Promoting Open Science in Educational Data Mining: An Interactive Tutorial on Licensing, Data, and Containers

Aaron Haim
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
01609, USA
ahaim@wpi.edu

Stephen Hutt
University of Denver
2199 South University
Boulevard
Denver, Colorado, USA
stephen.hutt@du.edu

Stacy T. Shaw
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
01609, USA
sshaw@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, Massachusetts
01609, USA
nth@wpi.edu

ABSTRACT

Across the past decade, the open science movement has increased its momentum, making research more openly available and reproducible across different environments. In parallel, data mining within education has provided a better understanding of the behaviors and interactions among students and teachers. However, there is a discernible gap between the understanding and application of open science practices in data mining. In this tutorial, we will expand the knowledge base towards open data and open analysis. First, we will introduce the complexities of intellectual property and licensing within open science. Next, we will provide insights into data sharing methods that preserve the privacy of participants. Finally, we will conclude with an interactive demonstration on sharing research materials reproducibly. We will tailor the content towards the needs and goals of the participants, enabling researchers with the necessary resources and knowledge to implement these concepts effectively and responsibly.

Keywords

Open Science, Reproducibility, Licensing

1. INTRODUCTION

Open science and robust reproducibility practices are becoming increasingly adopted within numerous scientific disciplines. Within subfields of educational technology, however, the adoption and review of these practices are sparsely implemented, typically due to a lack of time or incentive to do so [1, 15]. Some subfields of education technology have introduced open science practices (special education [4], gamification [6], education research [14]); however, others have seen little to no adoption. Authors have numerous concerns and minimal experience in what can be made publicly available, such as datasets and analysis code [7]. Additionally, research made publicly available is not typically reproducible without additional effort to fix unnoticed issues [8]. A lack of discussion can lead to repetitive communication, irrecoverable processes, or a reproducibility crisis within a field of study [2]. As such, there is a need for accessible resources, providing an understanding of open science practices, how they can be used, and how to mitigate potential issues that may arise at a later date.

Following the success of previous tutorials at the *13th International Conference on Learning Analytics* [10], *24th International Conference on Artificial Intelligence in Education* [12], *15th and 16th International Conference on Educational Data Mining* [16, 9], and *10th ACM Conference on Learning @ Scale* [11] along with an accepted tutorial to be presented at the *14th International Conference on Learning Analytics*[1], this tutorial aims to expand the knowledge base of participants towards two concepts of open science: open data and open analysis. First, we will discuss the issues of intellectual property and licensing within openly available work. Next, we will provide an overview of data sharing methods that preserve participant privacy and align with data collection agreements, through the use of data enclaves and anonymized datasets. Finally, we will conclude with an interactive example of how to share materials in a reproducible way using the current best practices. Throughout the tutorial, we will adapt to the needs and goals of participants, addressing concerns and providing resources tailored to them.

2. BACKGROUND

Open Science is a transformative movement that advocates for the democratization of scientific knowledge. At its core, Open Science seeks to make scientific research, data, and dissemination accessible to all, breaking down the barriers

A. Haim, S. Hutt, S. T. Shaw, and N. T. Heffernan. Promoting open science in educational data mining: An interactive tutorial on licensing, data, and containers. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 1017–1020, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730037

[1] https://doi.org/10.17605/osf.io/kja8r

of paywalls, proprietary databases, and closed-access publications. It is built on the principles of transparency, collaboration, and shared knowledge. The goals of Open Science are to advance the pace of discovery but also foster a more inclusive, equitable, and accountable scientific community.

As with many things, the translation from ideals and principles into real-world implementation comes with considerable challenges. For example, open access publication typically comes with a higher cost for the researcher (in turn damaging goals of equity and accessibility). Similarly, in education research, data sharing often poses challenges. Data are typically collected in partnership with educators, administrators, and students, who authorize the collection of data for a specific study/set of research questions, and often actively prohibit the distribution of data to third parties. Data can be deidentified, but given how intrinsically personal educational data can be, this task can be labor-intensive. Worse, some of the easier forms of deidentification (such as removing all forum post data prior to sharing[2]) lead to data no longer being usable for a wide range of research and development goals.

Sharing data on a by-request basis (e.g., Wolins, 1962 [19]) and carefully crafting data agreements has long been a potential solution, but it is often ineffective. For example, (Wicherts, Borsboom, Kats, & Molenaar, 2006) [18] contacted owners of 249 datasets, only receiving a response from 25.7%., a response rate similar to that noted in (Wolins, 1962) [19] following requesting data from 37 APA articles (though many years earlier and prior to email). Within education technology, (Haim et al., 2023) [8] contacted the authors of 594 papers, only receiving a response from 37, or 6.2%, of which only 19 responded that their dataset is public or could be requested. Some of the reasons cited were a lack of rights necessary to release the dataset, personally identifiable information was present, or that the dataset itself was part of an ongoing study. The task of sharing data requires a time investment from researchers, typically with no incentive. Moreover, the process can be stalled by changes in email addresses or institutions.

Work has been done to address Open Science principles specifically in education research, through Open Education Science [17], a subfield of Open Science [5]. This movement seeks to address problems of transparency and access, specifically in education research, addressing issues of publication bias, lack of access to original published research, and the failure to replicate. The practices proposed by Open Education Science fall into four categories, each related to a phase in the process of educational research: 1) open design, 2) open data, 3) open analysis, and 4) open publication. Of most relevance to the current tutorial are Open Data and Open Analysis.

Open Data ensures research data and materials are freely available on public platforms, aiding in replication, assessment, and close examination. However, there can be challenges, especially with educational data. There might be initial agreements that prevent data sharing or issues related to personal identifiable information (PII) which restrict what can be made public.

Open Analysis, on the other hand, emphasizes that analytical methods should be reproducible. This is commonly achieved by sharing the code used for analyses. Such code is typically shared on platforms like GitHub or preregistration websites. But there is a catch: the code is often of limited value without the associated data. Simply put, without Open Data, achieving Open Analysis can be tough. Moreover, there are challenges like ”code rot” and ”dependency hell” [3], where changing libraries can render older code unrunnable.

3. TUTORIAL GOALS

The tutorial focuses on introducing some common open science practices and their usage within education technology along with some interactive examples on how to apply the concepts in research. The target audience is researchers, as the practices offer structure and robustness. Based on past tutorials, we anticipate 5-10 participants and will design an interactive session tailored to their experiences and questions. This approach will allow us to present a responsive tutorial and foster additional community around open science topics.

3.1 Prior to the Conference

Prior to the conference, we will be compiling and organizing all relevant materials and resources. These will be published on a dedicated website, ensuring participants have easy access both during and after the tutorial. In addition, we will request all registrants to complete a pre-survey (using the participant registration list following the author registration/early registration deadlines). This survey aims to gather insights about participants' prior experience with the topics and their specific expectations from the tutorial. This data will be instrumental in allowing us to customize the tutorial, ensuring it meets the individual needs of participants and fostering an engaging and interactive session.

3.2 During the Conference

Our tutorial session will be an interactive and responsive session split into three sections. These sections are outlined below:

1. We will begin the tutorial by discussing how Intellectual Property (IP) intersects with the Open Science Framework. We'll tackle any questions or concerns from attendees about this topic. Our focus will be on code licensing, guided by the principles from Creative Commons. We will discuss why licensing code is important, strategies to safeguard a researcher's intellectual property, and provide guidelines for both Tech Transfer and University IP protection.

2. In the next segment of our tutorial, we discuss Open Data relative to the needs of participants. We anticipate opening this section by again addressing participant concerns to frame our future discussion. This will include identifying personal, moral, institutional, or legal concerns regarding open data.

   Participants will be introduced to the concept of Data Enclaves. This will cover understanding the primary

[2] https://edx.readthedocs.io/projects/devdata/en/latest/using/package.html

objectives of sharing data (including identifying the goals of the individual research team, the relationship between Data Enclaves and GDPR/Privacy legislation, and real-world examples of accessing information via these enclaves. Furthermore, we will provide valuable resources on establishing and efficiently using Data Enclaves.

We will also discuss how researchers can share data sets after they have been anonymized, ensuring the identity of participants remains confidential. We will also provide insight into the creation and sharing of synthetic datasets that mimic real datasets without using actual data.

As part of this section of the tutorial, we will also extend the discussion from EDM 2023 surrounding the right to erasure (commonly referred to as ”the right to be forgotten”) and how this may impact Open Science goals [13]. This discussion will complement the discussion of data enclaves, but also encourage attendees to consider how the Open Science guidelines and mandates (e.g., NSF) may be juxtaposed with privacy legislation (e.g., GDPR).

We will close this segment of the tutorial with a general discussion, weighing the advantages and drawbacks of each of the aforementioned approaches. This will not only deepen participant understanding but also help them draw parallels to their own research objectives and needs. Throughout this section, we will emphasize that there is not a “one size fits all” solution and that researchers should make choices based on individual goals and requirements.

3. Finally, we will provide instruction towards sharing materials in a reproducible manner, including best practices on storage, documentation, and privacy. This will be demonstrated with an interactive example using development containers via Visual Studio Code[3] and Docker[4]. The specific example used will depend on information gathered from the participants in the survey prior to the conference.

3.3 Following the Conference

After the conference, all additional resources created for the tutorial will be uploaded to the project's homepage for preservation. As this tutorial wants to repeat and expand upon open science and reproducibility at prior tutorials across conferences, an additional project will be created on the OSF website containing components pointing to all previous conferences and resources. A post-survey will be available at the end and after the tutorial to obtain feedback about the presentation for future use. An aggregate of the response will also be made public on the project's homepage. A community group on Discord will be created to collect, communicate, and discuss open science and reproducibility following the tutorial.

4. ORGANIZERS

Aaron Haim[5] is a Ph.D. student in Computer Science at Worcester Polytechnic Institute. His primary research focuses on reviewing, surveying, and compiling information related to open science and testing, documenting, and fixing the reproducibility of papers published at education technology and learning science conferences. His secondary focus is on developing software and running experiments on crowdsourced, on-demand assistance in the form of hints and explanations.

Stephen Hutt[6] is an Assistant Professor of Computer Science at the University of Denver. He has previously studied or worked in departments of computer science, cognitive science and education. His research is at the intersection of Artificial Intelligence, Learning Science, and Cognitive Science. Specifically, he considers how we can use state-of-the-art techniques to examine the complex internal cognitive and noncognitive processes commonly experienced in learning. He uses these insights to develop dynamic and adaptive learning technologies as well as contribute to broader theories.

Stacy T. Shaw[7] is an Assistant Professor of Psychology and Learning Sciences at Worcester Polytechnic Institute. She is an ambassador for the Center for Open Science, a catalyst for the Berkeley Initiative in Transparency in Social Sciences, and serves on the EdArXiv Preprint steering committee. Her research focuses on mathematics education, student experiences, creativity, and rest.

Neil T. Heffernan[8] is the William Smith Dean's Professor of Computer Science and Director of the Learning Sciences & Technology Program at Worcester Polytechnic Institute. He co-founded ASSISTments, a web-based learning platform, which he developed not only to help teachers be more effective in the classroom, but also so that he could use the platform to conduct studies to improve the quality of education. He has been involved in research papers containing some of the largest openly accessible data and materials in addition to convincing the Educational Data Mining conference to use the Open Science badges when researchers are submitting papers.

5. REFERENCES

[1] K. Armeni, L. Brinkman, R. Carlsson, A. Eerland, R. Fijten, R. Fondberg, V. E. Heininga, S. Heunis, W. Q. Koh, M. Masselink, N. Moran, A. O. Baoill, A. Sarafoglou, A. Schettino, H. Schwamm, Z. Sjoerds, M. Teperek, O. R. van den Akker, A. van't Veer, and R. Zurita-Milla. Towards wide-scale adoption of open science practices: The role of open science communities. *Science and Public Policy*, 48(5):605–611, 07 2021.

[2] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, May 2016.

[3] C. Boettiger. An introduction to docker for reproducible research. *SIGOPS Oper. Syst. Rev.*, 49(1):71–79, jan 2015.

[4] B. G. Cook, L. W. Collins, S. C. Cook, and L. Cook. A replication by any other name: A systematic review of replicative intervention studies. *Remedial and*

[3] https://code.visualstudio.com/

[4] https://www.docker.com/

[5] https://ahaim.ashwork.net/

[6] https://sjhutt.com/

[7] http://stacytshaw.com/

[8] https://www.neilheffernan.net/

Special Education, 37(4):223–234, 2016.

[5] B. Fecher and S. Friesike. *Open Science: One Term, Five Schools of Thought*, pages 17–47. Springer International Publishing, Cham, 2014.

[6] A. García-Holgado, F. J. García-Peñalvo, C. de la Higuera, A. Teixeira, U.-D. Ehlers, J. Bruton, F. Nascimbeni, N. Padilla Zea, and D. Burgos. Promoting open education through gamification in higher education: The opengame project. In *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM'20, page 399–404, New York, NY, USA, 2021. Association for Computing Machinery.

[7] A. Haim, C. Baxter, R. Gyurcsan, S. T. Shaw, and N. T. Heffernan. How to open science: Analyzing the open science statement compliance of the learning @ scale conference. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 174–182, New York, NY, USA, 2023. Association for Computing Machinery.

[8] A. Haim, R. Gyurcsan, C. Baxter, S. T. Shaw, and N. T. Heffernan. How to Open Science: Debugging Reproducibility within the Educational Data Mining Conference. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 114–124. International Educational Data Mining Society, July 2023.

[9] A. Haim, S. Shaw, and N. Heffernan. How to open science: Promoting principles and reproducibility practices within the educational data mining community. In M. Feng, T. KÃ¤ser, and P. Talukdar, editors, *Proceedings of the 16th International Conference on Educational Data Mining*, pages 582–584, Bengaluru, India, July 2023. International Educational Data Mining Society.

[10] A. Haim, S. T. Shaw, and I. Heffernan, Neil T. How to open science: Promoting principles and reproducibility practices within the learning analytics community, Jul 2023.

[11] A. Haim, S. T. Shaw, and N. T. Heffernan. How to open science: Promoting principles and reproducibility practices within the learning @ scale community. In *Proceedings of the Tenth ACM Conference on Learning @ Scale*, L@S '23, page 248–250, New York, NY, USA, 2023. Association for Computing Machinery.

[12] A. Haim, S. T. Shaw, and N. T. Heffernan. How to open science: Promoting principles and reproducibility practices within the artificial intelligence in education community. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, and O. C. Santos, editors, *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, pages 74–78, Cham, 2023. Springer Nature Switzerland.

[13] S. Hutt, S. Das, and R. S. Baker. The Right to Be Forgotten and Educational Data Mining: Challenges and Paths Forward. In *Proceedings of the 16th International Conference on Educational Data Mining, EDM 2023*. International Educational Data Mining Society, 2023. Publication Title: International Educational Data Mining Society ERIC Number: ED630886.

[14] M. C. Makel, K. N. Smith, M. T. McBee, S. J. Peters, and E. M. Miller. A path to greater credibility: Large-scale collaborative education research. *AERA Open*, 5(4):2332858419891963, 2019.

[15] B. Nosek. Making the most of the unconference, 2022.

[16] S. Shaw and A. Sales. Using the open science framework to promote open science in education research. In *Proceedings of the 15th International Conference on Educational Data Mining*, page 853–853. International Educational Data Mining Society, Jul 2022.

[17] T. van der Zee and J. Reich. Open education science. *AERA Open*, 4(3):2332858418787466, 2018.

[18] J. M. Wicherts, D. Borsboom, J. Kats, and D. Molenaar. The poor availability of psychological research data for reanalysis. *American Psychologist*, 61(7):726, 2006.

[19] L. Wolins. Responsibility for raw data. *American Psychologist*, 17(9):657–658, 1962.