Where's the Data? Exploring Datasets in Computing Education

Natalie Kiesler*
DIPF Leibniz Institute for Research and Information in Education
Frankfurt, Germany
kiesler@dipf.de

John Impagliazzo†
Hofstra University
Hempstead, New York, USA
john.impagliazzo@hofstra.edu

Katarzyna Biernacka
Humboldt-Universität zu Berlin
Berlin, Germany
katarzyna.biernacka@hu-berlin.de

Amanpreet Kapoor
University of Florida
Gainesville, Florida, USA
kapooramanpreet@ufl.edu

Zain Kazmi
Execusoft Solutions Inc.
Toronto, Ontario, Canada
zain.kazmi@execusoftsolutions.com

Sujeeth Goud Ramagoni
Marquette University
Milwaukee, Wisconsin, USA
sujeethgoud.ramagoni@marquette.edu

Aamod Sane
Flame University
Pune, India
aamod.sane@flame.edu.in

Keith Tran
NC State University
Raleigh, North Carolina, USA
ktran24@ncsu.edu

Shubbhi Taneja
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
staneja@wpi.edu

Zihan Wu
University of Michigan
Ann Arbor, Michigan, USA
ziwu@umich.edu

ABSTRACT

This working group aims to identify available datasets within the context of computing education research. One particular area of interest is programming education, and the data in question may include students' steps, progress, or submissions in the form of program code. To achieve this goal, the working group will review well-known data resources and repositories (e.g., DataShop, GitHub, NSF Public Access Repository, and IEEE DataPort) and recent papers published within the SIGCSE community. As a result of the review process, the working group will create an overview of available datasets and characterize them while reflecting on current data practices, challenges, and the consequences of limited access to research data. Additionally, the group intends to propose a path for the community to become more open and move toward open data practices. This proposal highlights the importance of sharing research data within the computing education research community to make it stronger and more productive.

CCS CONCEPTS

• **Social and professional topics** → **Computing education**; • **General and reference** → *Cross-computing tools and techniques.*

KEYWORDS

Open Data, Open Science, datasets, reusing data, computing education, secondary research, educational data mining

ACM Reference Format:
Natalie Kiesler, John Impagliazzo, Katarzyna Biernacka, Amanpreet Kapoor, Zain Kazmi, Sujeeth Goud Ramagoni, Aamod Sane, Keith Tran, Shubbhi Taneja, and Zihan Wu. 2023. Where's the Data? Exploring Datasets in Computing Education. In *Proceedings of the ACM Conference on Global Computing Education Vol 2 (CompEd 2023), December 5–9, 2023, Hyderabad, India.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3617650.3624951

1 INTRODUCTION

Computing education research aims to investigate teaching and learning processes to foster and enhance learning in the field. Such analysis usually relies on data to test hypotheses, build models, or improve systems. However, despite the increasing availability of data, it remains challenging to find open datasets in the context of computing education. Consequently, researchers may have to "reinvent the wheel" and gather data from scratch by themselves. This is usually a time-consuming, expensive process without any incentives or rewards for the researcher [11, 12].

Although open data strategies and guidelines have been promoted for years by big funding organizations in the U.S. and Europe [6, 14, 16], computing education researchers still struggle with finding and reusing data. For example, two recent working

*Working Group Leader
†Working Group Co-Leader

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CompEd 2023, December 5–9, 2023, Hyderabad, India

© 2023 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0374-4/23/12.
https://doi.org/10.1145/3617650.3624951

groups [7, 8] within the Innovation and Technology in Computer Science Education (ITiCSE) community expressed their challenges with finding publicly available datasets for their research and the limitations of available data. One grand challenge is openly publishing data from primary research [2, 3, 7]. Fortunately, some good practical examples are known where research data is provided to the community (e.g., [4, 5, 9, 10, 13, 15, 17]).

However, finding datasets for accessing and understanding them remains a barrier for several reasons [3, 12]. Among them are scarce resources for the initial search. Moreover, limited knowledge about, and access to (sensitive) data can hinder research in the field. So this working group intends to take action by addressing the following tasks (1) start developing a corpus/overview of openly available datasets in computing education research by reviewing well-known data resources and repositories (e.g., DataShop, GitHub, NSF Public Access Repository, and IEEE DataPort) and recent papers published within the computing education research (CER) community, (2) characterize available datasets, and (3) outline a path for more researchers willing to share their data. The overarching goal is thus to support researchers within the community who need data as a basis for their (secondary) research and who want to know examples of open datasets as a model for more data publications.

2 WORKING GROUP OBJECTIVES

This working group is motivated by the following objectives:

(1) Identify areas within CER and the types of data sought. The plan is to specify the contexts in alignment with the participants' interests and expertise. For example, programming education and data representing students' steps, progress, or submissions in program code is one area of interest.
(2) Find sources where researchers from the CER community can search for research data and identify available datasets relevant for computing education researchers interested in secondary research or needing data to investigate, for example, learning environments or other systems and models.
(3) Create an overview of available datasets to the CER community (e.g., in the form of the working group report) along with a new, qualitative characterization of the data.
(4) Reflect on and discuss current data practices, the limited access to research data, and related challenges when searching for data. The community perspective will be included based on a survey about current Open Data practices [1].
(5) Based on the community's input, we will propose options for a path forward for the CER community to become more open and move toward available data practices.

The working group activities will be aligned with the participants' interests. Nonetheless, the full working group report will yield an overview of available datasets, their characteristics, and potential for secondary research, along with insights into the CER community's data practices, and concerns regarding the publication of research data. Based on these milestones, the working group will discuss feasible options for open data practices at conferences, such as SIGCSE, ITiCSE, and CompEd in the future.

3 INTENDED RESULTS

The authors intend to deliver the following results:

(1) A collection of datasets and their characteristics useful for applications in the CER community.
(2) Methods and strategies for finding datasets useful in computing education and research.
(3) Examples of ways to apply the identified datasets in computing education and research.
(4) An overview of the community's perspective regarding Open Data practices.
(5) Recommendations to the CER community to develop, publish, and use datasets in meaningful ways.

These results will help enhance computing education by developing practical aspects of finding and reusing datasets in computing.

REFERENCES

[1] Katarzyna Biernacka, Adrian Mulligan, Jonathan Zimmermann, and Rudi Rudiak. 2023. Research Data Sharing and Reuse 2020. Online. https://doi.org/10.17632/nr9n75cpv2.1 Mendeley Data, V1.
[2] Katarzyna Biernacka and Niels Pinkwart. 2021. Opportunities for Adopting Open Research Data in Learning Analytics. In *Advancing the Power of Learning Analytics and Big Data in Education.* https://doi.org/10.4018/978-1-7998-7103-3.ch002
[3] Christine L. Borgman and Irene V. Pasquetto. 2017. Why Data Sharing and Reuse Are Hard To Do. https://escholarship.org/uc/item/0jj17309
[4] Neil Christopher Charles Brown, Michael Kölling, Davin McCall, and Ian Utting. 2014. Blackbox: A Large Scale Repository of Novice Programmers' Activity. In *Proceedings of the ACM Technical Symposium on Computer Science Education (SIGCSE).* https://doi.org/10.1145/2538862.2538924
[5] John Edwards. 2022. 2021 CS1 Keystroke Data. https://doi.org/10.7910/DVN/BVOF7S
[6] European Union. 2023. European Open Science Cloud. https://eosc-portal.eu/
[7] Petri Ihantola, Arto Vihavainen, Alireza Ahadi, Matthew Butler, Jürgen Börstler, Stephen H. Edwards, Essi Isohanni, Ari Korhonen, Andrew Petersen, Kelly Rivers, Miguel Ángel Rubio, Judy Sheard, Bronius Skupas, Jaime Spacco, Claudia Szabo, and Daniel Toll. 2015. Educational Data Mining and Learning Analytics in Programming: Literature Review and Case Studies. In *Proceedings of the 2015 ITiCSE on Working Group Reports (ITiCSE-WGR '15).* ACM, New York, 41–63.
[8] Johan Jeuring, Hieke Keuning, Samiha Marwan, Dennis Bouvier, Cruz Izu, Natalie Kiesler, Teemu Lehtinen, Dominic Lohr, Andrew Peterson, and Sami Sarsa. 2022. Towards Giving Timely Formative Feedback and Hints to Novice Programmers. In *Proceedings of the 2022 Working Group Reports on Innovation and Technology in Computer Science Education* (Dublin, Ireland) *(ITiCSE-WGR '22).* ACM, New York, 95–115. https://doi.org/10.1145/3571785.3574124
[9] Natalie Kiesler. 2022. Dataset: Recursive problem solving in the online learning environment CodingBat by computer science students. Online. https://doi.org/10.21249/DZHW:studentsteps:1.0.0 Datenerhebung: 2017. Version: 1.0.0. Datenpaketzugangsweg: Download-SUF. Hannover: FDZ-DZHW. Datenkuratierung: İkiz-Akıncı, Dilek.
[10] Natalie Kiesler and Benedikt Pfülb. 2023. Higher Education Programming Competencies: A Novel Dataset. In *Artificial Neural Networks and Machine Learning – ICANN 2023*, Lazaros Iliadis, Antonios Papaleonidas, Plamen Angelov, and Chrisina Jayne (Eds.). Springer Nature Switzerland, Cham, 319–330. https://doi.org/10.1007/978-3-031-44198-1_27
[11] Natalie Kiesler and Daniel Schiffner. 2022. On the Lack of Recognition of Software Artifacts and IT Infrastructure in Educational Technology Research. In *20. Fachtagung Bildungstechnologien (DELFI)*, Peter A. Henning, Michael Striewe, and Matthias Wölfel (Eds.). Gesellschaft für Informatik e.V., Bonn, 201–206. https://doi.org/10.18420/delfi2022-034
[12] Natalie Kiesler and Daniel Schiffner. 2023. Why We Need Open Data in Computer Science Education Research. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education Vol. 1* (Turku, Finland) *(ITiCSE 2023).* ACM, New York. https://doi.org/10.1145/3587102.3588860
[13] Michael Kölling and Ian Utting. 2012. Building an Open, Large-Scale Research Data Repository of Initial Programming Student Behaviour. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education (SIGCSE '12).* ACM, New York, 323–324. https://doi.org/10.1145/2157136.2157234
[14] National Science Foundation. 2013. Open Data at NSF. https://www.nsf.gov/data/
[15] Benjamin Paaßen. 2020. Python Programming Dataset. https://doi.org/10.4119/unibi/2941052 Bielefeld University.
[16] Dirk Pilat and Yukiko Fukasaku. 2007. OECD principles and guidelines for access to research data from public funding. *Data Science Journal* 6 (2007), OD4–OD11.
[17] Keith Quille and Keith Nolan. 2022. Predicting Success in CS1-An Open Access Data Project. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2.* ACM, New York. https://doi.org/10.1145/3478432.3499092