Data Mashups Privacy Preservation for Learning Analytics

Mercedes Rodríguez García [1], Antonio Balderas [2] and Juan Manuel Dodero [2]

[1] *Departamento de Ingeniería en Automática, Electrónica, Arquitectura y Redes de Computadores, Universidad de Cádiz, 11519 Puerto Real, Spain*
[2] *Departamento de Ingeniería Informática, Universidad de Cádiz, 11519 Puerto Real, Spain*

Abstract

The diversity of information sources available to educational institutions makes it necessary to mash up information in order to get the highest performance through learning analytics. Data mashup requires the implementation of data anonymisation methods in order to protect the privacy of the learners who appear in the data partitions. However, the process of anonymising this data mashup can lead to a loss of data utility. This paper presents a protocol for merging data mashups that preserves privacy by k-anonymising the data while preserving its analytical utility.

Keywords

Learning Analytics, Data Mashup, Data privacy, K-anonymity.

1. Introduction

Today, large datasets about students' activity are available to educational institutions from a variety of sources [16]. These datasets collect important data on student performance and learning, but also contain demographic data. To integrate and compile information from all sources, current e-learning environments rely on data mashups, which offer a broader view of the learner through the exploitation of Learning Analytics (LA) [28].

The confidence of the education community is fundamental to the adoption of LA-based tools [13]. Mashing up information with personal content from a variety of sources is not welcome, as this may compromise individuals' privacy. Even if unique identifiers that identify the information are removed, correlation through potentially identifiable attributes (quasi-identifiers) could assist in re-identification of the individual [24]. Therefore, the need for a protocol to anonymise data and guarantee the usefulness of learning data is fundamental.

This paper presents and applies on a dataset of higher education students, a protocol to mashup data and then anonymise it without losing the statistical usefulness of the data [20].

2. Background

Data privacy is one of the biggest challenges in LA research [3]. The solutions that can be found in this context are based on approaches that prevent access to data by people who should not have access to it, either by defining role-based data access [4] or by storing information locally and avoiding cloud solutions [2].

Applying LA is essential if practitioners want a snapshot of their students' learning process, since LA is fundamental to exploit the large amount of information from the learners' work in the different virtual learning environments [17]. Thus, a mashup of the datasets contained in the partitions from different providers has to be performed while guaranteeing the learners' privacy.

[1]earning Analytics Summer Institute Spain (LASI Spain) 2022, June 20–21, 2022, Salamanca, Spain

EMAIL: mercedes.rodriguez@uca.es (A. 1); antonio.balderas@uca.es (A. 2); juanma.dodero@uca.es (A. 3)

The datasets provided can come from either vertical or horizontal partitioning. While in horizontal partitions, different datasets follow the same schema but store different users [15, 21]. In vertical partitions, different datasets store different sets of attributes of the same users (identified by a common attribute) [6]. Vertical partitioning is the typical configuration of datasets used to build next-generation of Virtual Learning Environments (VLEs). Databases to store and query e-learning data can be implemented with different storage techniques, including graph databases [22], e.g. RDF (Resource Description Framework) triple stores and relational databases [1].

Techniques used in previous work on vertically partitioned datasets achieve anonymisation by k-generalising the dataset [15, 21, 19, 9]. Generalisation techniques have the disadvantage that they either require high computational cost to find an optimal generalisation that minimises information loss [18], or they require an ad hoc taxonomic binary tree for each attribute to be anonymised [8]. It would be desirable to incorporate more practical k-anonymisation techniques in vertical data mashups, such as those based on microaggregation.

With respect to LA, the way in which learner data is represented in VLE is critical to the performance of LA methods [26]. One of the main goals of FAIR (Findability, Accessibility, Interoperability, and Reusability) [27] and open data principles is to improve data representation by enriching metadata with multiple attributes. However, intelligent computing techniques such as machine learning have ethical and security issues that may be discordant with compliance with these principles [23]. Hence, when applied to the field of technology-enhanced learning, FAIR and open data principles can be an advantage for the support of human learning, as well as a risk to human privacy.

The application of Privacy-by-Design (PbD) techniques is crucial for LA research and analytics in educational institutions. Given that current VLEs rely on data from cloud-based environments [16, 5], LA requires enhanced Privacy-Preserving Data Publishing (PPDP) methods capable of operating on data mashups, so that privacy constraints do not impose a limitation on LA solutions [10]. This research aims to address the problem that the PPDP solutions used for LA [14, 11] have not taken into account the actual mashup structure of current VLEs. For the sake of privacy-driven learning analytics, PPDP techniques have been limited to k-anonymity, since others as differential privacy have been proven to provide a worse balance between privacy and utility [12]. The limits and misuse of differential privacy regarding data publishing, which is the main purpose of this research, have been confirmed previously [7].

3. Privacy Preserving Data Mashup Protocol

In this section we present a protocol to mashup vertical data partitions from different data providers. The protocol consists of two phases: the setup protocol, and the anonymisation and integration protocol. In the first phase, the mashup coordinator identifies the data providers that could provide the data partitions to be used by the data consumer. While in the second phase, the data providers and the mashup coordinator anonymise and vertically integrate the data partitions to obtain the de-identified dataset.

We assume that the vertical data partitions contain three types of attributes: identifying attributes, quasi-identifying attributes —whose combinations may be identifying if cross-referenced with other sources of information— and confidential attributes.

3.1. Setup Protocol

As shown in figure 1, the mashup coordinator is responsible for initiating the setup protocol as soon as it receives a request from a data consumer. The mashup coordinator’s tasks include the following:

1. Identification of the providers that can contain the information required by the request. Providers publish their data schema, indicating: their identifying attributes, their quasi-identifying and confidential attributes.
2. Construction of the final mashup schema. This schema should include the identifier attribute that will be used for the join of the data partitions, the aggregate quasi-identifiers, the privacy level that will be applied to the aggregate quasi-identifiers and the set of confidential attributes.
3. Designation of the leading provider that will initiate the anonymisation and integration protocol.

This example aims to demonstrate how the setup protocol is implemented. To do this, we assume that the coordinator has received a request for information about the evaluations of a set of students along with their demographic data, and starts the setup protocol.

First, the mashup coordinator identifies potential data providers. In this ex-ample the mashup coordinator will consider two providers.

- Provider 1 (P 1): student demographic data that comes from an LMS database table (figure 2, left side).
- Provider 2 (P 2): a LRS containing the assessments of a set of students in an activity (figure 2, right side).

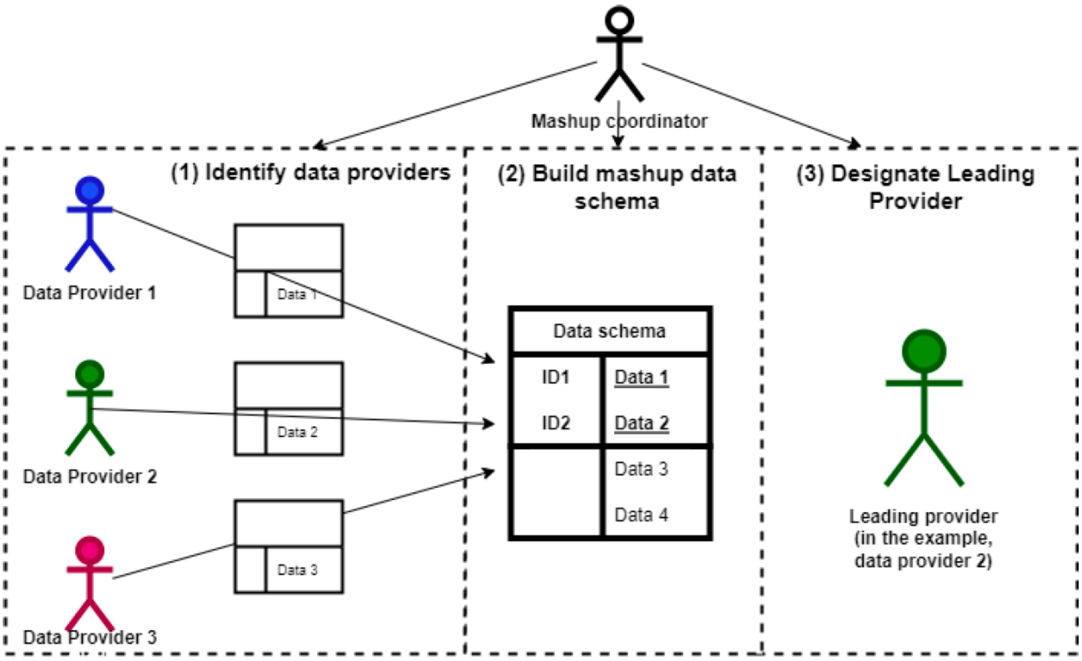
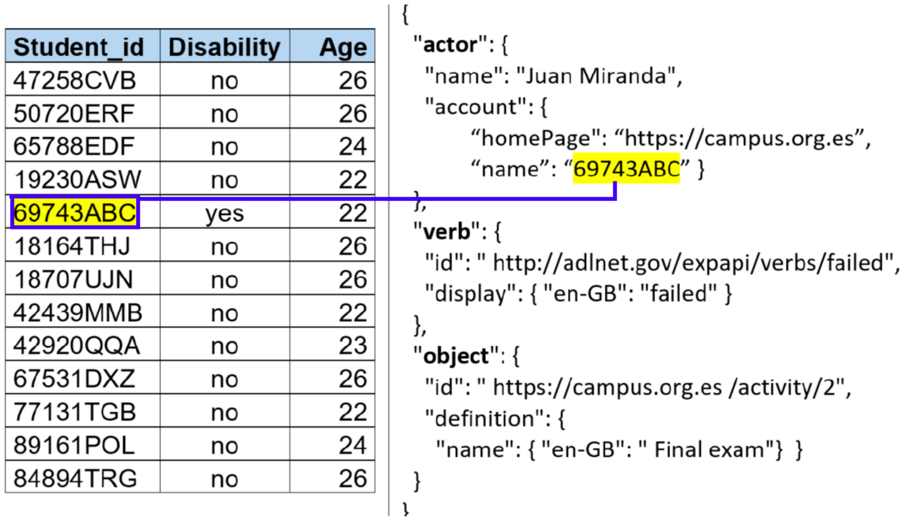

Mashup coordinator
(1) Identify data providers
(2) Build mashup data schema
(3) Designate Leading Provider
Data Provider 1
Data 1
Data Provider 2
Data 2
Data Provider 3
Data 3
Data schema
ID1
Data 1
ID2
Data 2
Data 3
Data 4
Leading provider (in the example, data provider 2)

Figure 1: Setup protocol.



Student_id	Disability	Age
47258CVB	no	26
50720ERF	no	26
65788EDF	no	24
19230ASW	no	22
69743ABC	yes	22
18164THJ	no	26
18707UJN	no	26
42439MMB	no	22
42920QQA	no	23
67531DXZ	no	26
77131TGB	no	22
89161POL	no	24
84894TRG	no	26

```
{
 "actor": {
  "name": "Juan Miranda",
  "account": {
      “homePage": “https://campus.org.es”,
      “name": “69743ABC” }
 },
 "verb": {
  "id": " http://adlnet.gov/expapi/verbs/failed",
  "display": { "en-GB": "failed" }
 },
 "object": {
  "id": " https://campus.org.es /activity/2",
  "definition": {
   "name": { "en-GB": " Final exam"} }
 }
}
```

Figure 2: List of demographic data obtained from the LMS database (left side) and record of a student in xAPI who has failed the activity (right side). Student from the xAPI record is linked to their database record.

Second, the mashup coordinator builds the data mashup scheme. In this ex-ample, the coordinator uses the RDF view strategy described in [25] and defines the mashup name-space to map the linked data attributes of the aforementioned schemes, as the linked data vocabularies, e.g. foaf and schema.org, might not be easily found or mapped to the attributes of the providers.

Each tuple *t* in P1.*demographic* produces the following set of RDF triples:

```
mup:student#t.student_id rdf:type foaf:Person
```

For each tuple *t* in P1.*demographic* and each local QI attribute identifiable as such in P1, generate one RDF tuple. For each local QI attribute, the protocol follow the following strategy:

- If a standard vocabulary exists to represent it, the attribute is mapped. For instance, gender.
- If it does not exist, it is defined directly in the namespace (mup). For in-stance, disability.

```
mup:student#t.student_id schema:gender mup:student#t.gender
mup:student#t.student_id mup:disability mup:student#t.disability
mup:student#t.student_id mup:age mup:student#t.age
```

The mashup coordinator can also use *foaf : age* as a valid mapping instead of using directly *mup : age* adding the following triple:

```
foaf:age owl:sameAS mup:age
```

For each tuple *t* in P1.*demographic* and u in P2.*activity* such that t.student id = u.student id, a triple of the following structure is generated:

```
mup:student#t.student_id mup:failed mup:student#u.activity_id
```

Thirdly, the mashup coordinator chooses the leading provider so that the latter can initiate the integration and anonymisation protocol.

3.2. Anonymisation and Integration Protocol

This protocol carries out the vertical integration of the data partitions identified in the setup protocol and the k-anonymisation of the aggregate quasi-identifier, which is built by vertically joining the quasi-identifier attributes of each partition. Privacy-preserving data collection and integration is achieved by decoupling the collection of quasi-identifiers from the collection of confidential data and by using what are known as privacy-preserving connectors (ppc) [20] —a pseudonym of that identifier attribute shared by all the vertical partitions. The ppc for a given record is computed as a collision-resistant hash function of the value that the identifier attribute holds in the record and a nonce common to all records. The nonce—one-time arbitrary number—is used to prevent reusing the connector and strengthen the connector against dictionary attacks.

Two ppc are used in the protocol: one to integrate the data partitions received in the quasi-identifier collection, named Qppc, and another to integrate the data partitions received in the confidential data collection, named Cppc. This segregated collection of attributes contributes to anonymising data because it allows confidential attributes to be disassociated from quasi-identifiers and, thus, prevents the mashup coordinator from linking the original values of the quasi-identifiers with sensitive information.

The anonymisation and integration protocol is summarised as follows:

1. The leading provider generates the nonces Qnonce and Cnonce used to build the privacy-preserving connectors.
2. The leading provider shares the nonces with the other data providers participating in the process by using a secure channel between communicating parties, such as TLS (Transport Layer Security).
3. Each provider derives the connectors Qppc and Cppc for each of the records in the partition.

4. Each provider sends the quasi-identifier attributes of its partition, along with the corresponding Qppc connectors, to the mashup coordinator via a secure channel.
5. The mashup coordinator vertically integrates the received data partitions through the connector Qppc to build the aggregate quasi-identifier.
6. The mashup coordinator initiates the anonymisation process of the aggregate quasi-identifier. Any PPDP method that satisfies k-anonymity, such as those based on aggregation or generalisation mentioned in Section 2, can be used to anonymise the quasi-identifier attributes.
7. The mashup coordinator sends the anonymised aggregate quasi-identifier to each data provider. Because the anonymisation of the quasi-identifiers has been delegated to the mashup coordinator, the data providers must make sure before reporting confidential information that the result satisfies the requirements of k-anonymity.
8. Each provider integrates the anonymised aggregate quasi-identifier with its confidential data through the connector Qppc.
9. Each provider sends its confidential data, along with the connectors Cppci and the anonymised aggregate quasi-identifier, to the mashup coordinator via a secure channel.
10. The mashup coordinator vertically integrates the received data partitions through the connector Cppci to yield the de-identified dataset provided to the data consumer. This dataset satisfies k-anonymity because at least k records share the same values in the aggregate quasi-identifier.

4. Conclusions

This contribution has shown a new PPVD (Privacy-Preserving Vertical Data) protocol with the following features.

- It serves requests for learning datasets from data consumers.
- Identifies learning data sources, i.e. the different data providers that can satisfy a particular information request.
- Vertically integrates learning data from different educational sources without revealing the learners' identities referenced in the data.
- Finally, it provides the resulting k-anonymised dataset to the data consumer.

The protocol provides an effective integration of learning data and a PbD solu-tion for educational interoperable data architectures, while reconciling LA with privacy. The protocol can be used in any field of application beyond LA systems.

5. Acknowledgments

Supported by the Spanish National Research Agency (AEI), through the project CRÊPES (ref. PID2020-115844RB-I00).

6. References

[1] Ali, W., Yao, B., Saleem, M., Hogan, A., Ngomo, A.C.N.: Survey of RDF stores & SPARQL engines for querying knowledge graphs. TechRXiv (4 2021). https://doi.org/10.36227/techrxiv.14376884.v1

[2] Amo Filva, D., Prinsloo, P., Alier Forment, M., Fonseca Escudero, D., Torres Kom-pen, R., Canaleta Llampallas, X., Herrero Martın, J.: Local technology to enhance data privacy and security in educational technology. International journal of inter-active multimedia and artificial intelligence 7(2), 262–273 (2021)

[3] Berg, A.M., Vrolijk, J., Mol, S.T., Fisher, A.: Anonymisation and synthetic data approaches to minimise privacy risks in highly scaled la intervention driven infras-tructure. In: Companion Proceedings 11th International Conference on Learning Analytics Knowledge (LAK21) (2021)

[4] Cesconetto, J., Augusto Silva, L., Bortoluzzi, F., Navarro-Caceres, M., Zeferino, C., R. Q. Leithardt, V.: PRIPRO-privacy profiles: User profiling management for smart environments. Electronics 9(9) (2020). https://doi.org/10.3390/electronics9091519

[5] Conde, M.A., Hernandez-Garcıa, A.: Data driven education in personal learning environments – what about learning beyond the institution? International Jour-nal of Learning Analytics and Artificial Intelligence for Education 1(1) (2019). https://doi.org/10.3991/ijai.v1i1.11041

[6] Domadiya, N., Rao, U.P.: Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data. Procedia computer science 148, 303–312 (2019). https://doi.org/10.1016/j.procs.2019.01.023

[7] Domingo-Ferrer, J., Sanchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). Communications of the ACM 64(7), 33–35 (2021). https://doi.org/10.1145/3433638

[8] Fung, B., Wang, K., Yu, P.: Top-down specialization for information and privacy preservation. In: 21st International Conference on Data Engineering. pp. 205–216 (2005). https://doi.org/10.1109/ICDE.2005.143

[9] Fung, B.C.M., Trojer, T., Hung, P.C.K., Xiong, L., Al-Hussaeni, K., Dssouli, R.: Service-oriented architecture for high-dimensional private data mashup. IEEE Transactions on Services Computing 5(3), 373–386 (2012). https://doi.org/10.1109/TSC.2011.13

[10] Griffiths, D., Drachsler, H., Kickmeier-Rust, M., Steiner, C., Hoel, T., Greller, W.: Is Privacy a Show-stopper for Learning Analytics? A Review of Current Issues and their Solutions. Learning Analytics Review, 2016, 6, 1–30, ISSN:2057-7494

[11] Gursoy, M.E., Inan, A., Nergiz, M.E., Saygin, Y.: Privacy-preserving learning an-alytics: Challenges and techniques. IEEE Transactions on Learning Technologies 10(1), 68–81 (2017). https://doi.org/10.1109/TLT.2016.2607747

[12] Joksimovic, S., Marshall, R., Rakotoarivelo, T., Ladjal, D., Zhan, C., Pardo, A.: Privacy-Driven Learning Analytics, pp. 1–22. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-86316-6 1

[13] Jones, K.M.: Learning analytics and higher education: a proposed model for estab-lishing informed consent mechanisms to promote student privacy and autonomy. International Journal of Educational Technology in Higher Education 16(1), 1–22 (2019)

[14] Khalil, M., Ebner, M.: De-identification in learning analytics. Journal of Learning Analytics 3(1), 129–138 (4 2016). https://doi.org/10.18608/jla.2016.31.8

[15] Kim, S., Chung, Y.: An anonymization protocol for continuous and dynamic privacy-preserving data collection. Future Generation Computer Systems 93, 1065–1073 (4 2019). https://doi.org/10.1016/j.future.2017.09.009

[16] Ko, C.C., Young, S.S.C.: Explore the next generation of cloud-based e-learning environment. In: Chang, M., Hwang, W.Y., Chen, M.P., M¨uller, W. (eds.) Inter-national Conference on Technologies for E-Learning and Digital Entertainment. Lecture Notes in Computer Science, vol. 6872, pp. 107–114. Springer, Berlin, Hei-delberg (2011). https://doi.org/10.1007/978-3-642-23456-9 20

[17] Martınez-Navarro, A., Moreno-Ger, P.: Comparison of clustering algorithms for learning analytics with educational datasets. International Journal of Interactive Multimedia and Artificial Intelligence 5(2), 9–16 (2018)

[18] Meyerson, A., Williams, R.: On the complexity of optimal k-anonymity. In: Pro-ceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. p. 223–228. PODS '04, Association for Computing Machinery, New York, NY, USA (2004). https://doi.org/10.1145/1055558.1055591

[19] Mohammed, N., Fung, B.C.M., Wang, K., Hung, P.C.K.: Privacy-preserving data mashup. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. p. 228–239. EDBT '09, Association for Computing Machinery, New York, NY, USA (2009)

[20] Rodriguez-Garcia, M., Balderas, A., Dodero, J.M.: Privacy preservation and an-alytical utility of e-learning data mashups in the web of data. Appliec Sciences 11(18) (2021). https://doi.org/10.3390/app11188506

[21] Rodríguez-Garcia M., Cifredo-Cachon, M. A., Quiros-Olozabal, A.: Cooperative privacy-preserving data collection protocol base don delocalized-record chains. IEEE Access 8, 180738-180749 (2020)

[22] Sakr, S., Bonifati, A., Voigt, H., Iosup, A., Ammar, K., Angles, R., Aref, W., Are-nas, M., Besta, M., Boncz, P.A., Daudjee, K., Valle, E.D., Dumbrava, S., Hartig, O., Haslhofer, B., Hegeman, T., Hidders, J., Hose, K., Iamnitchi, A., Kalavri, V., Kapp, H., Martens, W., ¨Ozsu, M.T., Peukert, E., Plantikow, S., Ragab, M., Ri-peanu, M.R., Salihoglu, S., Schulz, C., Selmer, P., Sequeda, J.F., Shinavier, J.: The future is big graphs: A community view on graph processing systems. Communications of the ACM 64(9), 62–71 (2021). https://doi.org/10.1145/3434642

[23] Sheth, A.: Internet of things to smart IoT through semantic, cognitive, and perceptual computing. IEEE Intelligent Systems 31(2), 108–112 (2016). https://doi.org/10.1109/MIS.2016.34

[24] U.S. Department of Education: Family Educational Rights and Privacy Act, 34 CFR 99 (FERPA). Online at https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

[25] Vidal, V.M.P., Casanova, M.A., Cardoso, D.S.: Incremental maintenance of RDF views of relational data. In: Meersman, R., Panetto, H., Dillon, T., Eder, J., Bel-lahsene, Z., Ritter, N., De Leenheer, P., Dou, D. (eds.) On the Move to Meaningful Internet Systems Conference. Lecture Notes in Computer Science, vol. 8185, pp. 572–587. Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41030-7 42

[26] Waheed, H., Hassan, S.U., Aljohani, N.R., Hardman, J., Alelyani, S., Nawaz, R.: Predicting academic performance of students from vle big data using deep learning models. Computers in Human Behavior 104, 106189 (2020). https://doi.org/10.1016/j.chb.2019.106189

[27] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scien-tific data 3(1), 1–9 (2016)

[28] Wolff, A., Moore, J., Zdrahal, Z., Hlosta, M., Kuzilek, J.: Data literacy for learn-ing analytics. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge. pp. 500–501 (2016)