques to be applicable.

Third party access to AFEL APIs

In addition to including personal data, the AFEL-created datasets described above do not only contain personal data, but are also highly dynamic. For the purpose of providing a fixed, stable and citable dataset for research, the dataset release currently includes - and will continue to include in the future - such dynamic datasets at a given date. However, in some scenarios, it can be imagined that real-time access to the constantly updating data might be necessary.

As described in D1.2 [AFdA16], the AFEL Data Platform is based on a set of services, applied currently for the purpose of internal development in the project, but that were originally developed to enable third-party access to datasets that are both heterogeneous and diverse in terms of provenance, ownership and policies [DDAM16]. The objective is for the next release of the AFEL Dataset to also use those mechanisms to include, in addition to the static, snapshot datasets, an ability for researchers to subscribe to some of the dynamic datasets, through obtaining a key for authorising queries to the AFEL Data platform's API.

Conclusion

This reports is a brief description of Deliverable D2.2, which is the first release of the AFEL Learning Analytics dataset. This dataset is made of data obtained from several data sources, covering metadata of various resources, user activities, games, assessments, etc. As such, it represents a useful, large scale data resource to support researchers and developers in their Learning Analytics studies/applications especially related to informal, online learning (on MOOCS, social platforms, etc). We also describe how this first release represents an initial step towards much larger future versions, which will also include data generated from ongoing work in the AFEL project, anonymised, real personal data about user activities on the GNOSS and other platforms, as well as ways to access dynamic data as streams generated from the AFEL data extractors.

References

[AFdA16] Adamou, Alessandro, Mathieu d'Aquin, and Besnik Fetahu. *D1.2 - Base data management infrastructure and core data model*. AFEL project deliverable (2016).

[AFGYL+16] Adamou, Alessandro, Besnik Fetahu, Ujwal Gadiraju, Ran Yu, Elisabeth Lex, Matthias Traub, Peter Holtz, Ricardo A. Maturana, Esteban Sota, Susana López-Sola, and Mathieu d'Aquin. *D1.1 - Specification of data to be collected*. AFEL project deliverable (2016).

[Agr00] Agrawal, Rakesh, and Ramakrishnan Srikant. *Privacy-preserving data mining*. ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.

[DDAM16] Daga, Enrico ; d'Aquin, Mathieu ; Adamou, Alessandro and Motta, Enrico (2016). *Addressing exploitability of Smart City data*. 2016 IEEE International Smart Cities Conference (ISC2), pp. 1–6.

[DTYBD17] Dietze, Stefan, Davide Taibi, Ran Yu, Phil Barker and Mathieu d'Aquin (2017), Analysing and Improving embedded Markup of Learning Resources on the Web, in Proceedings of the WWW 2017 Digital Learning Track.

[DMM03] Dhillon, Inderjit S., Subramanyam Mallela, and Dharmendra S. Modha. *Information-theoretic co-clustering*. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[Dwo08] Dwork, Cynthia. *Differential privacy: A survey of results*. International Conference on Theory and Applications of Models of Computation. Springer Berlin Heidelberg, 2008.

[FGYDA17] Fetahu, Besnik, Ujwal Gadiraju, Ran Yu, Stefan Dietze, Alessandro Adamou. *D2.1 - Data Analytics & Entity Linking for Learning Analytics*. AFEL project deliverable (2017).

[KHHZW15] Kuzilek, Jakub, Martin Hlosta, Drahomira Herrmannova, Zdenek Zdrahal, and Annika Wolff. *OU Analyse: Analysing at-risk students at The Open University*. Learning Analytics Review, no. LAK15-1, March 2015, ISSN: 2057-7494.

[RG14] Rossi, Lorenzo A. and Omprakash Gnawali. *Language independent analysis and classification of discussion threads in Coursera MOOC forums*. IEEE International Conference on Information Reuse and Integration (IRI) (2014).

[Swe02] Sweeney, Latanya. *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.

[Tho16] Thomas, Keerthi. *D6.3 - Data management plan*. AFEL project deliverable (2016)