Towards Open Data in Digital Education Platforms

Joana Soares Machado, Juan Carlos Farah, Denis Gillet
School of Engineering
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
{joana.machado,juancarlos.farah,denis.gillet}@epfl.ch

María Jesús Rodríguez-Triana
School of Digital Technologies
Tallinn University
Tallinn, Estonia
mjrt@tlu.ee

***Abstract*—Despite the traction gained by the open data movement and the rise of big data and learning analytics in education, there is limited support for researchers in education to generate, access, and share experimental data using openly-available digital education platforms. To explore how this gap could be addressed and elicit requirements, we conducted a survey with 40 researchers in the field of technology-enhanced learning, examining their experience and needs handling research data. Drawing on the results of our survey, we devised a set of features that educational platforms should provide to address the identified requirements, enabling researchers in education to run studies within typical learning environments, adhere to legal and ethical frameworks concerning privacy, and share their data confidently with a wider audience. We then categorized these features into five stages that represent the user flow, namely (1) Bootstrapping Research Studies, (2) Ensuring Consent, (3) Gathering Data, (4) Managing Data Sets, and (5) Supporting Open Research and Collaboration. Our aim is to guide forthcoming research and developments to relieve researchers of the burdens of conducting data-sensitive experiments, support the adoption of best practices, and pave the way for open data policies in digital education.**

***Keywords*-open data, data management, data sharing, privacy, education, learning analytics, open research**

I. Introduction

In recent years, there has been a number of initiatives to encourage the adoption of open data policies across research institutions [1], [2], encompassing different subjects such as data science [3], genomics [4] and physics [5]. Bolstered by the adoption of digital technologies and the proliferation of information generated by educational platforms [6], the field of education appears poised to take part in the open data movement. In fact, open data has already been used for evidence-based research in learning analytics (LA) [7]. Nevertheless, there are ethical and privacy concerns associated with handling data in education [6], [8], which may amplify as the volume of and access to information increases. These concerns are mainly addressed by regulations such as the European Union's (EU) General Data Protection Regulation (GDPR), which stipulate ethical and legal requirements for data privacy protection.

In this paper, we propose a set of features to enhance educational platforms that would allow researchers in education to conduct studies more easily, follow best practices when handling sensitive data, and publish their anonymized data sets in the spirit of open data. The paper is structured as follows. Section II provides the motivation behind our approach and highlights related work. In Section III we present our requirements elicitation process, which we conducted as a survey. In Section IV we discuss the results of our survey and the analysis through which we selected key features with which to support researchers. Section V presents these features and how we categorized them into the proposed user flow. We discuss the results in Section VI, drawing the conclusions that drive our future work.

II. Motivation and Related Work

Due to regulations such as the GDPR, there is a need for greater transparency in the way digital education platforms gather data from their users [9]. This requirement extends to researchers, who are often also subject to codes of conduct. Ensuring that research follows both ethical and legal frameworks is a challenge for researchers handling sensitive data [10]. Furthermore, the lack of confidence on whether the protocols followed are appropriate, together with the bureaucracy that ethics and privacy entail, hinder the path towards open data.

The main challenges of open data in education arise due to dispersion, unclear licensing, insufficient standardization of data, lack of incentives and infrastructure for data sharing, as well as ethical and data privacy issues [11]. Recent attempts to address these challenges include a data integration and sharing platform for digital education [12], standard data models for data collection in e-learning [13], as well as techniques for privacy-preserving LA [14].

Furthermore, as teachers have been shown to play an important role in selecting the tools used in their classrooms [15], researchers in education also face the potential challenge of having to adapt to platforms already in use by teachers. However, to the best of our knowledge, there is no educational platform that supports both learning processes in digital settings and transparent data handling policies in a way that could foster open data. This motivates our approach to encourage research and open data in education by focusing on enhancing existing learning technologies.

III. Requirements Elicitation

In order to understand how to enhance learning platforms with research functionalities, we conducted an online survey[1]. As our focus was on learning platforms already being exploited by teachers and students, we identified researchers as the key stakeholders in the requirements elicitation process. We therefore distributed the survey between December 2018 and January 2019 to 40 researchers in technology-enhanced

[1]Online Survey: http://bit.ly/2PcKH4G

learning, mainly from European institutions. Building on the challenges identified in Section II, the survey asked participants about their experience in the following areas: (A) *Usage of Open Data in Research*, (B) *Sharing Research Data*, (C) *Data Management and Sharing Features*, and (D) *Ethics and Data Privacy*. The survey combined multiple-choice and open-ended questions to better understand the rationale behind responses through quantitative and qualitative data.

IV. Survey Results and Analysis

In this section we analyze the results of the survey following the areas listed in Section III. Within each area, we present the requirements that emerged.

A. Usage of Open Data in Research

In our survey, 53% of participants used open data in their research. Specifically, they used open data to explore publicly available data (70%), to complement their own data (52%), to conduct secondary analysis (48%), to create visualizations (48%), and to reproduce other research results (35%). From these results we can infer that tools for data exploration, visualization, and analysis are needed in open data platforms (*Requirement A1*). The aspects of open data that respondents found most problematic were that the data format is not always easy to use (78%), the license for using the data is not always clear (53%), and data authenticity cannot always be ensured (45%). Platforms should thus support interoperable data formats, clear licensing, and data authenticity certificates (*Requirement A2*).

B. Sharing Research Data

Research data was shared by 48% of respondents. Their motivations for sharing included improving the transparency and accountability of their research (80%), increasing the exposure of their work (70%), allowing others to reuse and reinterpret their data (65%), fostering collaborations (60%), getting feedback on their research (60%), and allowing others to reproduce their work (55%). These motivations emphasize the need for transparency and accountability when sharing data (*Requirement B1*), for enabling collaboration in data platforms, and for linking new contributions to open data sets (*Requirement B2*). The reasons highlighted for not sharing research data included ethical and legal constraints (67%), the lack of standards and data infrastructure for data sharing (57%), the cost of preparing data and documentation for sharing (52%), and the lack of training to manage data effectively (38%). Also, 73% of participants agreed that they would be more inclined to share their research data if platforms provided guidelines and tools for data management and sharing. Thus, platforms should integrate features that help researchers handle the sharing process in a more efficient and automated way (*Requirement B3*). Finally, 80% of respondents would be willing to share their research data with colleagues from their research group, with the participants of the study in question (55%), and with trusted peers (50%), while 45% would want to share data openly with the public. Platforms should therefore support not only open data in the broad sense, but also sharing data with different degrees of exposure (*Requirement B4*).

C. Data Management and Sharing Features

On a scale of 1 (not interested) to 5 (very interested), participants were asked to rate the features presented in Fig. 1. Considering ratings 4 and 5, the following requirements emerge. A total of 78% of participants were interested in platforms providing a way to ask users to participate in research studies within the platform itself (*Requirement C1*). A consent management tool was considered useful by 81% of respondents, indicating that data platforms are in a suitable position to help researchers comply with existing ethical and legal regulations (*Requirement C2*). Moreover, 71% were interested in the automated removal of data when consent is withdrawn and 43% in allowing students to disable tracking, showing the need for tools designed to help researchers manage data privacy more effectively *(Requirement C3)*. These results also confirmed *Requirements A1*, *A2*, and *B2*, as participants reported their interest in an open data repository (68%), in tools for interacting with data sets (63%), in tools to import (63%) and export (83%) data in multiple formats, and in certifying data authenticity (58%). Finally, 53% of participants showed interest in tagging data sets with a Digital Object Identifier (DOI) (*Requirement C4*).

D. Ethics and Data Privacy

While 58% of respondents followed an explicit code of conduct in their research, such as their institutional or national codes, 35% were not sure, and 7% did not follow any code of conduct at all. These results reveal a lack of awareness of ethical and legal requirements guiding research practices, which was reinforced by the fact that only 62% of participants had an ethics committee in their institution. These figures support *Requirement C2*. Furthermore, 50% of respondents tracked the consent given by subjects for their research studies, mostly on paper (80%). Only 40% of respondents had strategies or

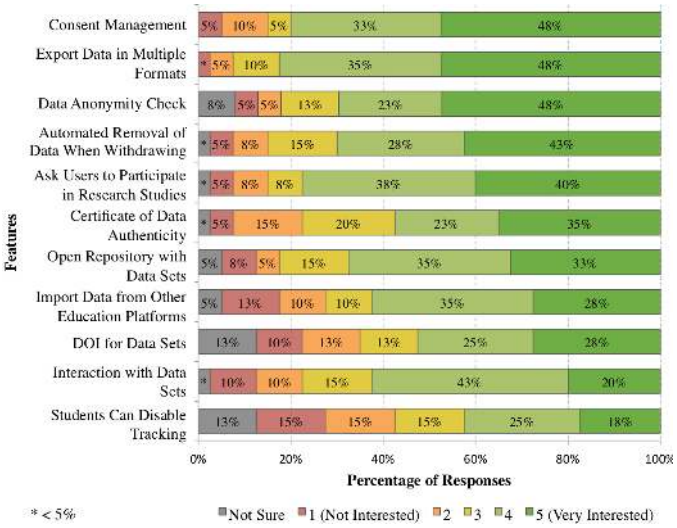

Fig. 1. Perceived interest on data management features (1 = not interested; 5 = very interested).

TABLE I
PROCESSES AND FEATURES TO SUPPORT THE REQUIREMENTS THAT EMERGED FROM THE SURVEY RESULTS.

Process	Requirements	Features
Bootstrapping Research Studies	C1	- Allow researchers to ask teachers to participate in research studies.
Ensuring Consent	C2, C3	- Allow participants to provide consent. - Allow researchers and participants to view signed consent forms. - Allow participants to withdraw consent.
Gathering Data	C1	- Configure the data-related parameters of an experiment. - Collect data generated inside the educational platform. - Provide contextual information.
Managing Data Sets	A2, C2, C3, D1	- Dedicated access for participants to view data collected about them. - Automatic removal of data from participants who withdraw consent. - Compliance with data privacy protection requirements. - Store data in a custom location. - Verification of the authenticity of the data generated in the platform.
Supporting Open Research and Collaboration	A1, A2, B1, B2, B3, B4, C4	- Repository to expose data sets and associated resources. - Different levels of exposure and granularity to share data sets. - Citable identifier for data sets. - Data export in multiple formats. - Data import from external repositories and new contributions. - Clear specification of rights and terms of use of the data set. - Interaction with data sets in the repository.

methods in place to handle data privacy-related processes in a reproducible way and 28% did not allow subjects to access information collected about them, confirming the need for *Requirement C3*. Researchers also expressed the need to store experiment data in custom locations (40%), so platforms should provide the option to specify where the data collected is kept (*Requirement D1*).

V. USER FLOW

With a focus on the scientific method's steps of conducting experiments and communicating results, we defined the user flow of a digital education platform supporting experimental research according to the following five processes: (1) *Bootstrapping Research Studies*, (2) *Ensuring Consent*, (3) *Gathering Data*, (4) *Managing Data Sets*, and (5) *Supporting Open Research and Collaboration*. We then mapped the requirements that emerged from the survey to features that would address them within the context of our user flow. Table I highlights this mapping. In supporting a user flow with these features, a digital education platform could relieve researchers of the burdens of conducting data-sensitive experiments, which could be limiting the adoption of open data practices in education. Moreover, these features are meant to empower all stakeholders, as recommended in [8]. Firstly, by creating a direct communication channel between researchers and teachers to increase transparency of how research studies are framed. Secondly, by allowing students to provide consent directly, temporarily disable tracking, and potentially withdraw consent. Finally, external contributors can participate in research in the spirit of the open data movement.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we identified a set of features necessary to encourage open research and foster open data in digital education platforms. The results of our survey showed that around half of respondents used open data and shared their own research data with others. Nevertheless, the majority of respondents had a number of concerns regarding data sharing, which were mostly linked to ethical and legal requirements. Our findings suggest that researchers in education are willing to participate in the open data movement, but require support tools to manage and share their research data.

Although our analysis is constrained by the sample size of our survey, we aim to build on our work by implementing an architecture that supports the aforementioned features, and conducting a usability study. This will allow us to evaluate which features are more valuable for researchers, validate the user interface, and receive direct feedback from stakeholders.

ACKNOWLEDGMENT

This research has been partially funded by the European Union (grant agreement nos. 731685 and 669074).

REFERENCES

[1] A. Burton, D. Groenewegen, C. Love, A. Treloar, and R. Wilkinson, "Making research data available in Australia," *IEEE IS*, 2012.
[2] "EPFL Open Science Fund," https://research-office.epfl.ch/epflopensciencefund/, accessed: Jan 2019.
[3] A. C. Bart, J. Tibau, E. Tilevich, C. A. Shaffer, and D. Kafura, "Blockpy: An open access data-science environment for introductory programmers," *Computer*, 2017.
[4] "Human Genome Project," https://web.ornl.gov/sci/techresources/Human_Genome, accessed: Jan 2019.
[5] "CERN Open Data Portal," http://opendata.cern.ch, accessed: Jan 2019.
[6] A. Pardo and G. Siemens, "Ethical and privacy principles for learning analytics," *BJET*, 2014.
[7] S. Kellogg and A. Edelmann, "Massively open online course for educators (mooc-ed) network dataset," *BJET*, 2015.
[8] N. Sclater, "Developing a code of practice for learning analytics," *JLA*, 2016.
[9] J. C. Farah, A. Vozniuk, M. J. Rodríguez-Triana, and D. Gillet, "A teacher survey on educational data management practices: Tracking and storage of activity traces," *EP4LA Workshop @ EC-TEL*, 2017.
[10] J. P. Daries *et al.*, "Privacy, Anonymity, and Big Data in the Social Sciences," *Communications of the ACM*, 2014.
[11] S. Dietze, G. Siemens, D. Taibi, and H. Drachsler, "Datasets for learning analytics," *JLA*, 2016.
[12] J. Nicholson and I. Tasker, "Dataexchange: Privacy by design for data sharing in education," in *IEEE FADS*, 2017.
[13] A. del Blanco, A. Serrano, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "E-learning standards and learning analytics. Can data collection be improved by using standard data models?" in *IEEE EDUCON*, 2013.
[14] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-preserving learning analytics: challenges and techniques," *TLT*, 2017.
[15] M. J. Rodríguez-Triana, A. Martínez-Monés, and S. Villagrá-Sobrino, "Learning analytics in small-scale teacher-led innovations: ethical and data privacy issues," *JLA*, 2016.