Anonymizing student team data of online collaborative learning in Slack

1st Mario Madureira Fontes
UTAD, Vila Real, Portugal
https://orcid.org/0000-0002-5618-0616

2nd Daniela Pedrosa
Univ. Aveiro, CIDTFF, Aveiro, Portugal
https://orcid.org/0000-0001-9536-4234

3rd Leonel Morgado
INESC TEC, Porto, Portugal
Universidade Aberta, Coimbra, Portugal
https://orcid.org/0000-0001-5517-644X

4th José Cravino
CIDTFF, Aveiro, Portugal
UTAD, Vila Real, Portugal
https://orcid.org/0000-0002-5376-6128

***Abstract*—Research data on the activities of student teams in online learning environments are relevant for evaluating instructional methods, strategies, tools, and materials. For research data sharing and publication purposes, these personal data must be anonymized or pseudonymized as recommended by data protection and privacy policies. This paper addresses issues related to anonymizing and pseudonymizing student data on the Slack teamwork platform, one often employed in educational and business settings. Issues are discussed from two perspectives: data extraction and data transformation. Difficulties and challenges concerning data extraction and transformation are described. The complexities of these two processes are considered, and a starting point for developing more efficient methods is put forward.**

***Index Terms*—education research data, data privacy, data transformation, online education, Slack**

I. Introduction

Data protection has been acquiring more relevance in light of the recently established privacy laws worldwide, such as the European GDPR [1] or the Brazilian LGPD [2]. In educational environments, student data are rich research material [3]. The question is how to conduct research on student data, share results, and still protect personal data of student research subjects. For inclusion in research publications, these student data must be either anonymized or pseudonymized.

The Anonymization process consists in removing any identifiable information concerning personal data. Pseudonymization, on the other hand, replaces real personal information in data records with pseudonyms, limiting personal identification to holders of those pseudonyms' lookup tables. Both procedures comply with data protection and privacy policies.

Our discussion on anonymization and pseudonymization procedures in this paper derives from our experience in extracting data collected in an online collaborative education instruction [4] platform and in transforming this data [5]–[7]. We used the Slack platform, which is an online teamwork collaborative platform that helps participants communicate in a better-organized manner than more traditional tools, such as e-mail. This article addresses two topics: issues related to data extraction and issues related to data transformation. Data transformation is implemented in compliance with data protection and privacy law principles [8].

This work was financially supported by the National Funds through FCT – I.P., and CIDTFF (UIDB/00194/2020) - Universidade de Aveiro, Portugal, – via CEECIND/00986/2017 Individual Support 2017, and via project PTDC/CED-EDG/30040/2017. Our thanks to all collaborators on this research.

II. Data Extraction Issues

Before conducting the data extraction processes, it is necessary to ask students for permission to use their educational data for research purposes, via an informed consent form. This form advises students that their sensitive data will be anonymized or pseudonymized (as appropriate).

This discussion on anonymization and pseudonymization procedures conveys our experience extracting student teams' data about collaborative work activities on the Slack online teamwork platform.

The Slack platform organizes team discussions as channels. Each channel page has to downloaded and saved as HTML files. A channel may contains non-textual information, requiring content extraction into multiple files. This is our recommended procedure to export and retain the text structure generated by the students' interaction because it is saved in the same way as it appears on the Slack platform. Another method to extract the content is using a text JSON format provided by the Slack exporting tool. However, in the JSON format the images, messages, and other information uploaded by the students on the Slack platform are not retained. The channel export tool provided by the Slack website is limited in that it can only be used by the student who created the channels, as administrator of the whole Slack workspace. If the teacher/researcher is not granted administrator access of the entire workspace, he/she cannot export the Slack channel data. After data extraction is performed, its transformation can be implemented and the content exported in HTML can be compared against the original content on the Slack platform, to check that no extraction errors occurred. (This can be done as simply as loading the file into a text editor or word processor, e.g. Microsoft Word or LibreOffice text editor.)

2161-377X/22/$31.00 ©2022 IEEE
DOI 10.1109/ICALT55010.2022.00009

III. Data Transformation Issues

Slack data transformation is a mostly manual process requiring interpretation of data content and dealing with different kinds of media such as text, images, videos, and audio. In general, in-text data transformations are the easiest to implement, by using computational search and replace of strings. It still requires manual checks, due to misspellings of students' names when referring to one another, or due to nicknames and similiar cases. Other data formats such as images, audio, and video are more laborious to verify for personal data, demanding specific approaches for each file format.

During the data transformation process, it is necessary to check for information that must be anonymized or pseudonymized. The choice between anonymization or pseudonymization depends on the level of privacy and on the research purpose. Data targeted for anonymization are usually those related to sensitive student data, such as family names, addresses, screenshots containing student faces, desktop screenshots presenting personal elements (e.g., contents of email messages), personal files, links enabling access to individuals' identities, enabling downloading of student project files, etc. Data targeted for pseudonymization concern less sensitive aspects, which must not be rendered public but are relevant for research purposes. For example, to retain an association between distinct individuals and specific sentences, assignments, and other items, while not disclosing those individuals' identities. A common approach is to employ pseudonyms (hence "pseudonymization") with student names being replaced by labels such as [S1, S2, S3...]. Other kinds of data are approached similarly: for example, teacher names replaced by [T1, T2, T3...].

IV. Discussion

The process of extracting and transforming data from the Slack platform is mostly manual, as metioned above. Technological barriers include choosing the best text editor to replace strings and a method to find and replace information.

The transformation process can require difficult judgement and much attention to which content must be anonymized or pseudonymized. Some contents contain spelling errors or incomplete information patterns. For example, a project named My Great Secret Project may appears as My Grat Seret Project or Great Secret Project or yet Secret Project. Another difficulty derives from semantic and contextualization features of tokens to anonymize. For example, suppose we have a pseudonymization lookup table: [Anne(name) - S1(code)], [Maria - S3(code)] and [Marianne(name) - S7(code)]. We have to prioritize the token Marianne, otherwise it could be replaced by MariS1 or S3nne, not S7. Other prioritizing problems could happen if we added another token such as [Anne Maria(name) - S2(code)].

Another issue concerning the transformation process is when some data has to be aggregated, for instance, if some students do not authorize the use of pseudonymization codes to replace their names but we need to count the name of the student in some statistics, so an identifier must be applied to the students' names, for example, "[aggregated student]".

V. Conclusion

Due to data privacy laws, such as the European GDPR and the Brazilian LGPD, researchers need to find ways to implement anonymization or pseudonymization procedures in research data. This necessity is impacting the software solutions nowadays and forces the developers to do maintenance on a lot of legacy software.

The online teamwork collaborative platforms, such as Slack, could include tools to help the process of anonymization and pseudonymization, such as the students´ names or some other specific research data, during the exportation process. The improvement of data extraction and transformation tools can help researchers to extract and transform research data more effortlessly and simplify the process of anonymization and pseudonymization.

We have provided examples of current shortcomings and pitfalls conducting this process on Slack platform teamwork data.

References

[1] E. Union. "GDPR Lex." en. (), [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL%3A2016%3A119%3ATOC.

[2] R. P. S. Advogados. "Brazilian General Data Protection Law (LGPD, English translation)." en. (), [Online]. Available: https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/.

[3] R. Davies, G. Allen, C. Albrecht, N. Bakir, and N. Ball, "Using educational data mining to identify and analyze student learning strategies in an online flipped classroom," *Education Sciences*, vol. 11, no. 11, 2021, ISSN: 2227-7102. DOI: 10.3390/educsci11110668. [Online]. Available: https://www.mdpi.com/2227-7102/11/11/668.

[4] D. Pedrosa, L. Morgado, J. Cravino, *et al.*, "Challenges Implementing the SimProgramming Approach in Online Software Engineering Education for Promoting Self and Co-regulation of Learning," in *2020 6th International Conference of the Immersive Learning Research Network (iLRN)*, San Luis Obispo, CA, USA: IEEE, Jun. 2020, pp. 236–242, ISBN: 978-1-73489-950-4. DOI: 10.23919/iLRN47897.2020.9155183. [Online]. Available: https://ieeexplore.ieee.org/document/9155183/.

[5] J. Sabin and A. Olive, "Slack: Adopting social-networking platforms for active learning," *PS: Political Science & Politics*, vol. 51, no. 1, pp. 183–189, January 2018, ISSN: 1049-0965, 1537-5935. DOI: 10.1017/S1049096517001913. [Online]. Available: https://doi.org/10.1017/S1049096517001913.

[6] L. Fulton, "Slack in education: A case study of alternative communication for groupwork in graduate level online education," in *Proceedings of Society for Information Technology & Teacher Education International Conference 2018*, E. Langran and J. Borup, Eds., Washington, D.C., United States: Association for the Advancement of Computing in Education (AACE), March 2018, pp. 1458–1463. [Online]. Available: https://www.learntechlib.org/p/182721.

[7] E. Alvarez Vazquez, M. Cortes-Mendez, R. Striker, L. Singelmann, M. Pearson, and E. Swartz, "Lessons learned using slack in engineering education: An innovation-based learning approach," in *2020 ASEE Virtual Annual Conference Content Access Proceedings*, Virtual Online: ASEE Conferences, Jun. 2020, p. 34916. DOI: 10.18260/1-2--34916. [Online]. Available: http://peer.asee.org/34916.

[8] A. Kotsios, M. Magnani, D. Vega, L. Rossi, and I. Shklovski, "An Analysis of the Consequences of the General Data Protection Regulation on Social Network Research," en, *ACM Transactions on Social Computing*, vol. 2, no. 3, pp. 1–22, December 2019, ISSN: 2469-7818, 2469-7826. DOI: 10.1145/3365524. [Online]. Available: https://doi.org/10.1145/3365524.