An End-to-End Data Pipeline for Managing Learning Analytics

Juan Carlos Farah*, Joana Soares Machado*, Pedro Torres da Cunha*, Sandy Ingram†, and Denis Gillet*
*École Polytechnique Fédérale de Lausanne, Switzerland
Email: {juancarlos.farah,joana.machado,pedro.torresdacunha,denis.gillet}@epfl.ch
†College of Engineering and Architecture of Fribourg, Switzerland
Email: sandy.ingram@hefr.ch

***Abstract*—Despite the importance of learning analytics in digital education, there is limited support for researchers in education to generate, access, and share experimental data while complying with ethical and privacy legislation. We propose a set of related tools that support researchers with these tasks and present a blueprint for how these tools can be integrated with existing platforms, enabling researchers to run studies within learning environments, adhere to legal and ethical privacy frameworks, and share their anonymous or anonymized data with a wider audience. We demonstrate the integration of these features into an existing online learning platform.**

***Index Terms*—open data, data management, data sharing, privacy, education, learning analytics, open research, anonymization**

I. Introduction

Digital education platforms often record learner interactions within online lessons and activities, providing educators and researchers with detailed learning analytics (LA) [1]. LA offer significant potential to improve educational experiences. For example, LA can support content personalization, the provision of real-time feedback, and the early detection of at-risk learners [2]. Although work has been done to promote data sharing in education and make education-focused open data repositories publicly available, ethical concerns and recent privacy legislation pose a challenge for the wider adoption of open data practices in education [3]. Furthermore, issues of data interoperability and contextualization make it particularly hard to support cross-platform LA [4].

In this paper, we propose a comprehensive LA data pipeline aimed at supporting open research in education and present the implementation of this pipeline in a digital education platform. The goal is to outline the pipeline's architecture and the practical considerations encountered when incorporating the pipeline into a web-based e-learning environment, thereby demonstrating the technical feasibility of developing and deploying this architecture. Given that flexibility is often required for such implementations, we focus both on how our pipeline's components can be used as part of an integrated system and individually as standalone units. Building on a previous requirements elicitation process [5], we begin by reviewing related work and motivating the need for such a pipeline. We then introduce the design considerations and architecture of our data pipeline in Section III and Section IV, respectively. Section V presents a discussion of our implementation. We conclude and discuss limitations in Section VI, after which we highlight future work in Section VII.

II. Background and Related Work

In this section, we provide a brief overview of the topics underpinning our work and outline how they have guided the design and implementation of our pipeline.

A. Data Formats and Interoperability

Standards for data models and data collection in e-learning are still lacking [6], [7] and pose a barrier to the scaling and interoperability of LA. Our proposed pipeline is optimized to work with LA data conforming to the Experience API (xAPI) standard [8] in the JSON format, but is also adaptable to work with LA specifications and formats from other platforms.

B. Anonymization

Although anonymization is a common strategy to protect data privacy, there are some obstacles that hinder its adoption in practice [9]. First, there is an inherent tradeoff between attaining privacy and preserving the utility of an anonymized dataset, such that the granularity of anonymization can hinder the envisioned data analysis. For example, a well-documented limitation of the *k-anonymization* algorithm [10] is that—in processing a dataset to ensure that individual data points are no longer distinguishable—it removes useful information from the dataset [11]. Anonymization algorithms based on encryption also have well-documented limitations, such as the vulnerability to attacks whereby bad actors potentially reverse-engineer a hash function using brute force algorithms against hashed data attributes that have common non-hashed values [12]. Second, anonymization approaches depend on the data being anonymized and are not easily generalizable. For example, techniques to anonymize geolocation data will differ from those for anonymizing usernames, and two platforms may generate different formats for the same data attribute. Third, automatically identifying sensitive attributes in a dataset is a complex problem, meaning that researchers need to be familiar with a dataset and its schema beforehand [13]. Several techniques have been proposed to address these challenges. Privacy-preserving LA tools often include simple methods such as suppressing or hashing identifiers [9], [14],

[15]. Other robust techniques include the aforementioned *k-anonymity* [10], which ensures that the quasi-identifiers map to at least k users, and *l-diversity* [16], which guarantees that the sensitive values are diversified enough for each set of users with the same quasi-identifiers. Finally, *t-closeness* [17] adds the property that the distribution of sensitive values has to be similar for each of these sets. Although previously proposed LA data management solutions include different anonymization strategies in their design [13], [18], there is no approach that generalizes to the diversity of formats and data types present on education platforms. We address this gap by proposing a desktop tool compatible with different data types and algorithms for anonymization. Furthermore, this desktop tool allows end-users to import custom data schemas and algorithms, making it extendable to as of yet unseen learning scenarios.

C. Open Data Repositories

The lack of incentives and infrastructure for data sharing poses a challenge to the adoption of open data practices. Moreover, existing data-sharing platforms for digital education are not integrated into a full data life-cycle [19], making dataset sharing cumbersome. In our contribution, we propose a pipeline architecture for online learning platforms that integrates sharing functionality alongside data visualization and anonymization tools, helping promote open data practices.

III. Design Considerations

We build on our review of related work and on a previous requirements elicitation process to guide the design considerations for our pipeline. Outcomes from this elicitation process showed that researchers in education were most interested in 19 features across five different processes related to LA data collection, management, and dissemination [5]. These features were to be supported by a toolkit comprising tools for (i) data exploration, visualization, and analysis, (ii) data management and sharing, (iii) managing data privacy, (iv) interacting with datasets, (v) importing and exporting data in multiple formats, and (vi) certifying data authenticity. In this paper, we focus on the architecture and implementation of these tools within a comprehensive LA pipeline.

To place these tools in context, we center our pipeline on a digital education platform on which instructors can create learning spaces containing multimedia content and interactive applications. The platform's learning spaces and applications capture detailed LA data, which is potentially valuable for both instructors—for example, to assess engagement within their learning spaces—and researchers, who can use it for pedagogical research questions. Our goal is to provide instructors and researchers with a holistic pipeline to work with these data from the moment of their capture to the moment they are being openly shared with the pedagogical and educational research community. The pipeline design was therefore driven by three primary considerations:

1) **An integrated, end-to-end approach.** Although many digital platforms in education and beyond provide the functionality to store, visualize, and otherwise work with data, one of our contributions was to combine these functionalities into a unified and cohesive architecture for managing data over their life-cycle. With our tools, users can capture and visualize learning space data, download complete datasets for offline analysis, and anonymize datasets to make them shareable. However—although designed as part of an integrated process—users can also choose to use whichever subset of these pipeline components and tools is appropriate for their workflows and use cases.
2) **A versatile technical toolkit.** The platform's audience includes instructors and researchers with different backgrounds and levels of technical experience. By design, we built our tools to be accessible and useful for such a broad audience. For example, while the anonymization algorithms in our desktop application can be executed in a user-friendly GUI, more technical users can customize these algorithms with their own code.
3) **Transparency and openness.** Given the sensitivity of working with private data, we were keen to provide users with full transparency in every step of our pipeline. For example, data is collected only after user consent, and users are able to download and view all collected data directly through a web dashboard, without intermediary requests. Simultaneously, we were keen to promote and encourage open data practices and principles, and therefore to provide tools that encourage data sharing.

These three considerations guide our architecture design and implementation, which we discuss in the following section.

IV. Architecture

Our system architecture was conceived of and designed as an integrated end-to-end pipeline to capture, visualize, anonymize, and share the data generated by educational platforms. The blueprint outlined in Fig. 1 summarizes our architecture and shows how data flows through the different stages of the pipeline, satisfying the design considerations presented in Section III. In this section, we demonstrate how we implemented and integrated these core features into Graasp. Graasp is an online learning platform on which instructors create learning spaces containing multimedia content and interactive applications [20]. These learning spaces can then be shared with students using a link, allowing them to interact with the content prepared by the instructor. If the instructor has enabled LA, students' interactions with the learning space will generate learning traces. Given that Graasp has been active since 2006 and currently has over 400,000 users, it serves as an appropriate educational platform on which to scaffold a first implementation of our proposed architecture.

Each component of the system architecture has a standalone purpose, and can therefore function as a discrete unit. As an example, certain users may forego the components described in Sections IV-A to IV-E entirely and use only the data anonymization tool described in Section IV-F. However, combining the components into an integrated process offers a more

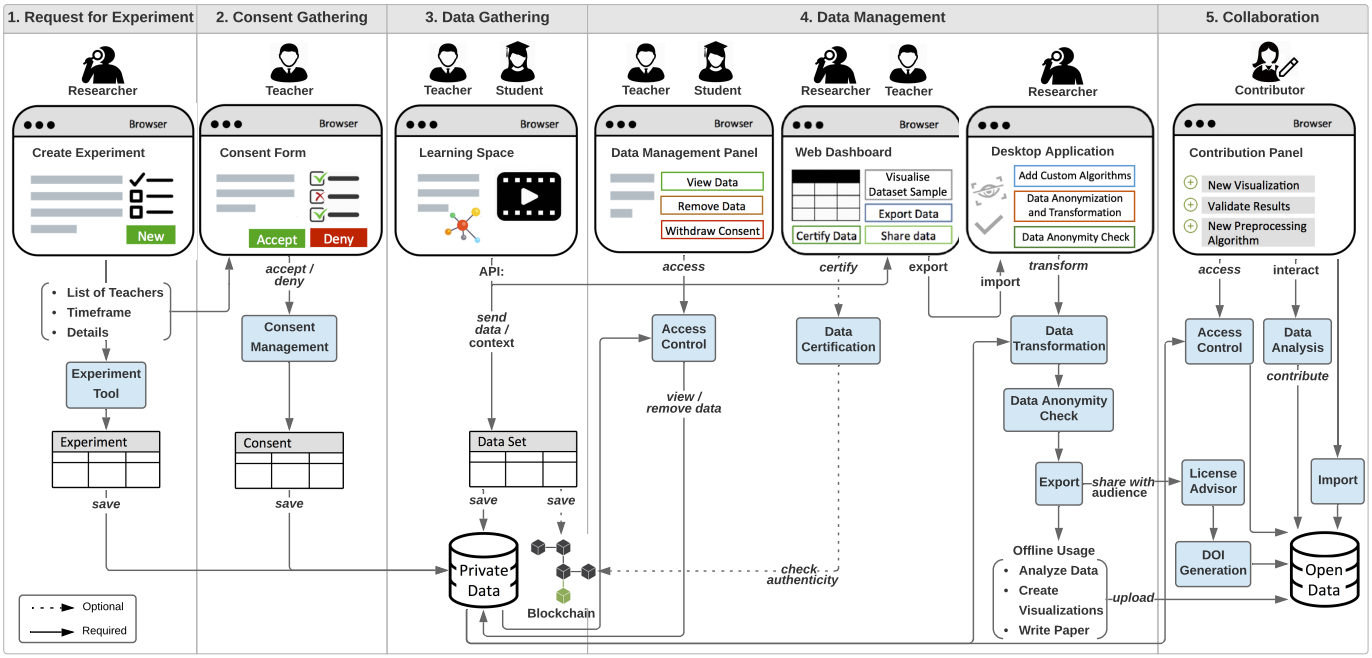

Fig. 1. The user flow in our architecture involves five main stages. First, there is a request for experiment, followed by consent gathering, data gathering, data management, and finally collaboration. The tools that we propose support stakeholders across all stages and are highlighted in blue.

comprehensive set of solutions for working with LA data. Indeed, although many digital platforms incorporate some of these functionalities, one of the innovations in our approach is combining them into a unified and cohesive architecture for managing data over their life-cycle.

A. Request for Experiment

The *Experiment and Consent Management Tool* establishes a connection between researchers and instructors. Using the tool, a researcher invites instructors to participate in their experiment, specifying parameters such as the experiment's purpose and duration, the data collected, and privacy and legal disclosures (Fig. 2). The tool allows the researcher to distribute this invitation to multiple instructors and makes the bootstrapping of a research experiment quick and efficient.

B. Consent Gathering

Having accepted the invitation to the experiment, instructors use the *Experiment and Consent Management Tool* to review the details of the experiment and provide their consent for participation and data collection. The tool timestamps and saves this consent, and therefore acts as a central repository for both researchers and instructors to review consent forms. The tool also allows an alternative, reversed workflow whereby teachers invite researchers to participate in their experiments. In that case, teachers can specify the conditions under which they accept granting access to their data.

C. Data Gathering

Provided consent has been given, the platform's learning spaces and the applications within them capture LA data conforming to the Experience API (xAPI) [8] standard. Predefined

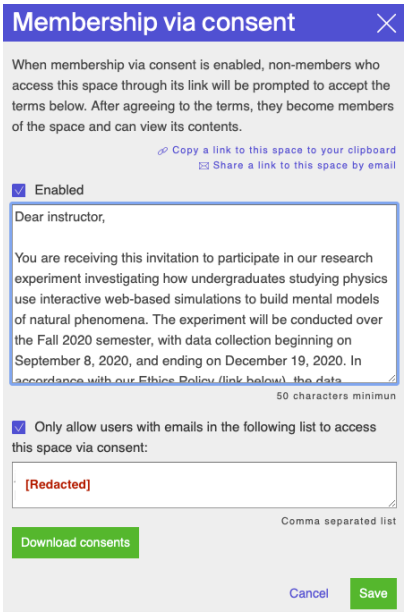

Fig. 2. Template of the *Experiment and Consent Management Tool*.

user interactions with a learning space or application, such as navigation between lessons or the typing of text into an input box, generate data points known as *actions*, which are sent to and stored on the platform's servers. Each action is a JSON object containing information on a discrete user interaction, such as its type, date, time, and content, and on the user performing the interaction, such as their username. Actions

also contain information regarding the context in which they occurred, such as the lesson, learning phase, or activity, as well as an approximate geolocation based on the user's IP address.

D. Data Exposure

To allow for seamless access to the LA data stored on the platform's servers, we built a dedicated RESTful Application Programming Interface (API) to expose data in a structured and well-defined format. The API includes endpoints for retrieving data on a learning space, user accounts, and the actions in a given space. Returned as a JSON object, the data can be analyzed by researchers in a well-supported and recognized file format, in addition to being easily consumed by front-end client interfaces.

E. Data Visualization

Once the LA data was made accessible via an API, we built a web dashboard to visualize it (Fig. 3). The dashboard was designed to provide a high-level overview of activity in a learning space. It contains charts of actions by day, time of day, and type, as well as an interactive map of actions coarsely clustered by geolocation. Additionally, it has functionality allowing the dashboard user to filter displayed actions by one or more users of the learning space.

The dashboard is particularly useful for non-technical learning space owners who would not otherwise be able to manipulate and extract insights from the JSON data returned by the API. However, the browser environment presented two important constraints to the dashboard's design. First, because some learning spaces may have accumulated a large volume of actions, the API call to retrieve these data was often prohibitively time-consuming, resulting in a poor user experience. Second, rendering such large volumes of data on the screen was liable to cause browser and system performance problems and lags. Given these constraints, we limited the number of displayed actions to a maximum of 5,000 randomly sampled actions, sufficient to provide a meaningful general overview of the learning space's usage. Simultaneously, we added an export functionality for users to request and download a space's complete dataset via the dashboard.

F. Local Data Anonymization

A final but integral component of the pipeline is a cross-platform desktop application[1] to transform, filter, and anonymize the datasets downloaded via the web dashboard. The primary motivation for developing this application was to provide researchers with tools to make their datasets more open and shareable, for which the ability to anonymize data is critical. The choice of platform to encompass these data anonymization tools—a desktop application—was an essential architectural decision for this component of the pipeline, as this allows users to anonymize and process their datasets *completely in their local environment*, without needing to share them with third-party services. The application, which we named *Graasp Insights*, works as follows:

[1]The application can be downloaded from https://insights.graasp.org.

1) **Dataset:** The JSON dataset downloaded from the web dashboard is loaded into the application, where it can be viewed alongside summary statistics and visualizations (Fig. 4).
2) **Schema:** A schema describing the dataset's structure is generated via the application's Schemas tab. This allows datasets with this schema to be automatically identified and tagged.
3) **Algorithms:** The application contains a number of pre-composed anonymization algorithms written in Python (Fig. 6). Some of these are optimized for the dataset schema generated by Graasp, but others work on general datasets. A user can review these algorithms and their code in a dedicated section of the application, where they can also add their own custom algorithms and scripts. A further discussion of these anonymization algorithms follows below.
4) **Executions:** Having loaded their dataset(s) into the application and reviewed and added anonymization algorithms, users proceed to an in-app *Executions* tab where they can run selected algorithms against their dataset(s) (Fig. 7). Users can view the generated results in the application (Fig. 8), or export the underlying JSON file to another location on their device.
5) **Pipelines:** In the *Executions* tab, users can only run one anonymization algorithm at a time against a single dataset. Via the application's *Pipelines* tab, users can chain multiple algorithms to be executed sequentially against a dataset, allowing them to efficiently perform various operations (e.g., hash or suppress fields) on their dataset in one quick step.
6) **Validation:** Before making their data available, users can select verification algorithms to be run against the anonymized dataset. These algorithms will check for the presence of sensitive attributes. Examples of such algorithms include *k-anonymity* and *l-diversity* verification, username scan based on regular expressions, among others (Fig. 9). If a test fails, additional anonymization algorithms can be applied, as previously described.
7) **Sharing:** Finally, users can share their anonymized datasets and algorithms on a repository of open educational datasets.

As highlighted above, the application contains pre-composed anonymization algorithms, some of which are optimized for the platform's schema, and others that can be used on general JSON datasets. The diversity of learning platform data and the complexity of JSON datasets, with deeply nested and potentially inconsistent structures, makes composing generalized algorithms a challenging problem and necessitates the composition of platform-specific algorithms. The algorithms built into the application are the following:

- **Data suppression:** With a dataset's schema identified by the application, users can select the fields to be suppressed (i.e., completely removed) from the dataset. For example, users may choose to suppress the *name* field

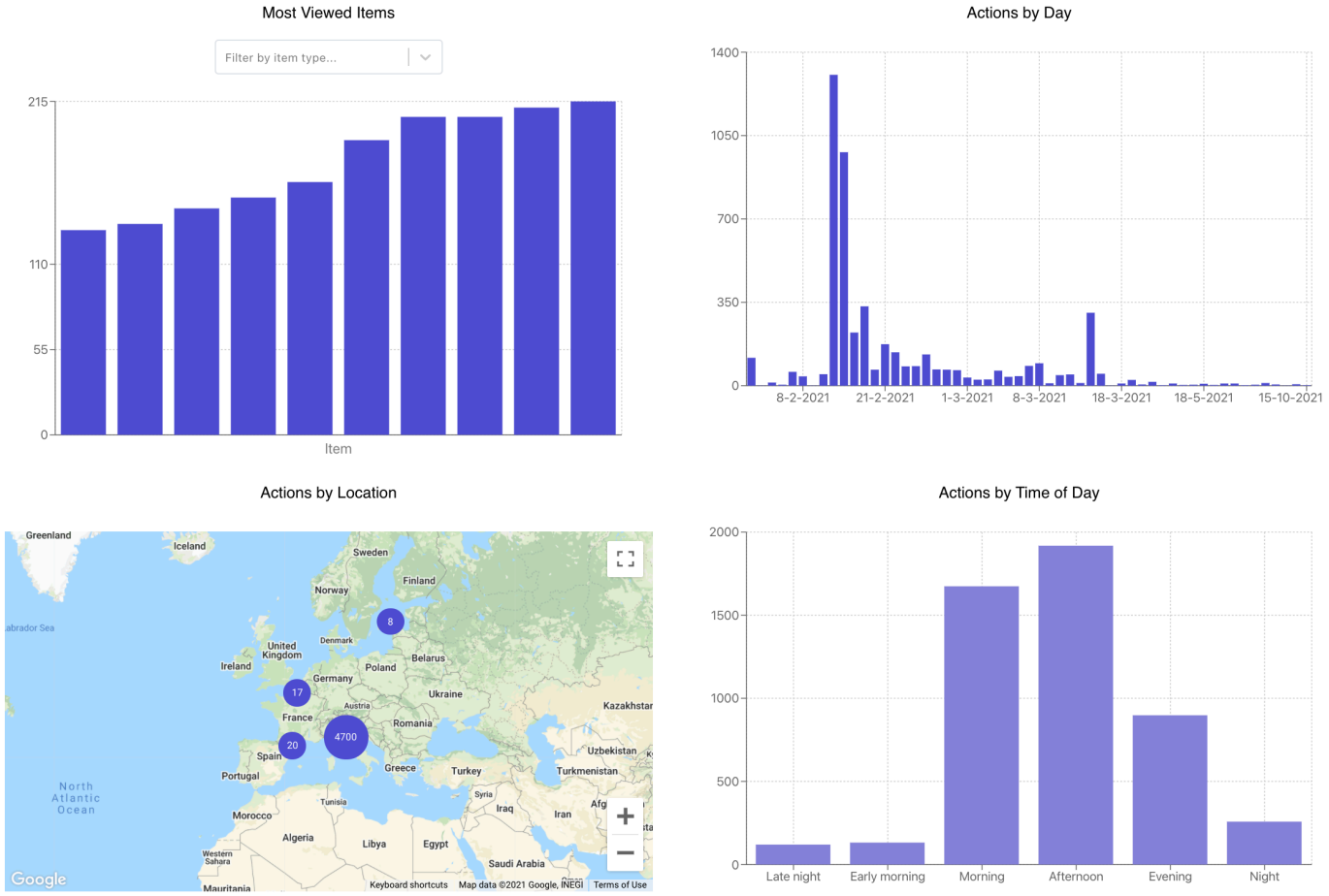

Fig. 3. A web dashboard visualizes the learning analytics data collected by the educational platform, providing high-level insights on learning space usage for technical and non-technical users alike.

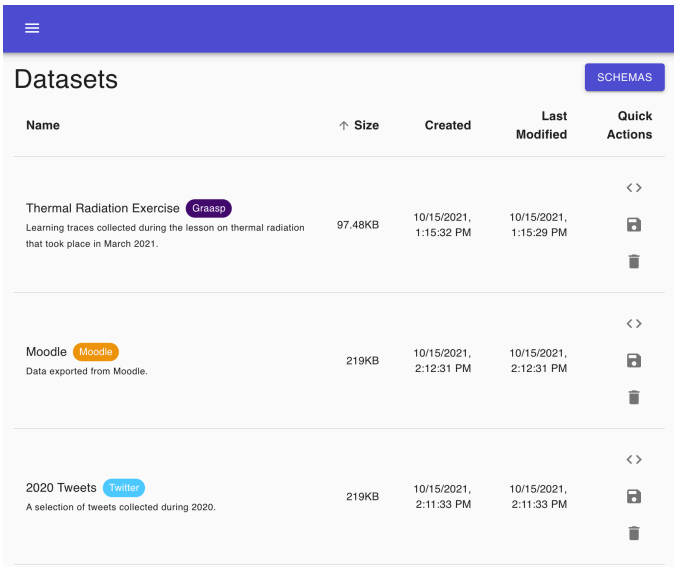

Fig. 4. The *Datasets* tab, where users load in the datasets they want to work with. A schema has been generated for each of these datasets, which are tagged with that schema's label.

from the action data, given its highly sensitive nature.

- **Data hashing:** With the dataset schema as a reference, users can select the fields to which they would like to apply a SHA-256 hash. For example, users may hash the *actionId* field from the action data.
- **Geolocation k-anonymization:** Optimized for the educational platform's schema, this algorithm gradually suppresses the *country*, *region*, and *city* keys within an action's *geolocation* key, until there are at least k users containing each combination, satisfying *k-anonymity* [21] (on the premise that, for the purposes of this algorithm, geolocation is the only identifying attribute).
- **User cleansing:** Optimized for the educational platform's schema, this multistep algorithm (i) replaces *userId*s in a dataset's *actions* key with a SHA-256 hash of each *userId*, (ii) removes identifying information, which may include names and emails, from a dataset's *users* key.
- **Advanced user cleansing:** This algorithm scans the entire dataset to detect remaining instances of identifying user information, such as names and IDs, replacing them with hashed versions thereof. This handles edge cases in which the data contains, for example, identifying

information captured via user-generated inputs. The scan is performed using regular expressions, hence matching close representations of user information (e.g., *John Doe* matches variations like *john.doe@institution.domain*).

Wherever applicable, algorithms were structured to support in-app parameterization, allowing users the freedom to select which data attributes to apply these algorithms to. Within the application, the algorithms are presented alongside plain language descriptions, allowing non-technical users to understand them. Additionally—in order to promote transparency—the code of each algorithm can be accessed within the application. More technical users can review and edit this code. Finally, as shown in Fig. 5, the application allows users to add their own custom algorithms, selected from their file system or written in the in-app editor.

G. Open Data Sharing and Annotation

An *Open Data Repository* was built to share datasets and associated contextual resources, such as metadata from the learning spaces that were part of an experiment. Resources and datasets in the repository can be tagged using a *DOI Generation Tool*, facilitating their citation in subsequent publications, and a *License Advisor Tool* provides recommendations for how to license the data. The *Access Control Tool* allows researchers to grant or deny access to their datasets according to how open they would like them to be. Optionally, the *Data Certification Tool* ensures that even if the data is stored outside the platform, its authenticity is verifiable via a certification mechanism, such as a blockchain-based solution [22].

V. Discussion

The three design considerations outlined in Section III fundamentally shaped the deployed system's architecture. First, the architecture encompassed a complete toolkit for managing experimental data, from seeking appropriate consent to being able to openly share anonymized datasets. Second, every component of the toolkit—as well as the pipeline as a whole—was designed to be accessible to non-technical users, without limiting the toolkit's ability to also provide more technical users with advanced features and customization. Third, transparency was rigorously enforced throughout the pipeline, from ensuring that data was gathered only after appropriate approvals, to making the data available to users without intermediary requests, to openly displaying the anonymization scripts built into the desktop application. The aim is to encourage feedback and audit reports from the users of our tools, promoting an open assessment mechanism whereby our pipeline becomes more robust and at the same time more pertinent to both instructors and researchers.

The architecture was seen to have significant practical implications. First, the requirement elicitation process [5] highlighted the diversity of tools researchers require to manage experimental data, and the devised pipeline provided these tools under one umbrella. This helped ensure that researchers (i) have all the necessary tools to manage their data over its life-cycle, (ii) do not need to find, configure, and adapt tools designed by different entities, and (iii) benefit from a consistent user experience and interface design. Second, given that data anonymization is a complex problem [23] and that existing LA anonymization tools are proposed only as proofs-of-concept [9], our desktop application is a significant step forward in giving researchers a starting point to anonymize their datasets. With pre-composed algorithms and a user-friendly GUI, researchers do not need to configure a development environment or write code to perform anonymization. Additionally, datasets can be viewed within the application, helping researchers verify the results of their anonymization pipelines. Finally, the application runs fully within a researcher's local environment, providing comfort that research data is not being transferred to or handled by remote servers.

VI. Conclusions

In this paper, we detailed a comprehensive architecture to manage the acquisition, visualization, anonymization, and sharing of LA data. Although the potential benefits of LA are well documented [2], realizing these benefits is made challenging by the diversity of the involved stakeholders [24] and the required tools [5]. Our contribution is to conceive of and deploy an integrated end-to-end toolkit that addresses these diverse stakeholders and their needs, and in doing so, helping to promote open data practices.

The architecture provides an extendable blueprint on how a digital education platform can be enhanced to include more comprehensive tools for its stakeholders. To highlight a few examples, in the outlined deployment we (i) created a direct communication channel between researchers and instructors, helping increase transparency in how research is framed and conducted, (ii) provided users with on-demand access to their complete datasets, and (iii) provided researchers with a user-friendly but versatile application to anonymize their datasets. While the architecture targets teacher-mediated contexts, its design is readily transferable to other domains, such as digital healthcare or e-government, and could also help support open data in these contexts.

One of the key limitations of the deployed architecture, as it stands, is its relatively narrow scope. Specifically, the desktop application is optimized for the anonymization of datasets generated by the associated educational platform. Although some of its algorithms are generalized and can be used on any dataset, more complete anonymization continues to require familiarity with the underlying dataset schema.

We highlight that—by design—the data collection, processing, and sharing process is fully controlled and under the responsibility of the instructors and researchers involved. They can personalize each step of the pipeline and validate the integrity of the final datasets using the tools and dashboards provided, or with custom-made visualization and validation algorithms.

VII. Future Work

Further generalization of the desktop application's anonymization algorithms is an area of work with significant

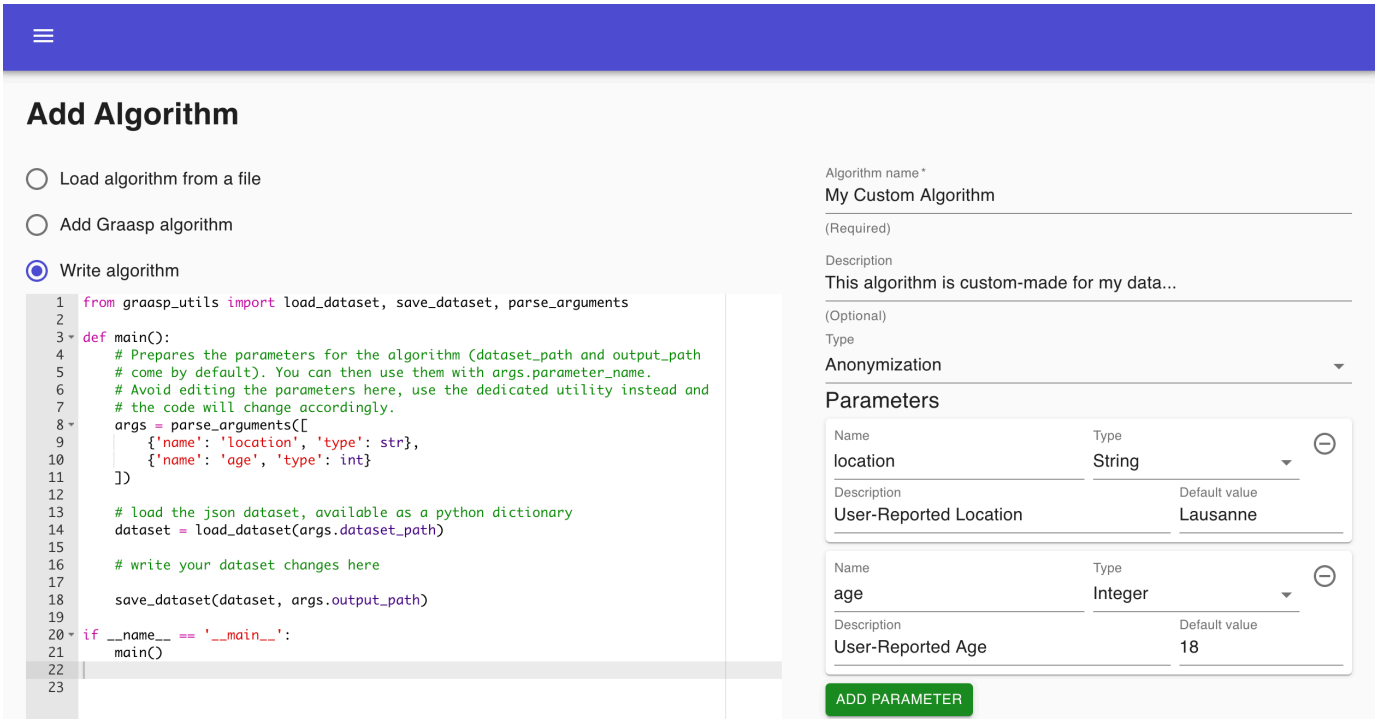

Fig. 5. Users can add their own custom algorithms using the built-in code editor. Algorithms can also be loaded from a file.

potential. User interface and workflow design can allow researchers to guide the application in (i) identifying and understanding a dataset's schema and (ii) tagging its quasi-identifying and sensitive attributes, after which generalized algorithms can be more accurately deployed, as implemented in ARX [25]. More extensive tests and validation regarding how effectively a dataset has been anonymized is another promising and important area of future work. Even after removing direct identifiers from a dataset, it is critical to perform de-identification checks to ensure that no record can be used to identify an individual. To that end, the application may introduce comprehensive data anonymity checks that quantify the privacy levels attained after applying anonymization algorithms, displaying such outcomes to users. In particular, deep learning solutions in the domain of natural language processing have been recently proposed for generically detecting values in text that could potentially be sensitive, achieving promising results [26], [27]. Those solutions could be an important asset in digital education but have yet to be tested in this domain. Finally, as data literacy grows, there is significant potential in making the desktop application available directly to individual users, who can use the application to view, understand, and potentially anonymize the personal datasets they acquire via *subject access requests* on the internet platforms they use. Following these and other use cases, we believe that our tools can inform the design of responsible data collection and dissemination policies not only for digital education, but also for other domains where guaranteeing user privacy is paramount.

Acknowledgment

This research has been partially funded by the European Union (grant agreement nos. 731685 and 669074), as well as the swissuniversities P5 program. We would also like to thank Hagop Taminian, Kim Lan Phan Hoang, and Badr Larhdir for their contributions to this project.

References

[1] S. Kellogg and A. Edelmann, "Massively Open Online Course for Educators (MOOC-Ed) Network Dataset: MOOC-Ed Network Dataset," *British Journal of Educational Technology*, vol. 46, no. 5, pp. 977–983, 2015.

[2] G. Siemens, D. Gasevic, C. Haythornthwaite, S. Dawson, S. Buckingham Shum, R. Ferguson, E. Duval, K. Verbert, and R. S. J. d. Baker, "Open Learning Analytics: An Integrated & Modularized Platform," Society for Learning Analytics Research, Tech. Rep., 2011.

[3] S. Dietze, G. Siemens, D. Taibi, and H. Drachsler, "Editorial: Datasets for Learning Analytics," *Journal of Learning Analytics*, vol. 3, no. 2, pp. 307–311, 2016.

[4] J. A. Ruipérez-Valiente, S. Halawa, R. Slama, and J. Reich, "Using Multi-Platform Learning Analytics to Compare Regional and Global MOOC Learning in the Arab World," *Computers & Education*, vol. 146, 2020.

[5] J. S. Machado, J. C. Farah, D. Gillet, and M. J. Rodríguez-Triana, "Towards Open Data in Digital Education Platforms," in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, vol. 2161-377X. New York, NY, USA: IEEE, 2019, pp. 209–211.

[6] 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 *2013 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2013, pp. 1255–1261.

[7] T. Hoel and W. Chen, "Learning Analytics Interoperability - Looking for Low-Hanging Fruits," in *Proceedings of the 22nd International Conference on Computers in Education*. Asia-Pacific Society for Computers in Education, 2014.

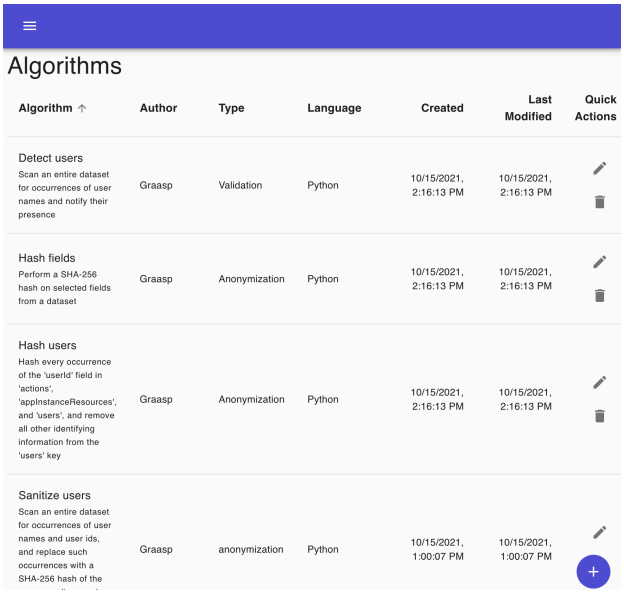

Fig. 6. The *Algorithms* tab, where users can add pre-composed or custom anonymization algorithms.

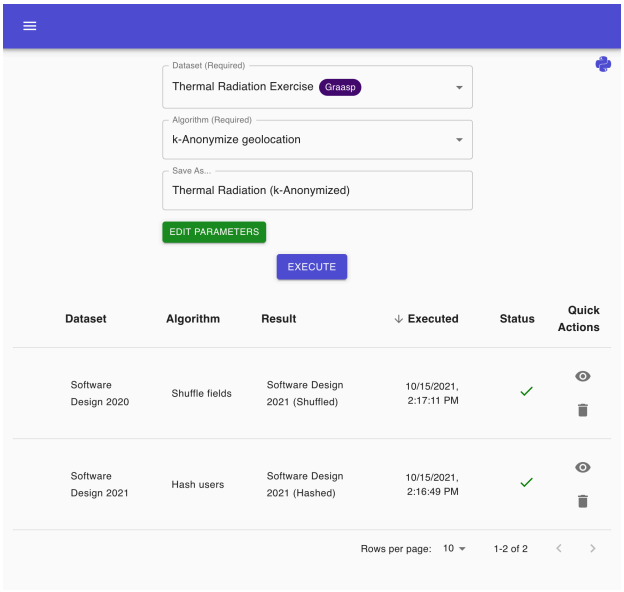

Fig. 7. The *Executions* tab, where users can select the datasets they have imported and run pre-composed or custom algorithms against them.

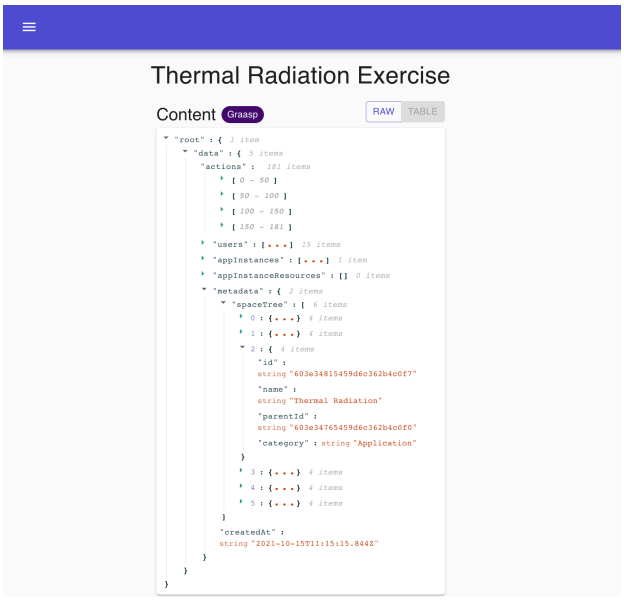

Fig. 8. A built-in JSON editor allows users to view and edit original and anonymized datasets within the application. Datasets can also be exported to be viewed and edited using external tools.

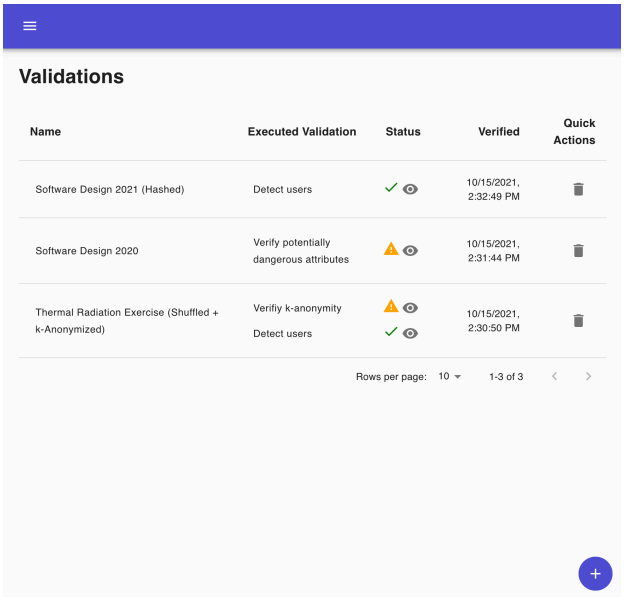

Fig. 9. The *Validations* tab, where users can run anonymity verification algorithms against the resulting datasets and see the outcome of the tests (success, warning, or failure).

[8] A. Berg, M. Scheffel, H. Drachsler, S. Ternier, and M. Specht, "Dutch Cooking with xAPI Recipes: The Good, the Bad, and the Consistent," in *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 2016, pp. 234–236.

[9] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-Preserving Learning Analytics: Challenges and Techniques," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 68–81, 2017.

[10] W. Jiang and C. Clifton, "Privacy-Preserving Distributed k-Anonymity," in *Data and Applications Security XIX*, S. Jajodia and D. Wijesekera, Eds. Springer Berlin Heidelberg, 2005, pp. 166–177.

[11] T. Komarova, D. Nekipelov, A. Al Rafi, and E. Yakovlev, "K-Anonymity: A Note on the Trade-off between Data Utility and Data Security," *Applied Econometrics*, vol. 48, pp. 44–62, 2017.

[12] L. Demir, A. Kumar, M. Cunche, and C. Lauradoux, "The Pitfalls of Hashing for Privacy," *IEEE Communications Surveys Tutorials*, vol. 20, no. 1, pp. 551–565, 2018.

[13] J. Chicaiza, M. C. Cabrera-Loayza, R. Elizalde, and N. Piedra, "Application of Data Anonymization in Learning Analytics," in *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, ser. APPIS 2020. ACM, 2020.

[14] J. Heurix, P. Zimmermann, T. Neubauer, and S. Fenz, "A Taxonomy for Privacy Enhancing Technologies," *Computers & Security*, vol. 53, 2015.

[15] C. M. Steiner, M. D. Kickmeier-Rust, and D. Albert, "LEA in Private: A Privacy and Data Protection Framework for a Learning Analytics Toolbox," *Journal of Learning Analytics*, vol. 3, no. 1, 2016.

[16] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,

"L-Diversity: Privacy beyond k-Anonymity," in *22nd International Conference on Data Engineering (ICDE'06)*. IEEE, 2006.

[17] N. Li, T. Li, and S. Venkatasubramanian, "T-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[18] R. Majumdar, A. Akćapınar, G. Akćapınar, B. Flanagan, and H. Ogata, "LAViEW: Learning Analytics Dashboard Towards Evidence-based Education," in *Companion Proceedings 9th International Conference on Learning Analytics & Knowledge (LAK19)*, 2019.

[19] J. Nicholson and I. Tasker, "DataExchange: Privacy by Design for Data Sharing in Education," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, 2017, pp. 92–97.

[20] D. Gillet, "Personal Learning Environments as Enablers for Connectivist MOOCs," in *2013 12th International Conference on Information Technology Based Higher Education and Training (ITHET)*. IEEE, 2013.

[21] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[22] J. C. Farah, A. Vozniuk, M. J. Rodríguez-Triana, and D. Gillet, "A Blueprint for a Blockchain-Based Architecture to Power a Distributed Network of Tamper-Evident Learning Trace Repositories," in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 2018, pp. 218–222.

[23] I. Wagner and D. Eckhoff, "Technical Privacy Metrics: A Systematic Survey," *ACM Computing Surveys*, vol. 51, no. 3, Jun. 2018.

[24] N. Sclater, "Developing a Code of Practice for Learning Analytics," *Journal of Learning Analytics*, vol. 3, no. 1, 2016.

[25] F. Prasser, J. Eicher, H. Spengler, R. Bild, and K. A. Kuhn, "Flexible Data Anonymization Using ARX—Current Status and Challenges Ahead," *Software: Practice and Experience*, vol. 50, no. 7, pp. 1277–1304, 2020.

[26] F. Hassan, D. Sánchez, J. Soria-Comas, and J. Domingo-Ferrer, "Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2019, pp. 358–365.

[27] G. Xu, C. Qi, H. Yu, S. Xu, C. Zhao, and J. Yuan, "Detecting Sensitive Information of Unstructured Text Using Convolutional Neural Network," in *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 2019, pp. 474–479.