Where's the Data? Finding and Reusing Datasets in Computing Education

Natalie Kiesler*
Nuremberg Tech
Nuremberg, Germany
natalie.kiesler@th-nuernberg.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

Computing education research (CER) is a rapidly advancing discipline, offering vast potential for data-driven, secondary research or replication studies. Although gathering and analyzing data for research seem straightforward, making research data publicly available to the community remains a challenge. Likewise, finding and reusing high-quality, prominent, and well-documented research data proves to be a daunting task. In this working group paper, the authors present their search for available datasets in the CER context (e.g., in databases and repositories). The available datasets are further analyzed using a newly developed metadata scheme and presented to the community as a resource. The second component of this work is a summary of the community's perspective and concerns on publishing their research data, which has been gathered through a survey among 52 computing education researchers. Based on this status quo, this report presents recommendations for measures and future steps for the community to become more accessible and establish open data practices. We thus emphasize the potential of making research data available to enhance productivity, transparency, and reproducibility in the CER community.

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, programming process data, 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? Finding and Reusing Datasets in Computing Education. In *Working Group Reports of the 2023 ACM Conference on Global Computing Education (CompEd-WGR 2023), December 5–9, 2023, Hyderabad, India.* ACM, New York, NY, USA, 30 pages. https://doi.org/10.1145/3598579.3689378

1 Introduction

A core element of Computing Education Research (CER) is investigating students' learning processes, pedagogical practices, and their effects. The overall objective is usually to foster and enhance learning in the field. With educational processes subject to research, this area heavily relies on gathering data from students, educators, and educational institutions to develop and test hypotheses, evaluate learning environments, or improve instructional design patterns and curricula.

Such data is usually expensive, including prior research on the state-of-the-art, developing hypotheses and research instruments, and thoroughly preparing the study. The latter has to be conducted at a particular place at a specific time, and rooms may have to be booked. Survey questions must be prepared via digital tools, which

*Working Group Leader
†Working Group Co-Leader

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

ACM ISBN 979-8-4007-0222-8/23/12
https://doi.org/10.1145/3598579.3689378

should be piloted. Then, researchers need to find respondents, interview partners or subjects, or other types of data sources and documents. More importantly, as many computing education research studies occur within institutions, the data collection must follow the course schedule, and an Ethics Review Board application must be approved beforehand. If researchers gather verbal data, these must be transcribed before being analyzed. In short, running studies and collecting data can be expensive. Nevertheless, the resulting data from these studies can be described as a treasure.

Moreover, donating or publishing research data can be time-consuming and challenging [97]. If researchers do not publish their research and paper publications, finding, accessing, and understanding data remains a barrier [15]. Too often, it is common practice to "reinvent the wheel", meaning that researchers gather data from scratch again, even though a similar study may have already been conducted [79, 100]. Not only is that an expensive task in terms of time, effort, and resources – it also decreases the potential for secondary research or validation studies in our community, which could contribute to a replication crisis in CER similar to other disciplines (e.g., psychology [5]) [33, 46, 58].

The following examples illustrate the struggle of computing education researchers in finding and reusing other researchers' data. Three recent working groups [79, 85, 136] within the Innovation and Technology in Computer Science Education (ITiCSE) conference expressed their challenges when searching for and reusing publicly available datasets to support their research. These working groups concluded the following:

> "*Only a few data sets describing authentic solutions to programming problems are publicly available. [...] It is challenging to find nicely comparable data sets, and even in data sets originating from our environments, there are many details we don't know* [79]."

> "*We found five datasets with the desired characteristics. [...] Surprisingly, such datasets were hard to find. Datasets mentioned in previous work were either unavailable, like the code.org dataset, or unsuitable for various reasons* [85]."

> "*replicating prior work using newer models is difficult, given that a wide variety of parameters, prompts, and evaluation approaches have been used, and not all methods are reported with sufficient detail. Producing a dataset that contains everything necessary for high-quality LLM research [...] is challenging and needs to be encouraged by the community* [136]."

It is thus not only a challenge to find datasets but to access and understand them as the data might be limited or different from expectations. Only a few practical examples are known where research data is provided to the community (e.g., [19, 47, 90, 91, 95, 106, 130, 138, 148, 149]). Yet, we need rich and big data [17, 89].

To counteract this situation, many funding institutions and academic organizations have started pushing for open science, such as the Open Science Framework [57]. Others expect open data practices as a requirement [52, 120, 134]. However, it remains common practice to share knowledge and research by written texts only, which increases the challenges in today's digital age with its increasing reliance on data [65].

The majority of publishing methods are limited to "text" contributions. Data tends to be perceived as a side product of the epistemological process, with peer-review processes explicitly excluding supplementary material (i.e., data) [92, 100].

A consequence of this process is that other researchers rarely review data and they are hardly published. Researchers are not incentivized to publish their data if it is not enforced. Therefore, continuing, validating, or replicating research from others becomes a challenge, as these activities rely on not only a written summary of the research (i.e., in the form of a paper article) but also the details of how the study was conducted, including metadata, mature primary data, and documentation of its provenance.

Hao et al. [70] found that less than 3% of all computer science education research is reproduced or is replication studies. The authors conclude that the research must be verifiable if research results lead to new policies or practices. McGill [113] even concludes that the lack of open research has led to substandard research practices. Since research results impact computing educators, practitioners should be able to independently verify results that have lacked replication before implementing recommendations based on these results. A critical methodology for facilitating such independent verification would be the publication of datasets where possible. A report on behalf of the European Commission published a conservative estimate in 2019, calculating 10.2 billion Euros for the opportunity costs of the lack of FAIR research data [36]. FAIR refers to data that is findable, accessible, interoperable, and reusable [171].

1.1 Research Objectives

The overarching goal of this working group is to support researchers within the CER community who need data as a basis for their (secondary) research. Thus, we addressed the need for an overview of openly available resources and datasets in the CER community and a characterization of available data. Specifically, the working group was motivated by the following objectives:

(1) Find resources 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.
(2) 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 in the form of metadata.
(3) Reflect on and discuss current data practices, the limited access to research data, and related challenges when searching for data; Include the community perspective based on a survey about current Open Data practices [11].
(4) Gather the CER community's perspective on the publication of research data. Based on the community's input, propose options for a path forward for the CER community and its flagship conferences to become more open and move toward available data practices.

1.2 Contributions

The authors accomplished the following results, which are delivered in this report.

(1) A **collection of resources**, their characteristics, target group and link, useful for applications in educational research in computing in **Appendix A**.
(2) A **collection of datasets** [93][1] based on existing study data along with characteristics useful for applications in computing education research, and the computing classroom. The collection comprises quantitative, qualitative, and interaction log data from educational settings and environments focusing on introductory programming contexts. An excerpt of the analysis is presented in **Table 2**.
(3) The identified data is described with a newly developed **metadata scheme**, available in **Appendix B** and applied in **Table 2**.
(4) A summary of the **community's perspective** regarding open data practices in **section 7**.
(5) **Recommendations** on meaningfully developing and publishing datasets in computing education research in **section 9**, as well as a **checklist** to support researchers in collecting, managing, and releasing datasets in **Appendix C**.

These results will help enhance computing education practically and theoretically by developing hands-on recommendations for finding and reusing datasets in computing, but also for the publication of research data in the future. The time for open science is now [133].

1.3 Paper Organization

To address related work, the authors present the concept of openness and FAIR research data along with current barriers for researchers trying to search, find, and reuse datasets gathered by others in Section 2. The remaining working group report consists of two major components: a review of existing datasets for computing education in section 3 to section 5, and a survey regarding data practices in the CER community in section 6 to section 8. We propose three research questions related to reviewing datasets in subsection 3.1 and three research questions regarding CER data practices in subsection 6.1.

Hence, approaches to finding meaningful datasets from computing education contexts are presented in terms of the methodology applied to several resources and the inclusion and exclusion criteria of datasets for further analysis. Next, the identified datasets for computing education researchers and educators are characterized by a newly developed metadata scheme for this context. Moreover, the authors introduce other related datasets, along with examples of how to use these datasets for secondary research and in classroom settings. A discussion of the results, including limitations, will wrap up this first part of the report.

The second major component is the presentation of the methodology and results from a survey conducted within the CER community. The survey aimed to gain insights into the current practices and concerns regarding the publication of research data in the CER community and to derive a feasible path toward open data practices. Accordingly, this paper summarizes the survey methodology and participants' responses before discussing ways to move towards open data practices in the CER community. This results in a set of recommendations for various stakeholders.

[1]Overview via CS-SPLICE https://splice.cs.vt.edu/datasetcatalog/

2 Openness and FAIR Data

Before reflecting on the search for available datasets from Computing Education Research (CER) contexts, we introduce the concepts of Open Data, Published Data, and FAIRness concerning research data. Moreover, this work presents some challenges related to sharing or publishing research data. In addition, we summarize the challenges from the secondary researcher's perspective and common difficulties encountered when searching for and trying to reuse data gathered by others.

2.1 Open, Published, and FAIR Data

In research data management, it is essential to distinguish between three fundamental concepts: Open Data, Published Data, and FAIR Data, each of which plays a distinct role in advancing the accessibility and reusability of data.

Open Data is a fundamental aspect of Open Science, aiming to provide equal access to research knowledge. According to the Open Definition [127], Open Data should be freely accessible, usable, modifiable, and shareable by anyone for any purpose. These datasets are typically accessible through open licenses and designed to foster transparency. After a considerable reproducibility crisis starting in the early 2010s (discussed, e.g., in [80]), research transparency and data sharing have become crucial to rebuild trust in science. Early publication of research data can help reduce misconduct, support replication, and foster collaborations. Furthermore, Open Data follows the principle that data collected or generated with public resources should be open and easily accessible to maximize its societal benefits. Sensitive data, or data under an embargo, are excluded from that discussion.

Published and open research data can be distinguished by its citability, a facet guaranteed through the attribution of persistent identifiers (PIDs). Nonetheless, it is imperative to note that not all data subject to publication, i.e., endowed with PIDs, necessarily adheres to open data principles. Hence, not all published data is also open. Instead, the data may entail restricted access protocols.

Various institutions, including governments and funders, demand data accessibility in line with the FAIR Principles [172], which are beyond openness (i.e., accessible data) or mere publications (which are not necessarily open). The FAIR principles add to these constructs by definition, as they demand the following from data:
To be Findable:

F1 (meta)data are assigned a globally unique and persistent identifier
F2 data are described with rich metadata (defined by R1 below)
F3 metadata clearly and explicitly include the identifier of the data it describes
F4 (meta)data are registered or indexed in a searchable resource

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2 metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1 (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2 (meta)data use vocabularies that follow FAIR principles
- I3 (meta)data include qualified references to other (meta)data

To be Reusable:

- R1 meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1 (meta)data are released with a clear and accessible data usage license
- R1.2 (meta)data are associated with detailed provenance
- R1.3 (meta)data meet domain-relevant community standards

Conversely, these principles go beyond openness and publication to ensure that data is also Findable, Accessible, Interoperable, and Reusable. After all, these are relatively abstract guidelines intended to improve the re-usability and actual reuse of data (or other digital resources). To achieve FAIRness of the data, making it openly available to others is insufficient. The data should be well-documented, tagged with metadata, stored in repositories, formatted to facilitate interoperability, structured for easy reuse, and assigned with a persistent identifier. FAIR Data principles are crucial in scientific research, where data discoverability and interoperability are paramount. However, published or FAIR Data is uncommon in many disciplines, such as computing education research [100].

2.2 Barriers to the Publication of Datasets

A study conducted with 13 scientists from Germany, Peru, India, and China from a learning analytics context revealed that most scientists had not published their data, citing uncertainties about what is allowed legally, potential risks in data anonymization, and the loss of crucial information during the anonymization process [12]. The barriers to publishing research data derived from these interviews were categorized into five dimensions [12]:

- Legal concerns - barriers related to legal constraints or considerations affecting data publication
- Concerns regarding loss of control - barriers concerning the loss of control over data when published
- Authority or practice considerations - barriers related to authoritative guidelines or established practices influencing data publication decisions
- Technical/processing barriers - technical challenges or impediments encountered in the process of publishing data
- Resource barriers - resource limitations or constraints affecting data publication endeavors

The concerns can vary across countries. The substantial time and effort required for data preparation and publication and insufficient funding and infrastructure also emerge as significant obstacles, particularly in low-income countries [12].

The legal challenges or uncertainties are especially relevant for the CER community whenever data gathered from students is involved. If that is the case, sharing research data in CER comes with two major (legal) issues: privacy concerns [44] and anonymization. According to Reidenberg and Schaub [141], data from educational contexts should not be individually identifiable when collected. Even when the data is anonymized or pseudonymized before or after the analysis, the publication of data is rarely planned from the study's onset and, therefore, usually not included in the informed consent or approved by an Ethics Review Board. This situation leads to problems in a later stage as it is challenging to obtain the permit after collecting the data. This may not be possible even though no personal data was collected during the study. Therefore, researchers may wish to work with the institution review boards to reduce the risk of privacy and data protection early in the research process. Further, organizations must take adequate security measures (e.g., authentification processes and contracts) to prevent others from disseminating research data. In response to high failure rates of replication studies and publication bias, van der Zee and Reich proposed a framework for available education research to increase the transparency and access to educational, scientific research [166]. The open education research framework consists of four phases: design, data collection, analysis, and publication, and addresses the entire data life cycle.

Another aspect worth mentioning is the lack of recognition within the CER community for the publication of research data, which is related to the resource concern. The focus of publications lies in a written paper summarizing findings. The data is usually neither available nor enforced for review. However, preparing research data for sharing or publishing requires at least the same amount of time and effort and remains unrecognized [97, 100]. Guidelines for data review are hardly available and require much more attention from the community [99]. Moreover, publications related to datasets may come with considerable costs, e.g., conference attendance, registration, or Open Access fees [101].

2.3 Challenges When Searching for and Reusing Data

Finding research data gathered by others remains time-consuming and has an uncertain outcome. For example, if data is announced in a publication as available upon request, a researcher trying to access them must rely on the corresponding author's availability and responsiveness. Even if one is fortunate enough to receive the requested data promptly, the data may inhibit further obstacles, as there is no standard open science infrastructure [133]. Kiesler and Schiffner [100] summarize several of these challenges when searching for and reusing data:

- Lack of (long-term) availability, e.g., researchers may have left academia, files may have been lost, proprietary software required, the software does not run anymore [63, 87, 88, 97].
- Lack of comprehensibility/maturity of data, e.g., high-quality documentation requiring time, effort, and resources.
- Lack of attribution for high quality, mature data, e.g., the effort needed for mature datasets is not a standard key performance indicator in academia [117].
- Lack of obligatory data reviews, e.g., publication formats and venues rarely demand the submission of research data; data reviews are uncommon, reinforcing a low data maturity level.
- Lack of data provenance, e.g., the origin of data, changes, data cleaning steps, errors, and other aspects relevant to derive meaning out of the data for secondary research [25].

To conclude, finding research data does not guarantee technical access to the data, understanding them, and being able to use them

in secondary research or for replication studies. The data may be poorly documented with metadata, lack maturity or provenance, or may not fit the research question (see, e.g., [84, 85, 100]).

2.4 Metadata

One crucial facet contributing to the findability of datasets involves structured data about data, commonly referred to as metadata. Simply put, metadata is data that describes other data. These descriptors elucidate the structure of objects, providing important administrative details concerning rights and ownership. In the best case, metadata consists of the minimum necessary information to (re-)use the research data.

Despite the existence of numerous metadata standards [116], there is currently no suitable schema available for data within the domain of CER. Some efforts are currently being made by researchers in this field, such as the CS-SPLICE[2] project [21], a working group that aims at providing reusable content, tools, and infrastructure for computing education. One of their goals is to provide formats and tools for analyzing learner data.

There is, however, a related proposal of a metadata model for the context of learning analytics (LAMM) [173]. In the implementation of this schema, established standards such as Dublin Core [45], DataCite [41], and RADAR [64] have already been considered. Subsequently, an expansion was implemented to incorporate discipline-specific properties.

To ensure the FAIRness of data, it is imperative to describe the data using specific attributes (i.e., access rights or persistent identifier) that should be integrated into the metadata schema. These attributes play a critical role in enabling the findability, accessibility, and comprehensibility of data, fostering their reusability across diverse domains, and facilitating seamless interoperability among different systems and platforms. Incorporating these characteristics into the metadata schema significantly enhances the overall quality and utility of the data, aligning with the FAIR principles and promoting their effective utilization by both humans and machines.

Utilizing markup languages such as XML, HTML, or JSON facilitates enhanced accessibility by embedding metadata. These languages enable structured representation, helping search engines comprehend and index content for improved discoverability. Using markup languages supports uniform data processing across diverse systems and applications, facilitating data exchange and interoperability among software and platforms. Overall, integrating markup languages for metadata provision can significantly enhance the visibility, accessibility, and utility of digital content within today's interconnected online landscape.

3 Method for Finding and Characterizing Meaningful Datasets

In the rapidly evolving CER landscape, open datasets have emerged as an essential resource for empirical investigation, theory validation, and replication for researchers. While the broader computing field has plenty of data resources, the CER community faces unique data availability and utility challenges. This pressing issue of data scarcity in CER exacerbates the already notable gap in the organized availability and accessibility of open datasets despite the prevalence of data-centric studies [18].

This section presents the methodology of a comprehensive, quasi-systematic data search and meta-analysis for identifying data relevant to CER. Our work is inspired by pioneering fragmented projects like Blackbox [19], which aimed to consolidate and house student data from specific tools (BlueJ in this case [105]) for future scientific inquiries. We strive to serve as an overarching resource for CER researchers interested in secondary data analysis by shedding light on the landscape of available datasets, their characteristics, and the notable gaps in the types of collected data. In doing so, we set the stage for more open, collaborative research efforts highlighting exemplary datasets and establishing data standards, thereby addressing the existing challenges and gaps in the field.

3.1 Research Questions

The data search and meta-analysis is this report's first significant component. The following research questions guide it:

- RQ 1 (*Landscape*). *Which are resources for datasets conducive to computing education and computing education research?*
- RQ 2 (*Characteristics*). *How can we describe the identified datasets in terms of metadata to support computing education researchers and educators in reusing the data?*
- RQ 3 (*Gaps*). *What are the current limitations and challenges associated with the search for available datasets?*

To address these research questions, we identify, consolidate, and characterize resources, as well as published and open datasets within the context of computing education. We thus develop a methodology for finding data and summarizing relevant resources and databases. In addition, we shed light on current practices, identifying challenges and limitations in data publication practices in computing education (CE) and computing education research (CER). Based on the findings, we derive workflows and recommendations for educators and researchers who plan to release their data in the CER and associated communities regarding how to describe their data with metadata.

While addressing these tasks might seem onerous, we believe that we can make an initial attempt to bring ideal datasets and practices into the limelight in teaching, learning, and research. This working group intends to pave the way for other computing education practitioners and researchers searching for primary data within computing contexts.

As part of our data search, we expect to encounter several datasets only partially relevant to CER. Even though the developed metadata scheme does not fully characterize these datasets, we mention and briefly describe them in this paper.

3.2 Methodology of the Search Process

This section introduces the approaches to finding and identifying relevant datasets for the CER community. Accordingly, the selection of databases and resources is presented along with the applied search terms and strategies. We further elaborate on the inclusion and exclusion criteria for selecting datasets.

3.2.1 Search for Data Related to Scientific Articles. Several data sources were utilized to identify meaningful datasets. We decided

[2] https://cssplice.github.io/

to search for papers published within the Computing Education Research (CER) community that are available in (1) the ACM Digital Library (Full-Text Collection), (2) IEEE Xplore, (3) Taylor & Francis Online, and (4) Sage Publications. This choice was guided by work on the analysis of academic databases in CS Education by Valente et al. [165] and the recently published book "Past, Present, and Future of Computing Education Research: A Global Perspective", and its chapter [6, pp 121-150] on venues that have shaped computing education research. The goal was thus to cover the following relevant venues and their publications and determine whether datasets are among or at least connected to these publications:

- ACM International Computing Education Research (ICER)
- ACM Innovation and Technology in Computer Science Education (ITiCSE)
- ACM SIGCSE Technical Symposium on Computer Science Education (SIGCSE TS)
- ACM Global Computing Education Conference (CompEd)
- ACM Transactions on Computing Education (ToCE)
- IEEE Frontiers in Education (FIE)
- IEEE Transactions on Learning Technologies (TLT)
- IEEE Transactions on Education (ToE)
- Journal of Educational Computing Research (JECR)
- Australasian Computing Education (ACE)
- Koli Calling (Koli)
- ACM India Compute (COMPUTE)
- Consortium for Computing Sciences in Colleges (CCSC)
- Taylor & Francis Computer Science Education (CSE)
- Workshop in Primary and Secondary Computing Education (WIPSCE)

3.2.2 Search for Data in Repositories. In addition to the digital libraries containing scientific articles, we searched for data in standard repositories in the community. The initial list was informed by the recent related literature presenting datasets and respective data sources (e.g., [79, 85, 136, 161]) and resources well-known to the authors, following the purposeful sampling approach [132].

- DataShop
- Harvard Dataverse
- GitHub
- IEEE DataPort
- Kaggle
- Mendeley Data
- NSF Public Access Repository
- Open Science Framework (OSF)
- Papers with Code
- Zenodo

This list was later expanded by snowballing, resulting in the resources in Appendix A. We thus included additional data sources identified as part of the scientific articles data search, as some publications had data in other repositories linked to them.

3.2.3 Search Terms and Strategy. The search for data was based on keywords and filter options of the resources as appropriate. The queries were refined depending on the possibilities of different databases. A typical list of keywords was used for the databases with paper publications (i.e., ACM Digital Library, Taylor & Francis Online, and IEEE Xplore), focussing on openly available datasets and datasets linked to a publication (see list below). These keywords were combined with the venues' titles listed in subsection 3.2.1, which were available as filters. As an example, an excerpt of the query used when searching the ACM Digital Library was as follows:

```
[All: "publicly available"] OR
[All: "available online"] OR
[All: "link to dataset"] OR
[All: "data collected"] OR
[All: "data collection project"] OR
[All: "dataset available"] OR
[All: "dataset is available online"] OR
[All: "dataset"] OR
[All: "datas et"] OR
[All: "open source"] AND
[[Publication Title: "iticse"]
...
```

The keyword search through the repositories additionally included, for example, "computing education", "CS education", "student data", "educational data", and "log data" as well as several programming languages, and other areas of interest to the authors (e.g., "Java", "Python", "introductory programming", "Parallel and Distributed Computing", and "High-Performance Computing").

During the search through the repositories (listed in subsection 3.2.2), the CER-related venues' titles were also used as keywords as appropriate. The authors iterated and refactored this query multiple times to capture known literature and accurate results. The search for datasets started in October 2023 and ended on November 8, 2023. The authors excluded regular expression queries from databases that did not support regular expressions, and they used alternate equivalent queries.

Moreover, the searches were conducted using a filter for dates beginning in January 2014, as we chose only to include data from the past ten years. This decision was made because we expected that older datasets were unlikely to be accessible/available (anymore). Another limitation concerned the language of the data, which is English.

The search resulted in several hits within each resource. Table 1 represents the number of results per resource selected based on our inclusion and exclusion criteria described next.

3.2.4 Inclusion and Exclusion Criteria. The authors meticulously crafted the inclusion and exclusion criteria, a crucial step in the research process, to align with the research questions. This criterion was further honed after the data extraction process, ensuring the study's rigor and the reliability of the results.

The resources that met any of the following exclusion criteria were discarded from the corpus with the complete meta-analysis. Yet, we still provide a few examples of datasets we consider helpful for the CER community.

- Data from K-12 educational contexts
- Datasets typically used for Machine Learning applications (e.g., MNIST [109], or SVHN [122]) or ML courses
- Data from systematic literature reviews

The following criteria led to a complete exclusion of the datasets from any further analysis.

Table 1: Sources for the data search and included results for the meta-analysis.

Source	Chosen
ACM Digital Library	8
IEEE Xplore	1
Taylor & Francis Online	1
CAROL Data	0
Corgis	0
Datashop	2
Dataverse	0
German Center for University and Science Research	2
GitHub	2
Hugging Face	12
IEEE DataPort	8
Kaggle	0
MDPI	2
Mendeley Data	0
National Center for Education Statistics	0
National Data Resources	0
NSF Public Access Repository	0
OSF	1
Papers with Code	2
UC Irvine Machine Learning Repository	0
Zenodo	5

- General student data, e.g., on dropouts, demographics, graduate numbers, etc.
- Generic data from educational contexts, without mentioning computing education
- Instruments
- Data with aggregated results only
- Private datasets we could not access

The resources that met the following inclusion criteria became part of the pool of resources selected for review.

- Data from higher education computing contexts
- Data gathered from or representing computing student actions or students' learning processes in educational contexts
- Data gathered or representing computing educators actions or their perspectives related to computing education
- Data representing the industry's perspective on computing graduates
- Qualitative and quantitative data
- Data relevant to computing education researchers, and computing educators for classroom use
- Datasets containing both primary data and aggregated data (i.e., and not just highly aggregated data)

3.3 Meta-Analysis of Identified Datasets in Computing Education Research

One of the significant challenges within the CER community is the highly contextualized nature of datasets, which makes them challenging to share and limits their public availability. This situation poses a significant barrier for researchers aiming to undertake meta-analyses or build upon existing work. Addressing this gap is crucial as it will help unravel the collected data types, which will inform and better support future data-enabled research in CER. This subsection presents the methodology used to screen and characterize the datasets.

The approach to developing a metadata scheme was iterative and consisted of several phases. In the first phase, we gathered a list of available CSed-specific public datasets from various repositories such as DataShop@CMU, Kaggle, and others (see Appendix A). At the same time, this working group gathered an understanding of any standards for dataset metadata and annotations that may already be in place (e.g., ProgSnap, or ProgSnap2 specification [73, 137]).

Given the absence of an established metadata standard for describing CER data generally, a dedicated model was designed based on existing standards and schemas. Initially, intuitive metadata requirements of the CER researchers (particularly those involved in this working group) were identified. These identified properties were subsequently mapped and compared with existing metadata schemata. Dublin Core [45], Data Cite [41], RADAR [64], LOM [78], LAMM [173], and PREMIS [37] were included for this purpose. The mapping process was further augmented by incorporating discipline-specific metadata.

In the next phase, the initially designed metadata were tested and explored by analyzing three datasets identified in the search (see Table 2). This piloting step aimed to determine whether the metadata can be applied to various CER datasets and to what extent the metadata categories and vocabulary needed adjustments. Specifically, we used different data types to explore mandatory metadata: programming log data, survey data, and qualitative research data. After applying the initial metadata schema (a group of three authors of this work tested its application for the three datasets), we revised it using discussions and a consensual approach. This iterative approach was used until a consensus was reached among all three.

The last step was to create a complete list of metadata, vocabularies, and definitions for the specific CER context (i.e., programming education) to assist the authors and CER researchers in characterizing their data. These will help authors meet SIGCSE and community expectations for metadata, annotation, and research data publication in the field. The complete schema developed by this working group is presented in Appendix B as an answer to RQ 2. It aims to constitute a collection of minimum, recommended, and optional requirements for characterizing CER data and, simultaneously, ensure the FAIRness of the data. It is a starting point for the CER community moving towards a standard metadata format.

By implementing this design, we aim to focus on both the availability and utility of datasets in the CER community, which enable more impactful secondary analysis research in the future.

3.4 Verification of the FAIRness of the Selected Datasets

Assessing the FAIRness of a dataset is a crucial process in research data management, ensuring that data is Findable, Accessible, Interoperable, and Reusable (cf. subsection 2.1). For this reason, we systematically incorporated this as a last step of our data analysis methodology.

For the verification of FAIRness, it is possible to use an existing tool (e.g., F-UJI [54]) or checklists such as the FAIR-Aware Tool [7].

The former automatically checks the FAIRness of the data, while the latter has to be done by the researchers themselves.

One should note that in the automated tools (such as F-UJI), the FAIRness check mainly refers to a defined set of machine-verifiable parameters. The FAIR assessment is performed based on aggregated metadata; this includes metadata embedded in the data (landing) page, metadata retrieved from a PID provider (e.g., Datacite), and other services. Still, there is no content check of, for example, the descriptive metadata (and a corresponding assessment mainly refers to the automatically verifiable features of the data set). Therefore, a manual check is required in the second step.

This approach to evaluating a dataset's FAIRness was applied, as it ensures that the selected datasets are valuable and promotes efficient data sharing, collaboration, and meaningful insights across the scientific and research communities.

4 Data Search Results (RQ 1, RQ 2, RQ 3)

As the demand for data-driven insights grows within the CER community, the need for a comprehensive search for and meta-analysis of existing resources and datasets becomes increasingly apparent. This subsection overviews publicly available/published datasets for computing education researchers and classroom use cases within the selected resources (see Appendix A).

The complete meta-analysis of all analyzed datasets and descriptions with metadata is available via an OSF project [93]. A short version of the dataset catalog is hosted on the CS-SPLICE website[3]. Table 2 further serves as an example of the meta-analysis, as it provides the metadata for three exemplary datasets.

4.1 Datasets for Computing Education Researchers (RQ 1)

In this section, we present the results of our search for datasets in a summarized form. In particular, we distinguish between datasets specifically for computing education researchers, pedagogical use cases in classroom scenarios, and other datasets we came across. Since the datasets for computing education researchers focusing on introductory programming education were our primary focus, we described these datasets by using the newly developed metadata scheme (see Table 5). The complete meta-analysis is available in an OSF project [93].

4.1.1 Datasets on Introductory Programming Education. In the context of introductory programming education and respective research, datasets serve as invaluable resources for educators to understand how students grasp fundamental programming concepts and apply them in programming tasks to solve problems. Respective datasets often encompass diverse information across various subdomains, thereby challenging data analysis. The datasets may offer a broad range of functions, which inhibits comparisons. Moreover, the identified datasets often have deviating metadata or documentation that elucidates the data's structure and characteristics. Precise task descriptions, test cases, accepted (model) solutions, or other task limitations may be missing, so it can be challenging for other researchers and educators to evaluate the student data.

[3]https://splice.cs.vt.edu/datasetcatalog/

In addition, the identified datasets cover different levels of granularity. Some learning environments captured keystroke data, others full/final submissions, or so-called tokens as intermediate steps [85]. It is not always apparent from the description of the datasets which information is provided and how detailed or fine-grained students' actions have been collected.

Yet another challenge is the accessibility of such data. Researchers often report on their data in paper publications, but the primary data is not findable. In the presented search for data and the respective meta-analysis, the authors only considered published datasets and those available upon request. The following examples illustrate some of these results:

- CS1 Keystroke Data Utah State [47]:
  - Deidentified keystroke data collected from CS1 student participants during 2019 at Utah State University.
- CodeBench Dataset 1.80 [125]:
  - This dataset contains logs collected from CS1 students from 2016 to 2023, whereas each academic year is divided into two semesters. CodeBench automatically logs all actions performed by students via an embedded IDE during their attempts to solve programming exercises.
- Recursive problem-solving in the online learning environment CodingBat by computer science students [90, 91]:
  - Dataset with students' steps during recursive problem solving of two standard problems. The data includes students' task processing time, interactions with the learning environment, and use of feedback options.

4.1.2 Data from Systematic Literature Reviews. Another valuable resource the authors identified during the search for data was related to the publication of systematic literature reviews (SLRs). Details of SLRs are usually omitted in the paper publication due to page limitations. Yet, there is plenty of data regarding a systematic search and review. For example, initial search terms, queries and results, preprocessing steps, categorizations, notes, calculations, or aggregations. Ideally, all of these steps are documented in the primary data so that other researchers can understand and apply inclusion or exclusion criteria as part of an SLR or even replicate it.

A systematic literature review is a collaborative effort to produce a comprehensive and unbiased summary of current knowledge on a particular subject. It's a shared responsibility to provide all the details so other researchers can quickly identify relevant papers and sources. With access to the complete primary data and records of the selection process, other researchers can gain even greater value from SLRs. They may, for example, be able to identify biases or find more related publications from adjacent disciplines.

The search for data in the CER context led to several examples where the authors of SLRs published the respective research data:

- How Creatively Are We Teaching and Assessing Creativity in Computing Education: A Systematic Literature Review [68]
  - This dataset captures the systematic literature review process of the researchers team on collecting computing education papers exploring the role of creativity. Additional attributes such as significant themes, measurement instruments, and remarks were captured during the process.
- Cybersecurity Literature Review [169]

Table 2: Metadata scheme for three exemplary datasets (survey, log, and qualitative data).

Entity Name	Survey Data	Log Data	Qualitative Data
Descriptive Data			
title	METRECC Africa 2020 data	2021 CS1 Keystroke Data	Group work in Learning Programming (GAPL)
creator	Sentance, Sue; Tshukudu, Ethel; Quille, Keith	Edwards, John	Schulz, Sandra; Berndt, Sarah; Hawlitschek, Anja
givenName	Sue; Ethel; Keith	John	Sandra; Sarah; Anja
familyName	Sentance; Tshukudu; Quille	Edwards	Schulz; Berndt; Hawlitschek
nameIdentifier	https://orcid.org/0000-0002-0259-7408	https://orcid.org/0000-0002-1215-976X	https://orcid.org/0000-0002-2254-6579
affiliation	Raspberry Pi Computing Education Research Centre, University of Cambridge; University of Botswana; Technological University Dublin	Utah State University	Humboldt-Universität zu Berlin; Otto-von-Guericke Universität Magdeburg
URL	https://doi.org/10.17863/CAM.87121	https://doi.org/10.7910/DVN/BVOF7S	https://doi.org/10.21249/DZHW:dipit2020:1.0.0
urlType	DOI	DOI	DOI
Source	[151]	[47]	[150]
publisher	Apollo - University of Cambridge Repository	Dataverse - Harvard University Repository	DZHW German Centre for Higher Education Research and Science Studies
publicationYear	2022	2022	2023
rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)	CC0 1.0 Universal	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Germany (CC BY-NC-SA 3.0)
description	The study addresses K-12 computing education in four African countries (Botswana, Kenya, Nigeria, and Uganda). The available data comprises the survey structure and questions, responses from the 58 study participants to the survey questions on demographics, years of teaching experience, qualifications, classroom time, topics covered in computer science teaching ...	Keystroke data collected from CS1 student participants during fall 2021 semester at Utah State University.	The research project "Digital Programming in Teams" (DiP-iT) investigates how collaborative learning in computer science studies can be didactically developed and supported with digital tools. The project focuses on the use and implementation of learning analytics methods. The DiP-iT project aims to develop didactic and technical support ...
keywords	Botswana, computing education, computing teachers, K-12 computing, Kenya, Nigeria, Uganda	Keystroke, CS1, computing education	learning programming, computer science education, computer science, collaborative and cooperative learning, higher education
language	EN	N/A	DE/EN
version	N/A	6.0	1.0.0
availability	open	open	restricted
format	xlsx, txt, csv	csv, pdf, py, tsv	docx
dataType	Cross-sectional survey	Log data	Qualitative data
relatedPublication	[162]	[48]	[145]
CER-Specific Metadata			
collectionStart	2020-12-01	2021-10-18	2020-05-01
collectionEnd	2021-01-31	2021-12-10	2020-07-31
programmingLanguage	N/A	Python	N/A
population	teachers	students CS-1	Lecturers, students
sampleSize	58	44	25
sampleDemographics	Botswana; Kenya; Nigeria; Uganda	USA	Germany
country	Botswana; Kenya; Nigeria; Uganda	USA	Germany
educationalInstitution	N/A	Utah State University	TU Bergakademie Freiberg, Otto-von-Guericke-Universität Magdeburg und Humboldt-Universität zu Berlin
measurementType	Questionnaire	Keystroke data	Semi-structured Interviews
dataProcessing	N/A	N/A	N/A
unitsNumber	58	1048575	N/A
taskNumber	N/A	13 (8 part)	N/A
dataProtection	none	anonymized	anonymized
dataStandard	N/A	progsnap2	N/A
learningEnvironment	N/A	pyCharm/PyPhanon	N/A
aggregation	no	yes	N/A
aggregationLevel	N/A	Keystroke-level	N/A
Working Group Specific Metadata			
Research Questions	What is the capacity for delivering computing education in primary and secondary schools in four African countries from the teachers' perspectives?	What are the methods and challenges when creating a dataset for keystroke data?	1. To what extent are cooperative and collaborative learning activities/scenarios part of courses that aim to teach programming or improve students' programming skills? 2. What goals are pursued in connection with cooperative and collaborative learning activities? ...
FutureWorks	We intend to repeat the survey in subsequent years and analyze the data through the proposed capacity sub-components to develop a fuller picture as Africa develops its capacity for formal computing education...	What is a good programming process? Should students write their code linearly from start to finish, or should they move between different code sections to make changes?	Secondary research options and use cases: 1. Comparison with other locations in the same subject area. 2. Long-term development of the use of cooperative and collaborative learning activities and tools in foundation courses. ...
FAIRnessScore	Total 87% (advanced) Findable 7 of 7 Accessible 2 of 3 Interoperable 4 of 4 Reusable 8 of 10	Total 79% (moderate) Findable 7 of 7 Accessible 2 of 3 Interoperable 3 of 4 Reusable 7 of 10	Total 47% (initial) Findable 6 of 7 Accessible 1.5 of 3 Interoperable 1 of 4 Reusable 3 of 10

– These datasets present the result of a systematic literature process to find cybersecurity education papers from SIGCSE and ITiCSE Conferences. Specifically, the dataset contains 1) all the papers included in the automatic search, 2) manually excluded papers, and 3) papers included in the literature review.

4.1.3 Benchmarks in Computing Education Research. Computing education researchers develop new methods or tools for teaching and learning in computing, addressing, for example, loops, functions, or conditionals. However, when researchers try to compare methods or tools, they do not have readily available datasets of everyday examples to compare their new approaches [85]. If we

develop a new way to teach a concept such as loop termination, how do we know it improves over earlier methods, and in what context? It would be helpful to have benchmarks for such cases. In other fields, such methods exist. For instance, ImageNet [156] is a famous database of images in which words relate to related images. This dataset has proved immensely useful in comparing different computer vision methods and profound learning. Similar benchmarks exist for comparing advances in SAT solvers [83].

In computing education, some benchmarks have evolved organically. One famous benchmark of sorts is the "Rainfall problem"[152], which serves as a marker to measure students' progress in introductory courses. Another set of benchmarks is various "Concept Inventories" [158], especially validated concept inventories [135]. A concept inventory is a set of questions that validate whether a student has understood a programming concept, such as parameter passing. While benchmarks in CER cannot be as deterministic and accurate as with ImageNet or SAT solvers, they will help us as a field to study and calibrate advances in teaching methods or tools.

A recent ITiCSE working group report summarized 17 prominent datasets as benchmarks to evaluate, for example, Large Language Models and related tools [136]. Among them are the following examples:

- Search-Based Pseudocode-to-Code (SPOC) [107]:
  - Pseudocode descriptions of coding problems
- Blackbox [16, 19]:
  - Traces of editing and IDE interactions
- Deepfix [69]:
  - Student-generated code with syntax errors

4.2 Additional Datasets for Educators and the Classroom (RQ 1)

The following datasets are an additional product of our search for CER data. They provide educators and students with real-world examples and applications. However, we did not analyze and describe these datasets with the developed metadata scheme as they are beyond the scope of this working group's search for data. Nonetheless, these additional entities foster a greater understanding of complex concepts and potentially transform the learning and teaching processes. This section explores the possibilities of utilizing datasets in computing education.

4.2.1 Real-World Datasets. Computing education researchers and teachers sometimes use datasets to explore subject matter and ways of learning. In class, students can use real-world datasets to work on practical problem projects and gain experience in data analysis and visualization. Real-world datasets contain actual data compared to synthetic ones that people generate artificially. Such datasets are helpful in computing education because they expose students to meaningful data encountered in professional settings that provide context and relevance to the material taught. These datasets can contain large or small data content and relate to science, finance, and other subject areas presented in various structured or unstructured forms.

Real-world datasets can help students understand complex problems and formulate practical problem-solving skills across multiple computing topics from real-world settings. Students can learn about data privacy along with ethical and legal considerations. They also can learn about data pre-processing skills and tools commonly used in the workplace. They can also explore online data sources from repositories and other resource portals while they excel in studying real-world problems. Additionally, instructors can monitor student data work for assessment and development. Hence, using real-world datasets in computing education can provide students with an in-depth experience to understand computing concepts in the real world. In the following, we present some examples

- Canadian Institute for Cybersecurity [22, 23]:
  - The Canadian Institute for Cybersecurity (CIC) is a comprehensive multidisciplinary training, research and development, and entrepreneurial unit that draws on the expertise of researchers in the social sciences, business, computer science, engineering, law, and science.
- OpenAI's GPT-3 Playground Usage Data [20, 126]:
  - OpenAI has released usage data for its GPT-3 Playground, which is helpful when analyzing how developers and students interact with AI-based educational tools for coding and programming.

4.2.2 Machine Learning Datasets. Creating a curriculum with ever-changing market demands and a heterogeneous student body is non-trivial. To comply with the evolving needs of graduating students, computing educators must introduce courses on, for example, Machine Learning (ML) and Artificial Intelligence (AI). The rapid development of new models in the field demands constant curriculum extensions and changes in an already compact curriculum. If not adequately prepared, many computing undergraduate students may not have first-hand experience with a formal introduction to an industry-standard ML class. Hence, there is an urgent need to identify, create, improve, and distribute ML datasets to expose students to using ML datasets in existing computing courses as identified by Way et al. [170].

Using ML datasets for multidisciplinary research can help students better understand other fields outside of computing. In contrast, in the field of education specifically, these datasets can provide personalized learning, predict course grades, provide intelligent tutoring, etc. From these perspectives, the inclusion of ML datasets in computing courses has the potential to address challenges associated with teaching computing courses and provide opportunities to enhance learning. In the following, we present some examples.

- Iris (via UC Irvine's Machine Learning Repository) [55, 163]:
  - The UC Irvine Machine Learning Repository is a collection of datasets, domain theories, and data generators used by the ML community, and one of these datasets is Iris. Iris is one of the classic datasets based on evaluating classification methods and is used widely in statistics and ML.
- TensorFlow's movielens [71]:
  - Movielens dataset contains a set of movie ratings from the MovieLens website, a movie recommendation service. GroupLens, a research group at the University of Minnesota, collected and maintained this dataset. There are five different versions of this dataset available of the following sizes -"25m", "latest-small", "100k", "1m", and "20m".

Many datasets exist in repositories such as Kaggle, Zenodo, the UC Irvine Machine Learning Repository, or GitHub. Finding datasets that fit a curriculum's specific needs may be challenging. Additionally, many educational curricula are scaffolded or have micro-curricula embedded within their course. Thus, ML datasets would need to be modified to fit these goals. We encourage educators to share datasets they have used in their courses, gathered, or processed themselves in conjunction with the learning activities, assessments, and course objectives to help other educators reuse their datasets in the classroom.

4.3 Other Datasets (RQ 1)

In addition to our primary scope of datasets, we present other datasets that may be relevant to the CER community. Educational Data Mining (EDM) is a growing body of research that often overlaps with CER, creating research sub-fields such as "CS-EDM". The primary goal of the field is the development of analytics, tools, and technologies using the underlying dataset, as opposed to the more general goals of studying the data itself. Moreover, we present examples of K-12 computing education and industry datasets. The datasets from K-12 education are often limited in their availability, as they relate to minors, and are usually of particular interest to primary and secondary education researchers.

4.3.1 Educational Data Mining. EDM essentially falls under the data mining and machine learning area [8, 153]. However, its distinguishing feature lies in the exclusive utilization of educational datasets. In this context, "educational" encompasses any data for learning purposes, from essential alphabet acquisition to complex subjects like rocket science. EDM's primary objective revolves around understanding how humans learn within specific settings and enhancing the effectiveness of this learning process, which can take various forms [174].

EDM primarily centers its attention on the educational landscape of universities and schools, dissecting data related to students' learning mechanisms and outcomes. These data can encompass the subjects they are studying, their learning methods (including the role of teachers), their learning behavior, and their assessment performance [42].

The availability of a dataset is a pivotal element within the EDM workflow. Once a dataset is acquired, the subsequent steps in the research methodology involve identifying research issues, designing and implementing data analysis pipelines, and presenting validation results [42]. As the EDM field continually expands with the emergence of new tools and technologies, a significant bottleneck issue arises in the form of a well-documented review of publicly available datasets.

Researchers have identified three primary data sources for EDM: well-known data sources, datasets employed in EDM competitions, and standalone EDM datasets [8]. The future success of EDM data sources hinges on their ability to manage proposed approaches efficiently and their experimental outcomes as a benchmarked reference. In this context, the reproducibility of data analysis pipelines and the benchmarking of proposed algorithms become crucial for the research community. This approach allows for more accessible advancements in the EDM domain.

Ultimately, the critical outcome lies in continuously improving existing data analysis pipelines by addressing EDM tasks reliant on publicly accessible datasets and benchmarking these pipelines using open-source implementations. This approach fosters progress and innovation within the EDM field.

Given the importance of this topic, it is imperative to narrow the focus of the data mining process specifically to CER. Providing open access to datasets related to CS education opens the door for researchers, offering them a chance to delve into and understand the primary challenges unique to various regions. Some of the notable EDM data resources comprise the UCI Machine Learning Repository [142], VisualData [168], CMU Libraries [110], the NLP Index [81], and Google's Dataset Search [66]. Some examples of datasets for EDM comprise the following:

- Massive Open Online Courses Datasets [4, 30]:
  - MOOC platforms like Coursera, edX, and Udacity often release datasets related to student engagement, performance, and course content. These datasets can be valuable for research on online computing education.
- Open University Data [108]:
  - The Open University Online Learning Platform (known as "Virtual Learning Environment(VLE)") collects data from off-campus students and their access to course content, forum discussions, assessments, etc.

4.3.2 K-12 Educational Data. Of the education data in data sets such as the National Center for Educational Statistics [119], they are concerned with the issues of K12 education. K12 education is on an immense scale, is publicly supported, and is a significant public policy matter. The data of concern in this context is related to teaching quality, student performance, and other public policy measures. This type of data is not particularly relevant for CER in higher education. However, some CER researchers may be interested in this. We, therefore, present a few datasets we were able to identify during the search process (but excluded from our final presentation of results in the form of the meta-analysis):

- Code.org's Annual State of Computer Science Education Report [34, 35]:
  - Code.org releases annual reports on the state of computer science education in the United States. These reports include data on the availability and access to computer science education in K-12 schools.
- International Computer and Information Literacy Study (ICILS) 2018 Dataset [82]:
  - ICILS is a large-scale assessment of students' computer and information literacy skills. It provides data on students' use of computers, their problem-solving abilities, and attitudes toward technology.
- Program for International Student Assessment (PISA) dataset [129]:
  - The PISA dataset includes information about students' mathematics, science, and reading performance in various countries. It helps study the effectiveness of computer science education across countries.

4.3.3 Industry Datasets. Exploring industry datasets (which were not the focus of our research) relating to computing graduates

yielded minimal results. Datasets relating to computing graduates primarily consisted of summary degree statistics published by National Governments or academic hiring trends. We present examples below:

- Computer Science Open Data [74]:
  - Professor hiring trends, stipends and best papers in the USA
- National Centre for Education Statistics [118]:
  - Degree conferrals by post-secondary institutions in the USA
- Pakistan Intellectual Capital [164]:
  - Dataset containing University Computing Professor data across 89 universities in Pakistan

Although valuable in some cases, these datasets do not provide the much-needed insight educators require when developing industry-aligned curricula. Educators require industry perceptions and details regarding student struggles and performance to create meaningful learning objectives within their courses [131, 157]. This could include some combination of a) interviews and surveys with employers regarding the perception of knowledge/skill gaps of new hires, b) perceptions of new graduates of their struggles in the industry, or c) a dataset of elements that should be added to curricula, that may enhance CS graduate experiences when entering the industry. Researchers may also be interested in data on their students' career selection process.

It is possible to find some information in datasets that cover various aspects of computing education, technology adoption, and workforce trends. For example, the Integrated Postsecondary Education Data System (IPEDS) [56] provides data on U.S. colleges and universities, including computing programs, enrollment, and graduation rates. Additionally, the Tech Industry Workforce Reports provide information from organizations like the Computing Technology Industry Association (CompTIA) [38] or the Bureau of Labor Statistics (BLS) [124] with insights into trends in the technology workforce. The European Data Portal - Education Dataset [51] provides education-related datasets in European countries, including data on technology adoption. Also, the LinkedIn Economic Graph [112] has insights into workforce trends, skills demand, and job market dynamics through its Economic Graph initiative.

In some cases, researchers use industry data to create industry-aligned curricula. For example, Knapp et al. [104] discusses a cybersecurity curriculum set in a US university based on industry data.

4.4 Characteristics of Datasets (RQ 2)

The dataset search has clearly shown no standard for describing data in CER. Even the basic information that would be necessary for a correct citation is missing in some cases. Despite some significant effort, it was sometimes difficult or impossible to find specific characteristics about the data, despite an extended, in-depth search outside the landing page or the ReadMe file of the data.

In particular, the identified CER datasets demonstrated significant deficiencies concerning their adherence to the FAIR Data principles. Specific attributes crucial to meeting these principles, such as persistent identifiers and basic information like the publication year, are absent in the dataset description. Of the 38 datasets checked for FAIRness, only one was labeled with the *advanced* level with 87% score achievement. A *moderate* level was reached by 34% (n = 13). More than 60% of the datasets (n=24) did not exceed the *initial* level, which shows the datasets' shortcomings in findability, accessibility, interoperability, and reusability.

For this reason, a dedicated metadata schema was developed for CER data (cf. Section 3.3 and Appendix B), which was used to collect and supplement the information on the datasets found. This metadata schema can be used to characterize data in the CER community.

4.5 Limitations and Challenges Associated with the Search of Datasets (RQ 3)

The standard practice in CER involves limited metadata provision and lacks repositories optimized for searchability, resulting in enhanced complexity in data retrieval. Search engines struggle to accurately index the data due to the absence of specific keywords (i.e., metadata). Furthermore, the lack of persistent identifiers like DOI frequently leads to broken links. So, even if the data is findable through metadata, it may not be accessible via the provided link. Access limitations, commonly imposed by paywalls or login barriers, present another obstacle in obtaining the queried data.

4.5.1 Challenges and Gaps in Open CER Datasets. Understanding the existing challenges and gaps is crucial for developing supportive infrastructure and guidelines that facilitate more effective use of open datasets in CER.

One of the challenges associated with open data in the field of CER pertains to the absence of metadata and, frequently, the absence of persistent identifiers. The data typically lacks metadata and is often deposited on websites without such essential descriptive information. Consequently, conventional search engines do not index them. Although this data is open broadly, it is not findable. Furthermore, the prospects for subsequent data reuse remain ambiguous in numerous cases, as clear usage licenses are frequently absent, contributing to challenges and insecurity regarding secondary use.

While some of the data in CER is deposited in repositories widely recognized within the community, their accessibility often falls short of facilitating scholarly citation or open access to the data itself. Many datasets are behind paywalls or login barriers, impeding free and open exploration. Additionally, specific datasets are only available upon request, yet establishing contact with the original data creators proves challenging or, at times, nearly impossible. This limited accessibility restricts the broader utilization of these valuable datasets. It inhibits the crucial aspect of replicating and verifying research findings, hindering CER's progress and collaborative nature.

Based on the aspects above, it is evident that datasets in CER do not adhere to the FAIR principles (the FAIRness score according to [54] thus remains 'Initial' for most datasets). Consequently, this complicates the findability, accessibility, interoperability, and reusability of the data. This discrepancy underscores the need to enhance the adherence of CER datasets to FAIR principles, thereby fostering a more robust and accessible foundation for future research endeavors.

4.5.2 Lack of Datasets (in Parallel and Distributed Computing). Unfortunately, there were several areas in which we could not identify any datasets. For example, there is a lot of effort to introduce the

concepts of parallel and distributed computing (PDC) early in computing courses. The NSF supports a center called CDER, which helps faculty develop PDC curricula and introduces PDC topics in their existing courses. Computing curricula are already so diverse that there is almost no room to teach PDC as a standalone course, so there is a need to introduce PDC concepts in the existing courses offered across U.S. institutions and worldwide [31, 140].

Introducing PDC to CS undergraduate students is, however, challenging. One obstacle for faculty is accessing a distributed cluster where students can run their programs. If access to high-performance computing (HPC) systems is possible, the next challenge is the lack of suitable datasets for students to learn more about the PDC Concepts. A recent paper [167] discusses the lack of rich datasets that can help students understand various operational metrics related to the operational aspects of data centers housing these HPC systems. Let us consider the case of graphics processing units (GPU), the most pervasive component of HPC systems today, without exclusive access to GPUs (which is very common). Researchers and students rely on data center operators to provide hardware-level information to the cluster users. Rich data sets that provide system-level or hardware-level information are currently lacking and thus can prevent researchers from developing optimization or hardware provisioning algorithms. Without such optimization or hardware provisioning algorithms, HPC systems will remain energy-inefficient and low-performing.

4.5.3 Limitations of the Data Search. We have searched various resources (e.g., papers, databases, etc.) using existing data. Since fields such as machine learning, psychology, public policy, and economics are inherently data-driven, these (sub-)disciplines somewhat dominate existing data resources and repositories. Thus, searching for CER-oriented data has produced slightly limited results. Data-driven scientific research disciplines dominate the datasets associated with open data initiatives. Currently, CER-relevant data does not seem overtly present or easily findable in the investigated resources, among them a few research data centers. Moreover, we excluded those datasets from the meta-analysis that were not openly available (e.g., private or available upon request).

The data search has been driven by keywords and filter options provided by the various resources. Our selection of keywords is discussed in section 3.2.3. However, not all resources offered the same filter options to refine the search results. Therefore, applying the same search strategy to all resources was impossible. Unless sufficient metadata or a paper describing the datasets is available, it is also unclear how much discrimination is possible with such a keyword-based approach. This was particularly evident in large-scale repositories such as GitHub. Our search revealed the inconsistency of repository characterizations containing data in this context. For instance, a search of "CS education" and "datasets" in GitHub yielded a combination of results, mainly relating to datasets used in CS courses as opposed to datasets about the courses. Moreover, the results of keywords used in other repositories such as "education", "programming data", or "student data" were broad search specifications. Results did not necessarily include datasets when attempting to narrow searches by including labels such as CS1, CS2, or conference titles (e.g., ITICSE, SIGCSE). Most of the results included either tools, code used to analyze datasets, or Machine Learning datasets, which we defined as out of scope. Due to this limitation, this working group perceives it as even more critical for the CER community to develop a protocol for reporting results in a standard format.

5 Discussion of the Search for and Access to Datasets

While open datasets offer many research opportunities, they also present challenges that inhibit their full potential. Drawing on case examples such as Blackbox, this subsection delves into the barriers researchers face in accessing and utilizing these datasets, from inconsistent metadata to varying data granularity and scope levels. We further discuss the implications of our search on using datasets in computing education.

5.1 Towards Standard Data Formats

Data-driven methodologies for analysis, content generation, and intelligent tutoring systems have been a significant focus of computer science education research. Recent developments in the field include, for example, learning environments providing feedback to students and counteracting limited human resources in supervising growing classroom sizes [85]. Investigating students' progress, however, can be challenging if the tracked data varies concerning its granularity (e.g., keystroke, line-by-line, or complete submission). As a result, computing education researchers have started to develop standardization, interoperability, and data-sharing tools.

One example is ProgSnap2, a proposed specification for datasets containing programming process data [137]. Many teams worldwide have adopted these standards as part of their online educational platforms/tools [59]. As a working group, we encourage this movement towards similar specifications and standard data formats.

5.2 Supporting the Publication of Open Datasets

During the data search, we noted that none of the queried repositories had been built specifically for the CER community. A first step to publishing research data in CER may be establishing a (specialized) data repository. Available repositories or data sources exist, including meaningful datasets (e.g., DataShop, GitHub, Zenodo, etc.). Still, they do not necessarily adhere to the FAIR principles (e.g., DataShop does not provide a PID). Moreover, generic repositories, such as Zenodo, which follows the FAIR principles, do not meet the subject-specific requirements of the community, as it does not address software publications, for example, the required maturity levels, and it does not align with the software development life cycle [97]. Moreover, we could not fill out every metadata field, either because the data was not collected or due to a lack of standardization across datasets. An alternative or short-term solution might be the development of guidelines for authors on utilizing existing repositories that follow the FAIR principles and support crucial metadata.

A second important point is the description of the data with metadata to enable (partly automated) retrieval and evaluation of the data on a meta-level. Due to the frequent personal reference of the data in CER and the associated assignment of usage rights to the data itself, automatic machine access - as recommended in the FAIR principles - is impossible.

By publishing computing education datasets, researchers can maximize the value of the data collection process and allow more secondary research to occur on top of the existing data. Sandve et al. [146], for example, suggests ten rules for reproducible computational research, so that other scholars can judge the provenance of data. As part of this working group, we also developed a checklist for publishing research data for CER researchers (see Appendix C). It can serve as a draft or potential solution to help address standard data description formats, quality assurance, legal issues, selecting a data infrastructure, and other general issues related to the publication of research data. The next step is to discuss this suggested checklist with other researchers and experts. Thus, the checklist will be subject to future research and work, and we encourage the community to use, evaluate, and refine it.

5.3 Using Datasets in Computing Education

Due to modern technologies allowing access to relevant material, the publication of large-scale datasets in educational contexts has increased [18]. The availability of datasets can advance active learning, whereby students can explore new knowledge by working with actual data. Students can also apply current knowledge to practical problems, thereby increasing their understanding of underlying (theoretical) principles and concepts.

Using actual data can make students more knowledgeable about expressing their thoughts and viewpoints and in-depth critical thinking, as it requires analytical skills. For example, educators can challenge (groups of) students with factual problems presented as one or more sets of data. Students can then analyze the situation and propose one or more resolutions to the problem. Regardless of whether students solve a problem, going through analytic thought processes and presenting a plausible solution become mentally enriching and, for example, increase students' awareness of biases represented within data. Thus, data literacy becomes crucial and allows students to become more proficient in a possible data-centric career [94].

Another advantage of datasets is their crossdisciplinary nature, making them a valuable resource for computing education. Such an approach encourages students to understand how concepts from one field can be applied to solve problems in another, promoting a holistic learning approach [139].

The use of datasets does not come without a price. One obstacle to dataset use is access to the technical infrastructure at home or an institution. Handling large datasets requires the necessary hardware and software to enable efficient use of the entity. Additionally, some worldwide laws and regulations protect individuals from public use of personal data. Using sensitive or personal data in educational contexts raises ethical concerns that could negate the use of some datasets. This raises the question of data literacy, where students and instructors may need training to use datasets in education effectively. Hence, although using datasets in education is noteworthy, educators must respect the concerns surrounding such use.

In conclusion, incorporating datasets into teaching and learning can become a dynamic approach that significantly enhances the educational experience. Datasets promote data literacy, enhance understanding, and facilitate interdisciplinary learning. They also can foster critical thinking and problem-solving skills. By using the potential of datasets, educators can equip students and future researchers with essential skills for a data-driven world.

However, issues such as privacy, ethics, accessibility, and other challenges may counteract the effective use of datasets. Therefore, each institution should evaluate the advantages and disadvantages of using datasets in its educational settings.

6 Surveying the CER Community

To understand the challenges computing education researchers and practitioners face when releasing or working with computing education-related datasets, we designed a survey-based study to elicit feedback from the broader CER community on data release and usage practices, concerns, preferences, and expectations. Data analysis from this study could inform the community on best practices related to dataset procurement, data management, and data release. In this section, we will describe the methodology for our survey-based study.

6.1 Research Questions

In the second part of our working group report, we aim to answer the following research questions through our study:

- RQ 4 (*Practices*). What practices do computing education researchers and practitioners follow regarding dataset usage, management, and release?
- RQ 5 (*Barriers*). What barriers and challenges are associated with publishing datasets in the CER community?
- RQ 6 (*Preferences*). What are the computing education researchers' expectations and preferences regarding published datasets in CER?

Each research question will be addressed by several survey questions to develop the respective answer (see Table 3).

6.2 Survey Development

To answer our research questions, we designed a survey-based empirical study. The primary objective of the survey in CER was to address the escalating need for openly accessible data within this scientific domain, aiming to realize the potential advantages, including increased transparency, reproducibility, and enhanced collaboration in research. The present study is based on a cooperative survey between the Humboldt-Universität zu Berlin, and Elsevier [11]. The motivation originated from the growing significance of data sharing in modern scientific methodologies and the pressing necessity to overcome barriers impeding the widespread dissemination of research data. To maintain consistency, the questionnaire items were formulated to align with the framework of the 2016 Open Data Survey [28] and expanded to include specific inquiries aimed at data publication practices, such as exploring various places for publication and sharing of data. Moreover, it incorporated an assessment of barriers encountered during the publication process, systematically categorized into five primary domains (cf. section 2.2).

Additionally, the survey included questions addressing data privacy concerns, encompassing aspects related to safeguarding sensitive information during the publication process.

Table 3: Categorization and Relationship of Research and Survey Questions

Category	Research Question	Survey Question (see Appendix D for details)
Usage and release practices (categories of data, formats, etc.)	RQ 4 (Practices). What practices are followed by computing education researchers and practitioners regarding dataset usage, management, and release?	Q1, Q2, Q4, Q5, Q6, Q7, Q15
Barriers (legal, technical, resources, etc.)	RQ 5 (Barriers). What are the barriers and challenges associated with the publication of datasets?	Q8, Q9, Q10, Q11, Q12, Q13, Q14
Preferences and expectations	RQ 6 (Preferences). What are the computing education researchers' expectations and preferences regarding published datasets in CER?	Q3, Q16, Q17, Q18

The original survey instrument [11] was then adapted to the CER community regarding projects and venues. In addition, the survey questions were selected and mapped to our research questions (see Table 3). The survey questions, including answer options, are available in Appendix D. Conducting the survey was approved by the Ethics Board of the DIPF | Leibniz Institute for Research and Information in Education in early August 2023.

The discipline-specific questions were adapted to the discipline of computing education for the survey on research data within the CER community. Moreover, we use the definition of research data from the German Research Foundation (DFG). They consider research data to include, among others,

> "*measurement data, laboratory values, audiovisual information, texts, survey or observation data, methodological test procedures, and questionnaires. Compilations and simulations can likewise constitute a key outcome of academic research and are therefore also included under the term research data. Research data in some subject areas is based on the analysis of objects [. . .]. The same applies if software is required to create or process research data.* [61]"

The authors excluded PDFs and other files merely for publication from this definition and provided this definition to the survey respondents before asking any questions.

For the survey, we asked participants to think of a representative research project they conducted within the context of Computing Education in which research data were produced or gathered.

6.3 Survey Distribution (Sampling)

To elicit feedback from computing education researchers and practitioners, we applied the purposeful sampling method [132]. Specifically, we distributed our survey from September to December 2023 to prominent community listservs associated with the following groups within the computing education research context:

- Special Interest Group on Computer Science Education (SIGCSE)
- Special Interest Group on Human Factors in Computing Systems (SIGCHI)
- Special Interest Group on Information Technology Education (SIGITE)
- Standards, Protocols, and Learning Infrastructure for Computing Education Project (CSSPLICE)
- Learning engineering Google group [121]
- Educational Data Mining (EDM)

In addition, the working group's members purposefully emailed known colleagues who worked in the computing education area (thereby covering the U.S., Europe, and Asia regions) to share their feedback on data practices. Moreover, we presented the survey to the attendees during the ACM Global Computing Education Conference (CompEd) in Hyderabad, asking for their participation.

6.4 Data Analysis Method

The data analysis was guided by the research questions presented earlier. Table 3 represents how the survey questions align with the research questions and address the main foci of the survey: Practices, Barriers, and Preferences.

To analyze the survey data, we use frequency analysis and descriptive statistics to describe quantitative results and thematic analysis [32] for coding open-ended qualitative responses. We also add representative quotes from respondents for each theme to contextualize the readers better and validate our themes.

7 Survey Results (RQ 4, RQ 5, RQ 6)

This section summarizes survey results to the questions described in Appendix D. The community's responses to the survey serve as an answer to RQ 4, RQ 5, and RQ 6.

The survey was available between the end of September and December 2023, leading to 76 responses. However, 24 researchers have not responded to any of the survey questions, leading us to provide an analysis of responses from 52 individuals. The research data, i.e., the replies to the survey, are also published as part of an Open Science Framework project [93].

7.1 Usage and Release Practices (RQ 4)

There are seven survey questions related to research question, RQ 4 (Practices), as indicated in Table 3. The related survey questions are Q1, Q2, Q4, Q5, Q6, Q7, and Q15. The following summarizes the outcome for each of these survey questions. The aggregated results are discussed with percentages rounded to the nearest integer.

Q1 addressed research data categories used by the researchers and had 52 respondents. Of these, 27 (52%) respondents had developed their own software, while 28 (54%) respondents used software developed by others. Regarding qualitative data, 35 (67%) used their own data, while 18 researchers (34%) used data created by others. Quantitative data was also used: 41 (79%) respondents used their own, while 19 (36%) used data by others. Regarding derived or compiled data, 11 (21%) respondents said they gathered their own data, while 7 (13%) collected the data of others. Regarding reference or canonical data, 6 (11%) respondents gathered that data independently, and 5 researchers (9%) used data from others.

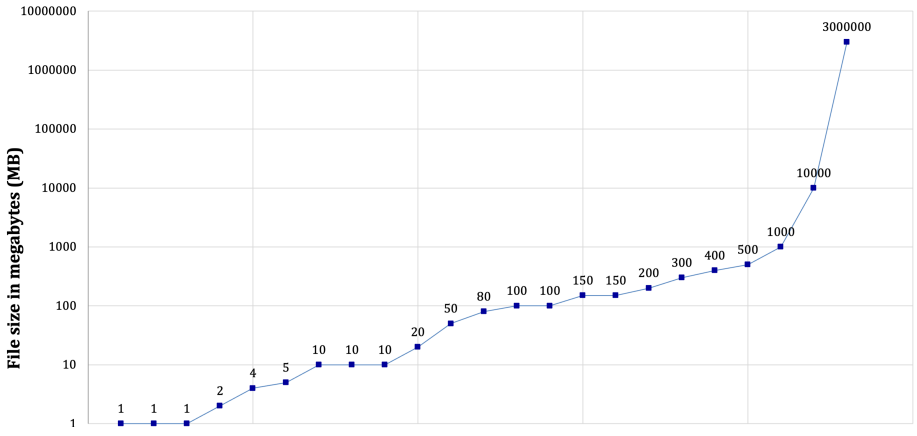

Figure 1: Responses indicating the volume of data created and shared for CSEd projects (Survey Question Q4, n=23).
Note: One respondent was excluded as they added 0 as the data volume, and the data volume was scaled to higher orders of bytes using a factor of 1000 rather than 1024.

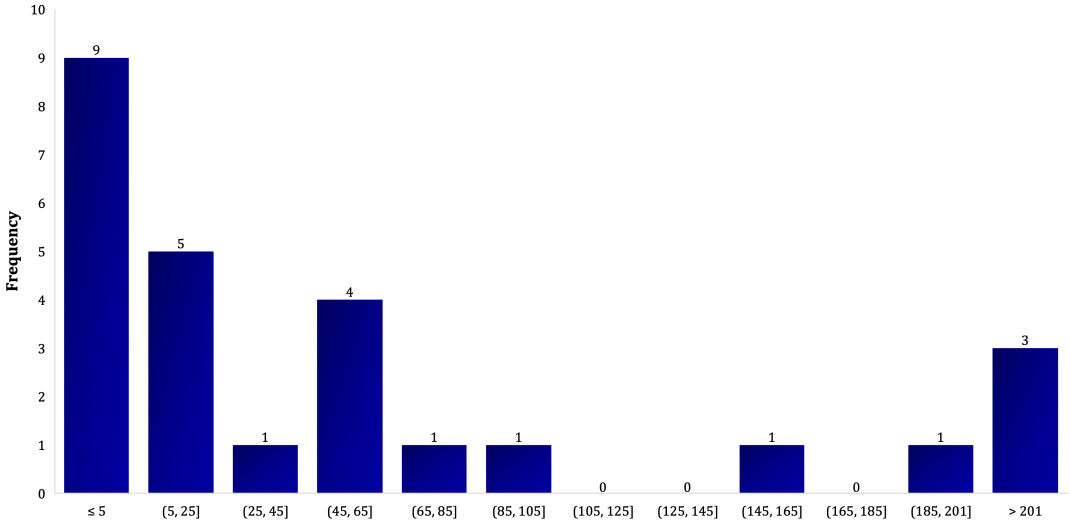

Figure 2: Histogram exhibiting the number of data files produced as part of the CER project (Survey Question Q5, n=26)

Q2 addressed the research formats used and had 52 respondents. Of these, 41 (79%) reported using unstructured text, while 24 (46%) used structured text. In addition, 29 respondents (55%) used general-purpose formats, while 7 (13%) worked with domain-specific formats. Also, 16 researchers (30%) used multimedia, 5 (9%) compiled binary artifacts, and 26 respondents (50%) said to use source code.

For Q4, 23 responses were gathered, which addressed the file size of the research data (see Figure 1) created and that may potentially be shared. By default, the minimum file size in the survey is zero. 500 was the maximum megabyte size, and 83 was the average size indicated by the respondents. For the gigabyte size answer option, 400 was the maximum, and 103 was the average. For the terabyte category, one dataset was reported with a volume of 3 TB and another with 1 TB.

The following survey question (Q5) asked for the number of data files produced as part of the last representative CER project. Here, 26 people responded to that question. Given the responses, 1 was the minimum, 7500 was the maximum number, and 400 files were identified as the average number. Figure 2 summarizes the number of data files produced by the respondents within the given ranges on the x-axis.

Q6 addressed the question of research data ownership, meaning who 'owns' the research data. We received 38 responses. The respondents were offered several reactions, as indicated in the list below. For each list item, multiple choices were allowed (e.g., before and after publication). The results are as follows, ranked from highest response rates to lowest:

- Myself: 23 (61%) before publication, 20 (53%) after publication
- Project Collaborator: 21 (55%) before publication, 19 (50%) after publication
- Institution: 20 (53%) before publication, 16 (42%) after publication
- Publisher: 3 (8%) before publication, 12 (32%) after publication

- Funder: 8 (21%) before publication, 7 (18%) after publication
- State/Government: 7 (18%) before publication, 6 (16%) after publication
- Don't know: 4 (11%) before publication, 3 (8%) after publication

The next question (Q7) asked if and how the research data that was used or created has been published. We gathered 39 responses, with multiple choices being allowed. The resulting options were selected as follows, with percents rounded to the nearest integer:

- Appendix to publication: 8 (20%)
- Stand-alone data publication: 5 (13%)
- Research data center: 3 (7%)
- Data repository by funder: 2 (5%)
- Data repository by a publisher: 1 (2%)
- Data repository by institution: 6 (15%)
- Software repository (e.g., GitHub): 10 (26%)
- Personal website: 2 (5%)
- Institutional website: 5 (13%)
- On another website: 1 (2%)
- Not published: 15 (38%)
- None of the above: 1 (2%)

Respondents also indicated the publication in the Snap!Cloud [143], the collection of projects for Snap!, and within institutional repositories.

The last survey question relevant to RQ4 was Q15, which asked for more details regarding the publication of the research data reported in this survey. Here, 37 responses were available. The results indicate that 15 (40%) respondents already published the data at a computing education conference, 15 (40%) plan to publish the data at a computing education conference, and 3 respondents (8%) do not believe the data fits within the scope of a computing education event. Other responses indicated that the researchers did not publish data for various reasons, such as prior rejections, data not in the scope of the conference, self-publication at webinars, publications in more data-centric venues, not being recognized in the field, and not being sure how to publish data.

7.2 Associated Barriers (RQ 5)

There are seven survey questions related to research question RQ 5 (Barriers), as indicated in Table 3. The related survey questions are Q8, Q9, Q10, Q11, Q12, Q13, and Q14. The following summarizes the outcome of these survey questions.

Thirty-seven respondents completed Q8, where we asked about the reasons for not publishing research datasets. Each respondent could select multiple reasons. The highest-ranked response was that the researchers did not feel that they had the obligation to publish their research data (n=21). The second most-voted barrier was that the researchers did not have time to publish the data (n=11), followed by the researchers being unsure how to make the data anonymous (n=10) and the research data not being documented sufficiently (n=9).

Question 9 was designed to elicit the limitations of specific barriers to publishing data. There were a total of 37 responses (see Figure 3), allowing users to select one or more of the following options:

- Legal concerns (e.g., ownership, privacy) (n=22, 59%)
- Resource constraints (n=16, 43%)
- Loss of control data (n=10, 27%)
- Authority or practice considerations (n=8, 22%)
- Technical constraints (n=6, 16%)
- None of the above (n=5, 13%)

Seven respondents provided more detailed explanations of their concerns. Four of these responses echoed the provided "legal concerns" option and expressed their concern for properly anonymizing their dataset. In addition to the possibilities, participants also offered specific situations that made publishing the dataset complex, such as prior agreements made regarding the dataset (*"promised to maintain confidentiality"*) or having multiple parties involved regarding the dataset (*"multi-institution project"*). One response also pointed out that *"There is no reward for overcoming the barriers"*.

Based upon the selections made in question 9, survey participants were presented with a series of follow-up questions (Q10-14) to understand the reasoning behind their choices. For brevity, we chose to analyze responses for the two most selected choices: "Legal concerns"(Q10) and "Resource constraints"(Q14). Although Q11, Q12, Q13, and Q15 also address reasonable barriers for researchers, it is a challenge to understand the reasoning behind the choices as we only received a low number of responses.

When respondents were asked about the legal concerns faced during the publishing of data (Q10), the most common answers included concerns about privacy and openness (n=16) and dissatisfaction with the anonymization process (n=13). Some respondents used the open question format to specify their concerns, submitting the following explanations:

> *"We are caretakers only of someone else's data. Institutional Review Boards typically note that we will destroy data one year after the project ends."*
> *"When I started collecting data, there was no ethical approval necessary, but things have changed quite a bit, so I am unsure what to do."*

Resource constraints were the second most common selection in Q9, and respondents were asked about these barriers in detail in Q14. The most common answer was that publishing research data requires too much work and effort (n=13). Additionally, some respondents specified this barrier in the open question format:

> *"I was rejected three times when trying to publish my data in the CER domain"*
> *"I do not trust companies like GitHub with things like my research data."*
> *"Time / reward tradeoff. Insufficient reward."*
> *"Prefer to be mentored by someone who has published material previously"*

7.3 Preferences and Expectations (RQ 6)

Four survey questions relate to research question RQ 6 (Preferences) regarding researchers' preferences and expectations, as indicated in Table 3. The related survey questions are Q3, Q16, Q17, and Q18. The following section summarizes the replies to each of these survey questions.

Q3 addressed the shared formats of research data, with 48 respondents. The choices with corresponding numbers and percentages are as follows, with percentages rounded to the nearest integer:

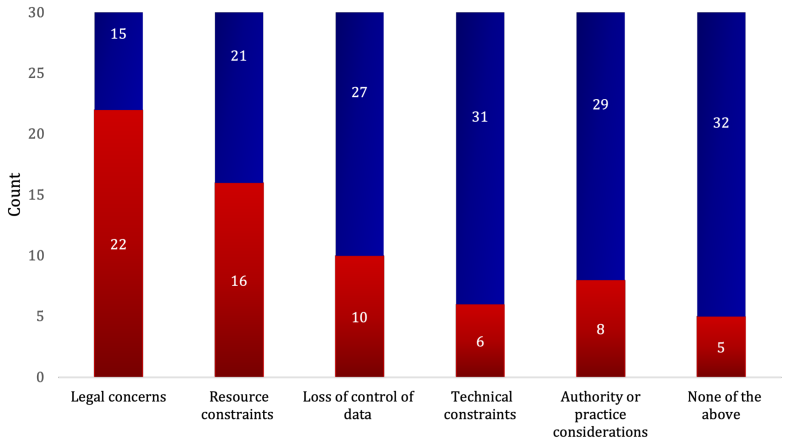

Figure 3: Barriers to publishing research data - Yes in red, No in blue (Survey Question Q9, n=37)

- Plain text: 26 (54%)
- Proprietary general-purpose formats: 25 (52%)
- Source Code: 22 (46%)
- Structured text: 19 (39%)
- Multimedia: 11 (23%)
- Compiled binary artifacts: 6 (12%)
- Proprietary domain-specific formats: 4 (8%)
- None of these: 1 (2%)

The respondents did not provide any other suggestions for the data formats they would share.

Q16 asked for preferred publication formats, such as various addenda to existing or separate data paper formats. It had 44 responses, and multiple selections were allowed. An addendum to existing paper formats was preferred by 66% respondents. For publishing software, 38% (n=14) believed that a 6-page explanation paper sufficed, and 30% (n=11) favored a 10-page format. More than 60% (n=23) favored the idea of a new format centered on data and explanations over the usual paper formats. When asked for suggestions beyond those offered, there were 4 responses. One suggested "Zenodo", one asked for a discussion with conference chairs, and one preferred just data and no papers. At the same time, a final respondent objected to publishing data due to privacy concerns.

The next question (Q17) asked whether the publication of research data should be mandatory upon accepting a paper. Here, 35 individuals responded, with 26 respondents selecting yes and 9 respondents selecting no.

Q18 asked for methods of conducting data reviews, which yielded 18 responses. 9 (50%) asked reviewers to check the completeness, usability, and quality of documentation and understandability of data. Here, 4 responses (25%) held that getting good data reviewers will be much more complicated than even ordinary reviewers, while 4 (25%) requested automated checks to anonymize data. One respondent asked merely that the existence of data should be verified, arguing that "too many promise but do not release".

8 Discussion of the Community's Perspective

In this section, the authors reflect upon and discuss how the community can interpret these results, thinking toward improved data practices and a path forward for the community. Accordingly, this section also addresses the survey's limitations.

8.1 Usage and Release Practices

RQ 4 addressed the practices that are followed by researchers and practitioners regarding data usage. The survey results showed that most researchers currently use their own qualitative and quantitative data. Regarding using the software for data, almost half of the respondents used their own software, and the other half preferred to use software written by others. Thus, attempts to make research data (including software) reusable could attract a target audience, as CE researchers seem to be open to that (see also [14]). The fact that respondents reported using their data primarily may also be due to the lack of FAIR research data in the community.

Moreover, the survey results indicated that most researchers used or created unstructured text data formats. This was followed by source code artifacts, structured text, general-purpose formats, multimedia data, domain-specific formats, and compiled binary artifacts. The survey also addressed file size; the most significant size was 3 terabytes. The vast majority were less than a gigabyte, which seems to be an easily manageable amount. The survey results also indicated that most respondents produced less than 50 files; some reached 200 files, while a few had several thousand files. Thus, managing and sharing these types and amounts of data should be reasonable.

Additionally, a significant majority of respondents believed that data is either owned by oneself or by collaborators before and after publication. On the one hand, researchers who believe that data is owned by themselves may be more inclined to publish data as they may be less concerned with legal issues. On the other hand, they may feel more assertive about their ownership, which could lead to less data donations. More research is required to identify a clear pattern and understanding of these practices and concerns.

Based on the survey findings for Q7, 46% of the researchers have selected to either add their data to an appendix in a publication or publish in a software repository (e.g., GitHub) as a preference. Both methods require low additional effort and are closely related to published research papers. The problem with this approach is that it does not follow the FAIR principles, making it difficult or even impossible to find or re-use the data. Moreover, only published data with PIDs can be cited. Interestingly, 38% of the researchers indicated they do not publish their data, making it impossible to find, reproduce the research results, or reuse the data (all addressed in Q7, multiple choice selection).

In addition, 34 of 37 researchers (91%) have either already published data or plan to publish the data at a computing education conference, and only 3 researchers do not believe the data fits within the scope of the CER community (addressed in Q15). There were various other valid responses for not publishing the data, such as data not within the scope of the conference, self-publication at webinars, published at data-centric venues, and insecurity on how to publish data. Nonetheless, most respondents seemed to have considered publishing their research data already.

8.2 Associated Barriers

Barriers and challenges associated with the publication of datasets were the focus of RQ 5. According to the survey responses to Q8 regarding the rationale for not publishing data, the publication process of research data is not too common in the CER community. Without external obligation, researchers seem to have little motivation to publish data. Part of the reasoning is the lack of recognition for data publication. Furthermore, there is a need for good research data management that helps avoid poor data documentation or even the loss of research data. Even if researchers want to publish their data, they face barriers, especially with the expected de-identification or anonymization process (i.e., what is it, how is it applied correctly, etc.). Thus, it seems crucial to train young researchers to realize this process.

Another interesting aspect was conveyed through the responses to Q9, focusing on the barriers to publishing data. There were 37 responses, and more than 60% indicated that significant publishing barriers are legal concerns and resource constraints. The results suggest that participants were unsure of the data-sharing process or making data publicly available. This situation should prompt the CER community to develop a framework for sharing data in a uniform format. Additionally, Institutional Review Boards (IRB) often create complications when sharing data with the community freely, so researchers should consider their options when applying for IRB approval. Another path forward may be to support replication studies. Prior work has shown that computing education researchers hold biases against replications and favor novel work [3].

8.3 Preferences and Expectations

The expectations and preferences of the surveyed researchers regarding published CER datasets were subject to RQ 6. As far as the format of research data, the authors observed that the top three preferences among researchers were (1) plain text, (2) proprietary general-purpose format, and (3) source code. Most people liked the idea of having a new track for "data paper" of approximately six pages, double column. People had divided opinions on having a tools or software publication track. In addition, most respondents (26 of 35) agreed to the mandatory publication of the research data upon acceptance of a paper. Therefore, it seems reasonable to adopt respective policies and tracks as part of upcoming community conferences.

The survey further asked for suggestions regarding ways to review datasets. Most people want the reviewers to focus on data comprehensibility. Some respondents believe that data review is impossible due to the shortage of qualified reviewers, with one outlier thinking that data should never be published for privacy reasons. The last comment is surprising since the CER community should encourage reproducible research with efforts to reduce the difficulty of responsible publishing. Of course, sensitive data should be protected and adequately anonymized. Hence, there is not necessarily a conflict as long as sensitive data is appropriately treated.

8.4 Limitations of the Survey

The authors circulated the survey in the CER community via the SIGCSE listserv and other mailing lists, as well as author contacts within the CER community. However, some of the following limitations are inherent. One of them concerned sampling. The target audience for this research was CER researchers and educators. Since the authors only surveyed the CER community as potential generators of such data sets, they did not reach individuals who are researchers in other fields who may have data but only occasionally publish it in CER conferences or journals. Moreover, we only had 76 respondents for the survey, which is a relatively small sample and likely due to the online distribution where low response rates are typical. It should further be noted that the response rates vary among questions.

Another limitation is the survey format, which did not allow queries to go in-depth with individual respondents. Thus, the pre-structured survey may not have fully captured all concerns within the CER community. In addition, limitations of self-reporting apply to the data-gathering method.

9 Recommendations for the CER Community

We must share data to ensure our research is *reproducible* and *cumulative*. This lets us improve the progress of CER as a discipline as we build our work on top of valid building blocks constructed by others in prior work.

To ensure these properties, the computing education community needs to introduce policies and build incentives within the CER ecosystem that are conducive to sharing research data. Based on what the authors found in this research on existing data and the conducted survey, the following changes and measures are proposed for various participants in the study and data ecosystem. Even though the authors address the stakeholders separately, it should be noted that only an interplay of all actors can ensure the CER community's movement towards open data practices.

9.1 Institutions

Research institutions and universities are crucial players in fostering Open Data practices. As such, they should establish openness

and transparency, comprising the following structures and support measures:

- Open Data or Open Science policies for researchers.
- Best practice examples and information on their websites.
- Guidance and support in the form of seminars and workshops.
- Data Steward position(s) and role models to consult individual requests.
- Reward data publications, e.g., via blog/news entries, badges, bonuses.
- Establish networking opportunities among researchers, e.g., formal or informal events on data publications.
- Provide the necessary infrastructure for conducting efficient research data management.

9.2 Institutional Review Boards (IRB)

A central concern for many researchers is the legal and ethical issues surrounding data sharing in the sense of publications. For this reason, we recommend that IRBs of organizations, in particular, ensure that there are

- Clear criteria regarding data anonymization, sharing, and the entire data management lifecycle.
- Accepted methods (and models) for data anonymization, ethical, and legal practices along the whole research data life-cycle.

To make these changes, IRB reviews for CER can use the experience of the Psychology community, for example, which has dealt with a replication crisis [123] already.

Another model we suggest is that of open licenses, including open-source licenses [128, 160]. We can, for example, develop data sharing "certifications" with various levels of sharing along the lines of Creative Commons licenses [40].

9.3 Publishers, Journals, and Conferences

Conferences and journals should incentivize data publication by

- Encouraging or requiring data sharing.
- Introducing new data-only tracks or tools and data tracks.
- Creating awards or badges for research data publication and quality.

Moreover, they should value and accept replication studies, not just novel work on new phenomena [3].

There are communities such as "Learning at Scale" [154] and "Software Engineering" who have taken steps in this direction, meaning CER can use these as starting points.

Publishers should also define minimal review criteria that make a dataset acceptable for publication and reuseable. For example,

- Data should be documented, have adequate metadata, and not be corrupted or otherwise unusable.
- As far as possible, use open-source tools to maintain data.

The last recommendation is that proprietary tools may render data inaccessible to many researchers. As tools and software undergo version changes, data may even become unusable. When proprietary software becomes necessary, it is strongly recommended that the proprietary format be complemented with an export to an open format whenever feasible. Unfortunately, issues can arise in version updates of open-source software as well. We thus note that using open-source tools also does not guarantee that data will be usable.

Additionally, if viable, preserving the software itself is advised. Comprehensive documentation should encompass all specifics, including the software version, ensuring thorough record-keeping.

9.4 Data Repository and Search Builders

It is essential that data sharing is available to the worldwide research community. Therefore, data repositories should be *open access* as far as possible, and it should not be prohibitively expensive to publish data. Platforms such as Zenodo provide a model we can emulate.

Furthermore, even in cases where data is not easy to share, their metadata should be open so that the researcher(s) can at least be contacted.

In addition to storing data, it is important to ensure that data is easy to discover and comprehend. We recommend the following features for any data repository:

- DOI or other identifier schemes.
- A high-quality, easily understandable minimal metadata and provenance scheme (cf. Appendix B).
- Support CER specific controlled vocabulary, ontologies ([26, 111, 155]), and standard data formats (e.g., ProgSnap2 [137]).
- Licensing and Rights management.
- Usability for different users such as students, teachers, researchers, data creators, and repository maintainers.

To support researchers, entering metadata should be as easy as possible [98, 102]. This encompasses, for example, definitions of the metadata entities, example vocabulary, metadata suggestions (generated by people or AI [39]), and an adequate scope of metadata. As the process of assigning metadata can take a lot of time, it is important to focus on the minimum requirements (mandatory fields) at this point. If possible, however, as much information as possible should be provided.

In addition, repository builders should support search and filter options using metadata to promote discoverability and findability.

9.5 Researchers and Educators

In addition to institutional incentives, it should be an expectation among researchers that

- They submit research proposals that incorporate resources (time and money) for proper research data management and publication.
- They use checklists to ensure appropriate (meta-) data collection and sharing practices early on in their projects, including IRB applications. We have designed a respective checklist to support researchers in collecting, managing, and releasing datasets (see Appendix C).
- They maintain data, including software, and try to share as much as possible.
- They properly cite data, including software that they re-use (and not only the related articles).
- They give and receive training in de-identifying or anonymizing their research data before publication [96].

It is essential to emphasize that data should be made as open as possible and as closed as necessary. Should compelling justifications exist for restricting access to the data and only disclosing metadata (e.g., gathering data from minors), this course of action can also be

considered a viable and responsible approach regarding the FAIR principles.

Those who train or supervise early career researchers should develop training materials to ensure that student researchers learn standard ways to conduct reproducible research based on published data [10, 13, 50]. They should further act as role models [67] to students when working with research data. The responsible use of data and data ethics should also be integrated into computing curricula [94, 103].

To support the computing education community, this working group developed a checklist for publishing research data within the CER context while considering other generic aspects. The checklist is available in Appendix C and can help guide researchers through the data documentation process, quality assurance, legal issues, and the selection of appropriate data infrastructure.

9.6 Industry

There is broad agreement that the needs of industry and the goals of academia should be aligned whenever feasible. To encourage such alignment, the industry should participate in supporting computing education research as follows.

- Work with the CER community to set up experiments and share data about students' performance in industrial tasks and skills.
- Support faculty and industry in updating curricula.

10 Conclusions

In this working group paper, the authors have presented a search for openly available and published datasets in computing education research. The goal was to develop a resource for CER researchers and educators interested in secondary data analysis or using data in the computing classroom. Therefore, we searched through several resources (Appendix A), identified publicly available datasets, and characterized them by applying a newly developed metadata scheme applicable to different types of collected data in the computing education context [93]. Moreover, in this report, we included other datasets relevant to computing educators and classroom settings, as well as from educational data mining, K-12 contexts, and industry contexts.

The second part of this report comprised an online survey and analysis of perceptions and efforts within the computing education research community. In particular, the goal was to understand their perception and concerns about sharing data and open data practices. Based on the survey responses, the authors presented recommendations to the CER community and analyzed computing education researchers' motivations, concerns, challenges, and ideas for future conferences and data publications. Most importantly, we identified a general openness towards the publication of research data, specifically new data tracks" as part of CER conferences. At the same time, this requires the development of adequate reviewing guidelines, which currently do not exist.

Based on the analyzed survey responses, the authors presented recommendations to the CER community. These recommendations address all stakeholders in the CER ecosystem, including research institutions and universities, institutional review boards, publishers, conferences, journals, data repositories and search builders, researchers and educators, and industry. Even though many of these recommendations may be perceived as generic, they are critical for open data practices in CER. Data policies and infrastructure are required for researchers to donate their data. Moreover, it is essential to continue working on data formats, standards, and metadata schema to ease the intersubjective understanding of research data. To support this development, we not only suggest a metadata schema for the CER context (Appendix B) but also a checklist for the publication of research data (Appendix C).

11 Future Work

There is a clear interest and endeavor in collecting and analyzing data in the computing education research community to gain new insights and increase the reuse of data for secondary research purposes. As discussed earlier, there are several challenges when working with data in general and thus there is a need to continue to address these challenges. This working group identified the following opportunities for future work.

Standardizing data collection and publication for better analysis and easier reuse. We noticed there is a significant disparity in how and what data is collected today. For example, the dataset available for CS1 programming courses at University X is so different from the data collected at University Y, that it is impossible to derive any conclusions and gain meaningful insights across institutions and countries. The lack of a common language (e.g., metadata) to describe the datasets contributes to the uncertainties of researchers aiming to conduct secondary research. In future work, we encourage educators and researchers to collect data in a more standardized way by advancing toward standards for harvesting and describing data and metadata.

Develop a central repository or infrastructure for CER artifacts and respective datasets. Another follow-up project this group encourages is the development and hosting of a repository for the CER community to publish and find respective research data.

Extend the search for data to other areas of computing education research. Due to the focus of this working group on certain areas of computing education (e.g., introductory programming, benchmarking), there is potential for future work on finding datasets from other areas by using other search terms and an adapted strategy.

Expand and evaluate the proposed metadata scheme. In this report, we presented a newly developed metadata scheme to describe the research data within computing education research contexts. This working group encourages the community to apply, evaluate, and expand upon this scheme as necessary to help other researchers describe their own research data, and, in turn, ease the understanding of research data published by others. We are aware that the proposed scheme is a minimal set, and that expansions may be feasible. Moreover, we intend to further verify the metadata with the support of the authors.

Finally, we want to encourage all of our fellow CER researchers to start thinking about ways to share their data in alignment with the FAIR principles, so that others can build upon prior work instead of reinventing the wheel all over – again and again.

Acknowledgments

This work builds on the efforts of previous projects. The authors wish to acknowledge the National Science Foundation of the United

States under Grant Numbers 2111435, 2111097, 2213792, and 1923597. They also want to acknowledge the support of the CCRI funding "An Infrastructure for Sustainable Innovation and Research in Computer Science Education", the HEADT Centre, and the Female Promotion of the Computer Science Department of the Humboldt-Universität zu Berlin. Moreover, we greatly appreciate the support of the DIPF Leibniz Institute for Research and Information in Education during this project.

References

[1] ACM. 2023. ACM Digital Library. https://dl.acm.org/

[2] David Aha and University of California. 2023. UC Irvine Machine Learning Repository. http://archive.ics.uci.edu/

[3] Alireza Ahadi, Arto Hellas, Petri Ihantola, Ari Korhonen, and Andrew Petersen. 2016. Replication in computing education research: researcher attitudes and experiences. In *Proceedings of the 16th Koli Calling International Conference on Computing Education Research* (Koli, Finland) *(Koli Calling)*. Association for Computing Machinery, New York, NY, USA, 2–11. https://doi.org/10.1145/2999541.2999554

[4] Carlos Alario-Hoyos. 2021. *Dataset MOOC Forum edX*. Universidad Carlos III de Madrid. https://doi.org/10.5281/zenodo.5115573

[5] Farid Anvari and Daniël Lakens. 2018. The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology* 3, 3 (2018), 266–286.

[6] Mikko Apiola, Sonsoles López-Pernas, and Mohammed Saqr. 2023. *Past, Present and Future of Computing Education Research: A Global Perspective*. Springer Nature, Cham. https://doi.org/10.1007/978-3-031-25336-2

[7] FAIR Aware. 2021. FAIRaware. https://fairaware.dans.knaw.nl/ Last access: 2023-10-08.

[8] RSJD Baker et al. 2010. Data mining for education. *International encyclopedia of education* 7, 3 (2010), 112–118.

[9] Austin Cory Bart, Ryan Whitcomb, Jason Riddle, Omar Saleem, Eli Tilevich, Clifford A. Shaffer, and Dennis Kafura. 2023. CORGIS. The Collection of Really Great, Interesting, Situated Datasets. https://corgis-edu.github.io/corgis/

[10] Katarzyna Biernacka, Ron Dockhorn, Claudia Engelhardt, Kerstin Helbig, Juliane Jacob, Tereza Kalová, Adienne Karsten, Kristin Meier, Andreas Mühlichen, Janna Neumann, Britta Petersen, Benjamin Slowig, Ute Trautwein-Bruns, Jeanne Wilbrandt, and Cord Wiljes. 2023. *Train-the-Trainer-Konzept zum Thema Forschungsdatenmanagement*. Zenodo. https://doi.org/10.5281/zenodo.10122153

[11] 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.

[12] 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*. IGI Global, Hershey, PA, 29–60.

[13] Katarzyna Biernacka and Sandra Schulz. 2022. *Forschungsdatenmanagement in der Informatik*. Logos Verlag. https://doi.org/10.30819/5490

[14] Jeremiah Blanchard, John R. Hott, Vincent Berry, Rebecca Carroll, Bob Edmison, Richard Glassey, Oscar Karnalim, Brian Plancher, and Seán Russell. 2022. Stop Reinventing the Wheel! Promoting Community Software in Computing Education. In *Proceedings of the 2022 Working Group Reports on Innovation and Technology in Computer Science Education* (, Dublin, Ireland,) *(ITiCSE-WGR '22)*. Association for Computing Machinery, New York, NY, USA, 261–292. https://doi.org/10.1145/3571785.3574129

[15] Christine L. Borgman and Irene V. Pasquetto. 2017. Why Data Sharing and Reuse Are Hard To Do. https://escholarship.org/uc/item/0jj17309

[16] Neil C. C. Brown, Amjad Altadmri, Sue Sentance, and Michael Kölling. 2018. Blackbox, Five Years On: An Evaluation of a Large-Scale Programming Data Collection Project. In *Proceedings of the 2018 ACM Conference on International Computing Education Research* (Espoo, Finland) *(ICER '18)*. ACM, New York, 196–204. https://doi.org/10.1145/3230977.3230991

[17] Neil C. C. Brown and Mark Guzdial. 2024. Confidence vs Insight: Big and Rich Data in Computing Education Research. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (Portland, OR, USA) *(SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 158–164. https://doi.org/10.1145/3626252.3630813

[18] Neil C. C. Brown and Mark Guzdial. 2024. Confidence vs Insight: Big and Rich Data in Computing Education Research. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (, Portland, OR, USA,) *(SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 158–164. https://doi.org/10.1145/3626252.3630813

[19] 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)*. ACM, New York, 000000–000000. https://doi.org/10.1145/2538862.2538924

[20] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[21] Peter Brusilovsky, Ken Koedinger, David A. Joyner, and Thomas W. Price. 2020. Building an Infrastructure for Computer Science Education Research and Practice at Scale. In *Proceedings of the Seventh ACM Conference on Learning @ Scale* (Virtual Event, USA) *(L@S '20)*. Association for Computing Machinery, New York, NY, USA, 211–213. https://doi.org/10.1145/3386527.3405936

[22] Canadian Institute for Cybersecurity (CIC). 2023. CIC Datasets. https://www.unb.ca/cic/datasets/index.html.

[23] Canadian Institute for Cybersecurity (CIC). 2023. University of New Brunswick. https://www.unb.ca/cic/datasets/index.html.

[24] Carnegie Mellon University. 2023. Datashop@CMU. https://pslcdatashop.web.cmu.edu

[25] Arturo Casadevall and Ferric C. Fang. 2010. Reproducible Science. *Infection and Immunity* 78, 12 (2010), 4972–4975. https://doi.org/10.1128/IAI.00908-10 arXiv:https://journals.asm.org/doi/pdf/10.1128/IAI.00908-10

[26] Lillian N. Cassel, Gordon Davies, William Fone, Anneke Hacquebard, John Impagliazzo, Richard LeBlanc, Joyce Currie Little, Andrew McGettrick, and Michela Pedrona. 2007. The Computing Ontology: Application in Education. In *Working Group Reports on ITiCSE on Innovation and Technology in Computer Science Education* (Dundee, Scotland) *(ITiCSE-WGR '07)*. Association for Computing Machinery, New York, 171–183. https://doi.org/10.1145/1345443.1345439

[27] Center for open Science. 2023. Open Sciene Framework. https://osf.io/

[28] Centre for Science and Technology Studies, Elsevier and Leiden University. 2017. *Open Data. The researcher perspective*. Technical Report. Centre for Science and Technology Studies, Elsevier and Leiden University. https://www.elsevier.com/open-science/research-data/open-data-report

[29] CERN. 2023. Zenodo. https://zenodo.org/

[30] Lee Chaw. 2022. *Dataset related to MOOCs*. UCSI University. https://doi.org/10.17632/v398vj34h6.1 V1.

[31] Juan Chen, Sheikh Ghafoor, and John Impagliazzo. 2022. Producing competent HPC graduates. *Commun. ACM* 65, 12 (2022), 56–65.

[32] Victoria Clarke and Virginia Braun. 2014. *Thematic Analysis*. Springer New York, New York, NY, 1947–1952. https://doi.org/10.1007/978-1-4614-5583-7_311

[33] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. 2020. Threats of a Replication Crisis in Empirical Computer Science. *Commun. ACM* 63, 8 (jul 2020), 70–79. https://doi.org/10.1145/3360311

[34] Code.Org. 2023. Code.org's Annual State of Computer Science Education Report. https://code.org.

[35] Code.Org. 2023. Code.org's Annual State of Computer Science Education Report. https://code.org/research.

[36] European Commission, Directorate-General for Research, and Innovation. 2018. *Cost-benefit analysis for FAIR research data – Cost of not having FAIR research data*. Publications Office. https://doi.org/10.2777/02999

[37] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. http://www.loc.gov/standards/premis

[38] Computing Technology Industry Association (CompTIA). 2023. Tech Industry Workforce Report. https://www.comptia.org/

[39] Edward M Corrado. 2021. Artificial intelligence: The possibilities for metadata creation. *Technical Services Quarterly* 38, 4 (2021), 395–405.

[40] Creative Commons Corporation. 2023. Creative Commons. https://creativecommons.org

[41] DataCite Metadata Working Group. 2021. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs v4.4. https://doi.org/10.14454/3W3Z-SA82

[42] Carnegie Mellon University DataLab. 2023. What is Educational Data Mining (EDM)? https://www.cmu.edu/datalab/getting-started/what-is-edm.html Last access: 2023-11-10.

[43] Anusuriya Devaraju, Robert Huber, Mustapha Mokrane, Patricia Herterich, Linas Cepinskas, Jerry de Vries, Herve L'Hours, Joy Davidson, and Angus White. 2022. FAIRsFAIR Data Object Assessment Metrics. https://doi.org/10.5281/zenodo.6461229

[44] Hendrik Drachsler and Wolfgang Greller. 2016. Privacy and analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM, New York, 89–98. https://doi.org/10.1145/2883851.2883893

[45] Dublin Core™ Metadata Initiative (DCMI). 2023. DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

[46] Florian Echtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395

[47] John Edwards. 2022. *2021 CS1 Keystroke Data*. Utah State University. https://doi.org/10.7910/DVN/BVOF7S

[48] John Edwards, Kaden Hart, and Raj Shrestha. 2023. Review of CSEDM Data and Introduction of Two Public CS1 Keystroke Datasets. *Journal of Educational Data Mining* 15, 1 (March 2023), 1–31. https://doi.org/10.5281/zenodo.7646659

[49] Elsevier. 2023. Mendeley Data. https://data.mendeley.com/

[50] Claudia Engelhardt, Raisa Barthauer, Katarzyna Biernacka, Aoife Coffey, Ronald Cornet, Alina Danciu, Yuri Demchenko, Stephen Downes, Christopher Erdmann, Federica Garbuglia, Kerstin Germer, Kerstin Helbig, et al. 2022. *How to be FAIR with your data.* Universitätsverlag Göttingen, Göttingen. https://doi.org/10.17875/gup2022-1915

[51] European Union (EU). 2023. European Data. https://data.europa.eu/en/

[52] European Union. 2023. European Open Science Cloud. https://eosc-portal.eu/

[53] Expanding Computing Education Pathways (ECEP) Alliance. 2023. National Data Resources. https://ecepalliance.org/cs-data/national-data-resources/

[54] FAIRsFAIR. 2023. FAIRsFAIR Research Data Object Assessment Service. https://www.f-uji.net/ Last access: 2023-10-20.

[55] R. A. Fisher. 1988. Iris. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56C76.

[56] National Center for Education Statistics (NCES). 2023. Integrated Postsecondary Education Data System (IPEDS). https://nces.ed.gov/ipeds/

[57] Erin D Foster and Ariel Deardorff. 2017. Open science framework (OSF). *Journal of the Medical Library Association: JMLA* 105, 2 (2017), 203.

[58] Dolores Frias-Navarro, Juan Pascual-Llobell, Marcos Pascual-Soler, Jose Perezgonzalez, and Jose Berrios-Riquelme. 2020. Replication crisis or an opportunity to improve scientific production? *European Journal of Education* 55, 4 (2020), 618–631.

[59] Ge Gao, Samiha Marwan, and Thomas W Price. 2021. Early performance prediction using interpretable patterns in programming process data. In *Proceedings of the 52nd ACM technical symposium on computer science education.* ACM, New York, 342–348.

[60] German Centre for Higher Education Research and Science Studies (DZHW). 2022. Find Higher Education and Science Research Data Packages. https://metadata.fdz.dzhw.eu/en/start

[61] German Research Foundation (DFG). 2023. Handling of Research Data. https://www.dfg.de/en/research_funding/principles_dfg_funding/research_data/index.html

[62] GitHub. 2023. GitHub Archive. https://github.com/datasets.

[63] Jeremy Goecks, Anton Nekrutenko, and James Taylor. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* 11, 8 (2010), 1–13.

[64] Goeller, Sandra and Soltau, Kerstin. 2022. Dokumentation RADAR Metadatenschema. https://radar.products.fiz-karlsruhe.de/sites/default/files/radar/docs/info/RADAR_Metadaten_Dokumentation_v9.1.pdf

[65] Alejandra González-Beltrán, Peter Li, Jun Zhao, Maria Susana Avila-Garcia, Marco Roos, Mark Thompson, Eelke van der Horst, Rajaram Kaliyaperumal, Ruibang Luo, Tin-Lap Lee, et al. 2015. From peer-reviewed to peer-reproduced in scholarly publishing: the complementary roles of data models and workflows in bioinformatics. *PLOS one* 10, 7 (2015), e0127612.

[66] Google. 2023. Dataset Search. https://datasetsearch.research.google.com/

[67] Virginia Grande, Päivi Kinnunen, Anne-Kathrin Peters, Matthew Barr, Åsa Cajander, Mats Daniels, Amari N. Lewis, Mihaela Sabin, Matilde Sánchez-Peña, and Neena Thota. 2022. Role Modeling as a Computing Educator in Higher Education: A Focus on Care, Emotions and Professional Competencies. In *Proceedings of the 2022 Working Group Reports on Innovation and Technology in Computer Science Education* (Dublin, Ireland) *(ITiCSE-WGR '22)*. ACM, New York, 37–63. https://doi.org/10.1145/3571785.3574122

[68] Wouter Groeneveld, Brett A. Becker, and Joost Vennekens. 2021. *How Creatively Are We Teaching and Assessing Creativity in Computing Education: A Systematic Literature Review.* https://doi.org/10.5281/zenodo.5752559

[69] Rahul Gupta, Soham Pal, Aditya Kanade, and Shirish Shevade. 2017. DeepFix: Fixing Common C Language Errors by Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 31, 1 (Feb. 2017), 7 pages. https://doi.org/10.1609/aaai.v31i1.10742

[70] Qiang Hao, David H. Smith IV, Naitra Iriumi, Michail Tsikerdekis, and Amy J. Ko. 2019. A Systematic Investigation of Replications in Computing Education Research. *ACM Trans. Comput. Educ.* 19, 4, Article 42 (aug 2019), 18 pages. https://doi.org/10.1145/3345328

[71] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. https://doi.org/10.1145/2827872

[72] Harvard University Library. 2023. Dataverse Project. https://dataverse.org/

[73] David Hovemeyer, Arto Hellas, Andrew Petersen, and Jaime Spacco. 2017. Progsnap: Sharing Programming Snapshots for Research (Abstract Only). In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education* (Seattle, Washington, USA) *(SIGCSE '17)*. Association for Computing Machinery, New York, NY, USA, 709. https://doi.org/10.1145/3017680.3022418

[74] Jeff Huang. 2022. Computer Science Open Data — jeffhuang.com. https://jeffhuang.com/computer-science-open-data/. [Accessed 22-11-2023].

[75] Hugging Face. 2023. Hugging Face. https://huggingface.co/

[76] IEEE. 2023. IEEE DataPort: Dataset Storage and Dataset Search Platform. https://ieee-dataport.org/

[77] IEEE. 2023. IEEE Xplore. https://ieeexplore.ieee.org/

[78] IEEE Computer Society. 2020. IEEE Standard for Learning Object Metadata. , 50 pages. https://doi.org/10.1109/IEEESTD.2020.9262118

[79] 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.

[80] Darrel Ince. 2011. The Duke University scandal - what can be done? *Significance* 8, 3 (Aug. 2011), 113–115. https://doi.org/10.1111/j.1740-9713.2011.00505.x

[81] The NLP Index. 2023. The NLP Index. https://index.quantumstat.com/

[82] International Association for the Evaluation of Educational Achievement (IEA). 2023. International Computer and Information Literacy Study (ICILS) 2018 Dataset. https://www.iea.nl/studies/iea/icils-2018.

[83] Matti Järvisalo, Daniel Le Berre, Olivier Roussel, and Laurent Simon. 2012. The international SAT solver competitions. *Ai Magazine* 33, 1 (2012), 89–92.

[84] Johan Jeuring, Hieke Keuning, Samiha Marwan, Dennis Bouvier, Cruz Izu, Natalie Kiesler, Teemu Lehtinen, Dominic Lohr, Andrew Petersen, and Sami Sarsa. 2022. Steps Learners Take When Solving Programming Tasks, and How Learning Environments (Should) Respond to Them. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 2* (Dublin, Ireland) *(ITiCSE '22)*. ACM, New York, 570–571. https://doi.org/10.1145/3502717.3532168

[85] 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

[86] Kaggle. 2023. Kaggle. https://www.kaggle.com/

[87] Daniel S. Katz, Morane Gruenpeter, and Tom Honeyman. 2021. Taking a fresh look at FAIR for research software. *Patterns* 2, 3 (2021), 100222. https://doi.org/10.1016/j.patter.2021.100222

[88] Daniel S. Katz, Fotis E. Psomopoulos, and Leyla Jael Castro. 2021. Working Towards Understanding the Role of FAIR for Machine Learning. In *2nd Workshop on Data and Research Objects Management for Linked Open Science.* ZB MED-Publikationsportal Lebenswissenschaften, Online, 1–6. https://doi.org/10.4126/FRL01-006429415

[89] Hieke Keuning. 2024. The interplay between rich and big data in programming education research. In *22. Fachtagung Bildungstechnologien (DELFI)*, Sandra Schulz and Natalie Kiesler (Eds.). Gesellschaft für Informatik e.V., Bonn, 19–21. https://doi.org/10.18420/delfi2024_01

[90] 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.

[91] Natalie Kiesler. 2022. *Daten- und Methodenbericht Rekursive Problemlösung in der Online Lernumgebung CodingBat durch Informatik-Studierende.* Technical Report. DZHW. https://metadata.fdz.dzhw.eu/public/files/data-packages/stu-studentsteps$/attachments/studentsteps_Data_Methods_Report_de.pdf

[92] 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* (Hyderabad, India) *(CompEd 2023)*. Association for Computing Machinery, New York, NY, USA, 209–210. https://doi.org/10.1145/3617650.3624951

[93] Natalie Kiesler, John Impagliazzo, Katarzyna Biernacka, Amanpreet Kapoor, Zain Kazmi, Sujeeth G Ramagoni, Aamod Sane, Keith Tran, Shubbi Taneja, and Zihan Wu. 2024. CompEd Working Group 2023 - Supplementary Material. https://doi.org/10.17605/OSF.IO/R83S5

[94] Natalie Kiesler, Simone Opel, and Carsten Thorbrügge. 2024. With Great Power Comes Great Responsibility: Integrating Data Ethics into Computing Education. In *Proceedings of the 2024 Conference on Innovation and Technology in Computer Science Education V. 2* (Milan, Italy) *(ITiCSE 2024)*. ACM, New York. https://doi.org/10.1145/3649217.3653637

[95] 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

[96] Natalie Kiesler, René Röpke, Daniel Schiffner, Sandra Schulz, Sven Strickroth, Matthias Ehlenz, Birte Heinemann, and Arno Wilhelm-Weidner. 2024. Towards Open Science at the DELFI Conference. In *22. Fachtagung Bildungstechnologien (DELFI)*, Sandra Schulz and Natalie Kiesler (Eds.). Gesellschaft für Informatik e.V., Bonn, 251–265. https://doi.org/10.18420/delfi2024_22

[97] 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

[98] Natalie Kiesler and Daniel Schiffner. 2023. Exploring and Improving Workflows for the Donation and Curation of Research Data. In *1st Conference on Research Data Infrastructure - Connecting Communities, CoRDI 2023, Karlsruhe, Germany, September 12-14, 2023*, York Sure-Vetter and Carole A. Goble (Eds.). TIB Open Publishing, Karlsruhe (Germany), 1–4. https://doi.org/10.52825/CORDI.V1I.284

[99] Natalie Kiesler and Daniel Schiffner. 2023. Open Science in den Bildungstechnologien: Zur Publikation und Begutachtung von Forschungsdaten inklusive Software im Rahmen der DELFI. In *Workshops der 21. Fachtagung Bildungstechnologien (DELFI)*. Gesellschaft für Informatik e.V., Bonn, 159–168. https://doi.org/10.18420/wsdelfi2023-40

[100] 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, 348–353. https://doi.org/10.1145/3587102.3588860

[101] Natalie Kiesler and Daniel Schiffner. 2024. Conferences are Exclusive by Nature. In *Proceedings of the 2024 RESPECT Annual Conference* (Atlanta, GA, USA) *(RESPECT '24)*. ACM, New York, 5 pages. https://doi.org/10.1145/3653666.3656077

[102] Natalie Kiesler, Daniel Schiffner, and Axel Nieder-Vahrenholz. 2023. Adapting RDMO for the Efficient Management of Educational Research Data. In *DELFI 2023, Die 21. Fachtagung Bildungstechnologien der Gesellschaft für Informatik e.V., 11.-13. September 2023, Aachen (LNI, Vol. P-338)*, René Röpke and Ulrik Schroeder (Eds.). Gesellschaft für Informatik e.V., Bonn, 271–272. https://doi.org/10.18420/DELFI2023-51

[103] Natalie Kiesler and Carsten Thorbrügge. 2023. Socially Responsible Programming in Computing Education and Expectations in the Profession. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023)*. ACM, New York, 443–449. https://doi.org/10.1145/3587102.3588839

[104] Kenneth J Knapp, Christopher Maurer, and Miloslava Plachkinova. 2017. Maintaining a cybersecurity curriculum: Professional certifications as valuable guidance. *Journal of Information Systems Education* 28, 2 (2017), 101.

[105] Michael Kölling, Bruce Quig, Andrew Patterson, and John Rosenberg. 2003. The BlueJ system and its pedagogy. *Computer Science Education* 13, 4 (2003), 249–268.

[106] 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

[107] Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy S Liang. 2019. SPoC: Search-based Pseudocode to Code. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., Red Hook, NY, USA. https://proceedings.neurips.cc/paper_files/paper/2019/file/7298332f04ac004a0ca44cc69ecf6f6b-Paper.pdf

[108] Jakub Kuzilek, Martin Hlosta, and Zdenek Zdrahal. 2017. Open University Learning Analytics dataset. https://doi.org/10.1038/sdata.2017.171

[109] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[110] Carnegie Mellon University Libraries. 2023. Carnegie Mellon University Libraries. https://guides.library.cmu.edu/az.php

[111] Thong Chee Ling, Yusmadi Yah Jusoh, Rusli Adbullah, and Nor Hayati Alwi. 2013. An Ontology for Software Engineering Education.

[112] LinkedIn. 2023. LinkedIn Economic Graph. https://economicgraph.linkedin.com/

[113] Monica M. McGill. 2019. Discovering Empirically-Based Best Practices in Computing Education Through Replication, Reproducibility, and Meta-Analysis Studies. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (Koli, Finland) *(Koli Calling '19)*. Association for Computing Machinery, New York, NY, USA, Article 7, 5 pages. https://doi.org/10.1145/3364510.3364528

[114] MDPI. 2023. MDPI Publisher of Open Access Journals. https://www.mdpi.com/

[115] Meta. 2023. The latest on Machine Learning | Papers with Code. https://paperswithcode.com/

[116] Metadata Standards Catalog. 2023. Index of subjects. https://rdamsc.bath.ac.uk/subject-index

[117] Barend Mons, Herman van Haagen, Christine Chichester, Peter-Bram't Hoen, Johan T den Dunnen, Gertjan van Ommen, Erik van Mulligen, Bharat Singh, Rob Hooft, Marco Roos, et al. 2011. The value of data. *Nature genetics* 43, 4 (2011), 281–283.

[118] N.A. 2019. Degrees in computer and information sciences conferred by postsecondary institutions, by level of degree and sex of student: 1970-71 through 2017-18 — nces.ed.gov. https://nces.ed.gov/programs/digest/d19/tables/dt19_325.35.asp. [Accessed 22-11-2023].

[119] National Center for Education Statistics (NCES). 2023. National Center for Education Statistics (NCES) Datasets. https://nces.ed.gov/datalab/index.aspx.

[120] National Science Foundation. 2023. Open Data at NSF. https://www.nsf.gov/data/

[121] n.d. 2023. Learning engineering. https://groups.google.com/g/learning-engineering/about. [Accessed 06-12-2023].

[122] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/37648.pdf

[123] Brian A Nosek, Tom E Hardwicke, Hannah Moshontz, Aurélien Allard, Katherine S Corker, Anna Dreber, Fiona Fidler, Joe Hilgard, Melissa Kline Struhl, Michèle B Nuijten, et al. 2022. Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology* 73 (2022), 719–748.

[124] Bureau of Labor Statistics (BLS). 2023. Technical Workforce Report. https://www.bls.gov/

[125] Institute of Computing (IComp of the Federal University of Amazonas. 2023. CodeBench. https://codebench.icomp.ufam.edu.br/dataset/

[126] Open AI. 2023. OpenAI's GPT-3 Playground Usage Data. https://github.com/openai/gpt-3.

[127] Open Knowledge Foundation. 2015. Open Definition. http://opendefinition.org/od/2.1/en/

[128] Open Source Initiative. 2023. OSI Approved Licenses. https://opensource.org/licenses/

[129] Organisation for Economic Co-operation and Development (OECD). 2018. PISA 2018 Dataset. https://www.oecd.org/pisa/data/2018database/.

[130] Benjamin Paaßen. 2020. Python Programming Dataset. https://doi.org/10.4119/unibi/2941052 Bielefeld University.

[131] James Paterson, Joshua Adams, Laurie White, Andrew Csizmadia, D Cenk Erdil, Derek Foster, Mark Hills, Zain Kazmi, Karthik Kuber, Sajid Nazir, et al. 2021. Designing dissemination and validation of a framework for teaching cloud fundamentals. In *Proceedings of the 2021 Working Group Reports on Innovation and Technology in Computer Science Education*. ACM, New York, 163–181.

[132] Michael Quinn Patton. 2002. *Qualitative Research & Evaluation Methods*. Sage, Thousand Oaks.

[133] Ana Persic, Fernanda Beigel, Simon Hodson, and Peggy Oti-Boateng. 2021. The time for open science is now. *UNESCO Science Report: The race against time for smarter development* 2021 (2021), 12.

[134] 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.

[135] Leo Porter, Daniel Zingaro, Soohyun Nam Liao, Cynthia Taylor, Kevin C Webb, Cynthia Lee, and Michael Clancy. 2019. BDSI: A validated concept inventory for basic data structures. In *Proceedings of the 2019 ACM Conference on International Computing Education Research*. 111–119.

[136] James Prather, Paul Denny, Juho Leinonen, Brett A. Becker, Ibrahim Albluwi, Michelle Craig, Hieke Keuning, Natalie Kiesler, Tobias Kohn, Andrew Luxton-Reilly, Stephen MacNeil, Andrew Petersen, Raymond Pettit, Brent N. Reeves, and Jaromir Savelka. 2023. The Robots Are Here: Navigating the Generative AI Revolution in Computing Education. In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education* (Turku, Finland) *(ITiCSE-WGR '23)*. Association for Computing Machinery, New York, NY, USA, 108–159. https://doi.org/10.1145/3623762.3633499

[137] Thomas W Price, David Hovemeyer, Kelly Rivers, Ge Gao, Austin Cory Bart, Ayaan M Kazerouni, Brett A Becker, Andrew Petersen, Luke Gusukuma, Stephen H Edwards, et al. 2020. Progsnap2: A flexible format for programming process data. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, New York, 356–362.

[138] 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, 1126. https://doi.org/10.1145/3478432.3499092

[139] Rajendra Raj, Mihaela Sabin, John Impagliazzo, David Bowers, Mats Daniels, Felienne Hermans, Natalie Kiesler, Amruth N. Kumar, Bonnie MacKellar, Renée McCauley, Syed Waqar Nabi, and Michael Oudshoorn. 2021. Professional Competencies in Computing Education: Pedagogies and Assessment. In *Proceedings of the 2021 Working Group Report on Innovation and Technology in Computer Science Education* (Virtual Event, Germany) *(ITiCSE-WGR '21)*. ACM, New York, 133–161. https://doi.org/10.1145/3502870.3506570

[140] Rajendra K Raj, Carol J Romanowski, John Impagliazzo, Sherif G Aly, Brett A Becker, Juan Chen, Sheikh Ghafoor, Nasser Giacaman, Steven I Gordon, Cruz Izu, et al. 2020. High performance computing education: Current challenges and future directions. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education*. Association for Computing Machinery, New York, NY, USA, 51–74. https://doi.org/10.1145/3437800.3439203

[141] Joel R Reidenberg and Florian Schaub. 2018. Achieving big data privacy in education. *Theory and Research in Education* 16, 3 (2018), 263–279.

[142] UC Irvine Machine Learning Repository. 2023. Datasets-UCI Machine Learning Repository. https://archive.ics.uci.edu/datasets

[143] Bernat Romagosa, Michael Ball, Jens Mönig, Brian Harvey, and Jadge Hügle. 2023. Snap! Build Your Own Blocks — cloud.snap.berkeley.edu. https://cloud.snap.berkeley.edu/. [Accessed 06-12-2023].

[144] Sage. 2023. Sage. https://us.sagepub.com

[145] Sarah Berndt Sandra Schulz and Anja Hawlitschek. 2023. Exploring students' and lecturers' views on collaboration and cooperation in computer science courses - a qualitative analysis. *Computer Science Education* 33, 3 (2023), 318–341. https://doi.org/10.1080/08993408.2021.2022361 arXiv:https://doi.org/10.1080/08993408.2021.2022361

[146] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten simple rules for reproducible computational research. *PLoS computational biology* 9, 10 (2013), e1003285.

[147] Schloss Dagstul. Leibniz-Zentrum für Informatik. 2023. dblp computer science bibliography. https://dblp.org/

[148] Andreas Scholl and Natalie Kiesler. 2024. Data: Analyzing Chat Protocols of Novice Programmers Solving Introductory Programming Tasks with ChatGPT. https://doi.org/10.17605/OSF.IO/WBKQV

[149] Andreas Scholl and Natalie Kiesler. 2024. Data: How Novice Programmers Use and Experience ChatGPT when Solving Programming Exercises in an Introductory Course. https://doi.org/10.17605/OSF.IO/6EN4Z

[150] Sandra Schulz, Sarah Berndt, and Anja Hawlitschek. 2023. Gruppenarbeit beim Programmieren lernen (GAPL). Datenerhebung: 2020. Version: 1.0.0. Datenpaketzugangsweg: SUF: Download. https://doi.org/10.21249/DZHW:dipit2020:1.0.0

[151] Sue Sentance, Ethel Tshukudu, and Keith Quille. 2022. *METRECC Africa 2020 data.* University of Cambridge. https://doi.org/10.17863/CAM.87121

[152] Otto Seppälä, Petri Ihantola, Essi Isohanni, Juha Sorva, and Arto Vihavainen. 2015. Do we know how difficult the rainfall problem is?. In *Proceedings of the 15th Koli Calling Conference on Computing Education Research.* 87–96.

[153] International Educational Data Mining Society. 2023. Educational Data Mining. https://educationaldatamining.org/ Last access: 2023-11-10.

[154] Daniel Spikol, Olga Viberg, Alejandra Martinez-Mones, and Philip Guo (Eds.). 2023. *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark). ACM, New York.

[155] Kristian Stancin, Patrizia Poscic, and Danijela Jaksic. 2020. Ontologies in education–state of the art. *Education and Information Technologies* 25, 6 (2020), 5301–5320.

[156] Stanford Vision Lab, Stanford University, Princeton University. 2021. ImageNet. https://www.image-net.org/

[157] Anna Stepanova, Alexis Weaver, Joanna Lahey, Gerianne Alexander, and Tracy Hammond. 2021. Hiring CS Graduates: What We Learned from Employers. *ACM Trans. Comput. Educ.* 22, 1, Article 5 (oct 2021), 20 pages. https://doi.org/10.1145/3474623

[158] Cynthia Taylor, Daniel Zingaro, Leo Porter, Kevin C Webb, Cynthia Bailey Lee, and Mike Clancy. 2014. Computer science concept inventories: past and future. *Computer Science Education* 24, 4 (2014), 253–276.

[159] Taylor and Francis. 2023. Taylor & Francis Online. https://www.tandfonline.com/

[160] The Open Knowledge Foundation. 2023. Conformant Licenses. https://opendefinition.org/licenses/

[161] Keith Tran. 2023. Systematic-Analysis of Open Access CSed dataset. https://go.ncsu.edu/csed-dataset Last access: 2023-11-04.

[162] Ethel Tshukudu, Sue Sentance, Oluwatoyin Adelakun-Adeyemo, Brenda Nyaringita, Keith Quille, and Ziling Zhong. 2023. Investigating K-12 Computing Education in Four African Countries (Botswana, Kenya, Nigeria, and Uganda). *ACM Trans. Comput. Educ.* 23, 1, Article 9 (jan 2023), 29 pages. https://doi.org/10.1145/3554924

[163] Antony Unwin and Kim Kleinman. 2021. The iris data set: In search of the source of virginica. *Significance* 18 (2021), 4 pages. https://api.semanticscholar.org/CorpusID:244763032

[164] Zeeshan-Ul-Hassan Usmani and Hussain Shahbaz Khawaja. 2021. Pakistan Intellectual Capital — kaggle.com. https://www.kaggle.com/datasets/zusmani/pakistanintellectualcapitalcs. [Accessed 22-11-2023].

[165] Aline Valente, Maristela Holanda, Ari Melo Mariano, Richard Furuta, and Dilma Da Silva. 2022. Analysis of Academic Databases for Literature Review in the Computer Science Education Field. In *2022 IEEE Frontiers in Education Conference (FIE).* IEEE, Uppsala, Sweden, 1–7. https://doi.org/10.1109/FIE56618.2022.9962393

[166] Tim van der Zee and Justin Reich. 2018. Open education science. *AERA Open* 4, 3 (2018), 2332858418787466.

[167] Laurens Versluis, Mehmet Cetin, Caspar Greeven, Kristian Laursen, Damian Podareanu, Valeriu Codreanu, Alexandru Uta, and Alexandru Iosup. 2023. Less is not more: We need rich datasets to explore. *Future Generation Computer Systems* 142 (2023), 117–130.

[168] VisualData. 2023. VisualData Discovery. https://visualdata.io/discovery

[169] Valdemar Švábenský, Jan Vykopal, and Pavel Čeleda. 2020. What Are Cybersecurity Education Papers About? A Systematic Literature Review of SIGCSE and ITiCSE Conferences. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (Portland, OR, USA) *(SIGCSE '20).* ACM, New York, 2–8. https://doi.org/10.1145/3328778.3366816

[170] Thomas Way, Mary-Angela Papalaskari, Lillian Cassel, Paula Matuszek, Carol Weiss, and Yamini Praveena Tella. 2017. Machine Learning Modules for All Disciplines. In *Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education* (Bologna, Italy) *(ITiCSE '17).* ACM, New York, 84–85. https://doi.org/10.1145/3059009.3072979

[171] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[172] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 9. https://doi.org/10.1038/sdata.2016.18

[173] Ian Wolff, David Broneske, and Veit Köppen. 2022. Towards a Learning Analytics Metadata Model. In *Companion Proceedings*, Alyssa Friend Wise, Roberto Martinez-Maldonado, and Isabel Hilliger (Eds.). 12th International Learning Analytics and Knowledge Conference (LAK'22), Online, 51–53. https://www.solaresearch.org/wp-content/uploads/2022/03/LAK22_CompanionProceedings.pdf

[174] Mustafa Yağcı. 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments* 9, 1 (2022), 11.

A CER Data Resources

Table 4: List of CER data resources and their characteristics.

Resource	Description	Target Group	Link
ACM Digital Library	The ACM Digital Library is a comprehensive digital database owned and published by the Association for Computing Machinery (ACM). It serves as a repository for research articles, conference papers, and other scholarly works primarily in the fields of computer science and information technology.	Researchers, academics, and professionals in computer science and information technology fields.	[1]
IEEE Xplore	IEEE (Institute of Electrical and Electronics Engineers) is an online repository of applied Computing and Electrical Engineering research papers. It mainly consists of journal articles as well as conference proceedings.	Applied Computing and Electrical Engineering researchers in academia and industry.	[77]
Sage	Sage publishes multiple peer-reviewed journals across multiple disciplines. A subset of the journals are open-access	Researchers from all disciplines	[144]
Taylor & Francis Online	Taylor and Francis online is an online repository of publications published by Taylor and Francis. It contains a variety of research articles, primarily journal articles, in multiple academic disciplines.	Researchers from all disciplines	[159]
Corgis	The CORGIS project provides over 40 datasets in areas including history, politics, medicine, and education. Its infrastructure supports the integration of new datasets with simple libraries for Java, Python, and Racket. In addition, several web-based tools allow learners to visualize and explore datasets without programming.	Computing educators who are looking for data to use in their classrooms.	[9]
Datashop	LearnSphere's DataShop is a service offering a repository to store research data and a set of analysis and reporting tools. DataShop is provided by the Pittsburgh Science of Learning Center of the Carnegie Mellon University.	Learning Science community, primarily researchers.	[24]
Dataverse	Dataverse is an open source software installation, which hosts virtual archives ("Dataverse collections") and is run by Harvard University. It has the goal to support researchers who want to share, preserve, cite, explore, and analyze research data.	Researchers, journals, data authors, publishers, data distributors, and affiliated institutions.	[72]
DBLP	The dblp computer science bibliography is a service offered by Schloss Dagstuhl - Leibniz Center for Informatics and the University of Trier Schloss Dagstuhl is a non-profit organization under German law. Its goal is to support and promote the worldwide computer science community. Schloss Dagstuhl is funded by the federal government of Germany and the federal states of Saarland and Rhineland-Palatinate.	Computer science researchers.	[147]
DZHW German Centre for Higher Education Research and Science Studies	The German Centre for Higher Education Research and Science Studies (DZHW) conducts application-oriented empirical research, and provides a research data center for such data. It thus provides a research infrastructure for both other data-collecting institutions and researchers. Its main focus is on higher education research and science studies.	Researchers in the above-mentioned disciplines (higher education research and science studies).	[60]
GitHub	GitHub provides datasets on code repositories, which helps study trends in coding practices, open-source contributions, and collaboration patterns.	Computing educators, researchers, and practitioners.	[62]
Hugging Face	Hugging face is a platform where the machine learning community collaborates on models, datasets, and applications.	Machine Learning researchers, engineers, and practitioners.	[75]
IEEE DataPort	Open source data repositories that is hosted by IEEE.	Researchers who want to upload their data or use other datasets for research or reproducibility.	[76]
Kaggle	Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC. It is a massive repository of community-published models, data, and code.	Data scientists and machine learning engineers.	[86]
MDPI	MDPI is an open access publishing platform based in Basel, Switzerland. It pursues the goal of fostering open scientific exchange forms regardless of the discipline. Currently, it offers 430 diverse open access journals.	Researchers from all disciplines.	[114]
Mendeley Data	Mendeley Data is a free and secure cloud-based communal repository specialized for research data. It is part of Elsevier and powered by Digital Commons Data. Mendeley Data harvests more than 20+ million datasets indexed from 1000s of data repositories regardless of the discipline.	Researchers from all disciplines.	[49]
National Center for Education Statistics	NCES provides a wide range of education-related datasets in the United States, including data on computer and technology use in schools and colleges.	(Computing) educators and computing education researchers.	[119]
National Data Resources	Computing education dataset by Expanding Computing Education Pathways (ECEP) Alliance. The datasets are mostly about CS participation in the US.	Computing education researchers and practitioners in the US	[53]
NSF Public Access Repository	NSF Public Access Repository (PAR) is a repository of research data produced by NSF funded projects	Researchers of NSF projects who are required to make findings or "deposits" publicly accessible	[120]
Open Science Framework (OSF)	OSF is a free and open-source project management tool that supports researchers throughout their entire project lifecycle. It supports finding papers and data associated with research, allowing researchers to share research materials as well as storing and analyzing data and sharing reports with others in the community.	Allow researchers to share data and auxiliary materials on studies.	[27]
Papers with Code	Papers with Code aims to provide free and open resources in the context of Machine Learning papers, and respective code, datasets, methods, and evaluation tables. The offer mainly addresses the disciplines of astronomy, physics, computer sciences, mathematics, and statistics.	Researchers in the aforementioned disciplines.	[115]
UC Irvine Machine Learning Repository	the UC Irvine Machine Learning Repository is a collection of datasets that have been used by researchers to train models successfully. These datasets are made available for future researchers to use through the repository.	Machine learning researchers	[2]
Zenodo	Zenodo is a general-purpose open repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artifacts.	Researchers from all disciplines.	[29]

B Metadata Scheme

Table 5: Minimal metadata specifications for the datasets (legend on the right).

Occurence (Occ)
0-n = optional and repeatable
0-1 = optional, but not repeatable
1-n = required and repeatable
1 = required, but not repeatable

Levels of Obligation (Obl)
Mandatory (M)
Recommended (R)
Optional (O)

No.	Element Name	Entity Name	Occ	Obl	Definition	Example Vocabulary
1	title	title	1	M	Title of the dataset	
2	creator	creator	1-n	M	Main researchers involved in producing the data	
2.1	given name	givenName	0-1	O	First name of the creator	
2.2	family name	familyName	0-1	O	Last name of the creator	
2.3	name identifier	nameIdentifier	0-1	O	Name identifier scheme	ORCID, GND
2.4	affiliation	affiliation	0-n	O	Organizational or institutional affiliation of the creator	
3	URL to Resource	URL	1	M	URL that resolves to the resource or to a landing page for the resource that contains important contextual information including the direct resolvable link to the resource, if applicable	
3.1	Resource URL Type	urlType	0-1	O	Designation of the identifier scheme used for the resource URL, e.g. DOI, ARK, Handle	
4	publisher	publisher	1	M	the name of the entity that holds, archives, publishes, distributes, releases, issues, or produces the resource	Zenodo, Kaggle etc.
5	publication year	publicationYear	1	M	Year in which the data was published	YYYY
6	rights	rights	0-n	R	Any rights information for this resource. The property may be repeated to record complex rights characteristics	Free text. Provide a rights management statemen for the resource or reference a service providing such information. Include embargo information if applicable. Use the complete title of license and include version information if applicable. May be used fo software licenses. Example: Creative Commons Attribution CC-BY 4.0 International
7	description	description	0-n	R	Description may include but is not limited to: an abstract, a table of contents, a graphical representation, or a free-text account of the resource and contained elements.	Free text
8	keywords	keywords	0-n	O	Allocation of Keywords	e.g., CS1, programming, student data
9	language	language	0-n	O	Language of the resource	ISO 639-2 or ISO 639-3 or IETF Best Current Practice 47; Example: eng, ger
10	version	version	0-1	O	Version number of the resource	
11	availability	availability	1	R	Information about the availability of the dataset	controlled vocabulary: open, restricted, closed
12	format	format	0-1	O	Technical format of the resource	e.g. txt, csv
13	data type	data gathering method	1-n	R	Information on the method used to collect the data	e.g., interview, cross-sectional survey, longitudinal survey, log data, learning artifacts, textbook log
14	related publication	relatedPublication	0-n	O	Bibliographic information about related publications	bibtex
CER-Specific Metadata						
15	data collection start date	collectionStart	1-n	R	Start date time (logged)	e.g., YYYY; YYYY-YYYY; YYYY-MM-DD, or Spring 2023
16	data collection end date	collectionEnd	0-n	R	End date time (logged)	e.g., YYYY; YYYY-YYYY; YYYY-MM-DD, or Fall 2023
17	programming language	programmingLanguage	0-n	O	Programming language the data was collected from	e.g. Snap!, Scratch, Java
18	population	population	0-n	O	Information about the study population(s); age/class/level	
19	sample size	sampleSize	0-n	O	Size of the observed population(s)	e.g. number of students, teachers, documents etc.
20	sample demographics	sampleDemographics	0-n	R	Demographics of the selected sample	
20.1	country	country	0-n	R	Country where the data was gathered	
20.2	educational institution	educationalInstitution	0-n	R	Educational institution where the learners got observed	
21	measurement type	measurementType	0-n	O	Type of measurement instrument	e.g., Software, Questionnaire, Performance test, Interview, Video recording, log data
22	data processing	dataProcessing	0-1	O	Indication of any further, secondary modifications to research data	
23	units number	unitsNumber	0-1	O	Number of units in the **entire** dataset	e.g., number of rows
24	task number	taskNumber	0-1	O	Number of unique tasks or problems that the dataset includes	e.g., number of tasks
25	data protection	dataProtection	0-1	R	Details about the level of anonymization on a dataset	e.g., fully anonymized, pseudonymized, none
26	data standard	dataStandard	0-n	O	Details about the data standards if applicable	e.g., ProgSnap
27	learning environment	learningEnvironment	0-n	O	Description of tools or learning platforms used	e.g. iTAP, FITech, CodingBat
28	aggregation	aggregation	0-1	O	Information whether the data is aggregated	yes/no
28.1	aggregation level	aggregationLevel	0-1	O	Information about the level of aggregation if applicable	e.g., keystroke, token, line-by-line, full submission (or blocks/snapshots if applicable to block-based programming)
Additional Properties						
29	research question	researchQuestion	0-1	O	Original research question	From the point of view of the resusers not always relevant
30	future work	futureWork	0-n	O	Research questions that may be answered with the help of this dataset	
31	FAIRness Score	FAIRnessScore	0-n	O	Assessment results of the automated FAIR Data Assessment Tool F-UJI [54] based on the FAIRsFAIR Metrics v0.5 [43]	Total score in percent [%] and scores earned for the individual principles

C Checklist for Researchers

Table 6: Checklist for the Publication of Research Data for CER Researchers

Data Documentation	Documentation for humans	Generic	M	write a ReadMe-File
			O	provide a Code Book
			O	provide a Data Dictionary
			R	use controlled vocabulary and ontologies
	Metadata for machine-readability	Generic	M	Title
			M	Creator
			M	URL to Resource
			M	Publisher
			M	Publication Year
			R	Rights
			R	Description
			R	Availability
			R	Data Type
			O	Keywords
			O	Language
			O	Version
			O	Format
			O	Related Publicatoin
		CER specific	R	Data collection start date
			R	Data collection end date
			R	Sample Demographics
			R	Country
			R	Educational Institution
			R	Data Protection
			O	Programming Language
			O	Population
			O	Sample Size
			O	Measurement Type
			O	Data Processing
			O	Units Number
			O	Task Number
			O	Data Standard
			O	Learning Environment
			O	Aggregation
			O	Aggregation Level
Quality Assurance		Generic	M	Self-control of the quality of the content of the data
			R	Self-control of the quality of the technical quality of the data • open formats • data size
			O	Peer-Review of the data
Legal Issues		Generic	M	Check if you have the right to publish the data
			R	Provide open licenses
			O	Data Protection, in particular: • anonymization • pseudonymization
			O	Check if other legal regulations are relevant, e.g. patents, confidentiality agreements etc.
Data Infrastructure		Generic	R	Choose a good repository that fulfill the criteria: • provides PID, e.g. DOI • provides rights management • provides a minimal metadata schema (preferably domain specific) • enables open data publications • provides license management (you can choose your license) • if applicable, provides versioning of data • fulfills the requirements of a trustworthy repository (has the CoreTrust-Seal oder Data Seal of Approval)
			O	Get a persistent identifier for researchers, e.g. ORCID
General		Generic	R	Publish your data as open as possible but as closed as necessary
			O	Start preparing your data for publication as soon as possible, in particular start with documentation at the very beginning of your project

D Survey Questions

Q1 **Categories of research data used**

Please select all the categories of research data that you used in one of your last computing education research projects:

- ☐ Software developed by myself
- ☐ Software developed by others
- ☐ Qualitative data by myself
- ☐ Qualitative data by others
- ☐ Quantitative data by myself
- ☐ Quantitative data by others
- ☐ Derived or compiled: e.g. text and data mining, compiled database, (3D) models gathered by myself
- ☐ Derived or compiled: e.g. text and data mining, compiled database, (3D) models gathered by others
- ☐ Reference or canonical: e.g. OULAD, gathered by myself
- ☐ Reference or canonical: e.g. OULAD, by others
- ☐ Other (please specify) (open question)
- ☐ None of these

Q2 **Research format(s) used**

Please select the research data formats that you used or created as part of your last computing education research project:

- ☐ Plain text: e.g., unstructured text, essays, etc.
- ☐ Structured text: e.g., XML, JSON, YAML, etc.
- ☐ (Proprietary) general purpose formats: e.g., Office products (Word, Excel, etc.), SPSS data, SQL dumps, etc.
- ☐ (Proprietary) domain-specific formats: e.g., SCORM, Sensor data, etc.
- ☐ Multimedia: e.g., Image, Video, Audio, etc.
- ☐ Compiled binary artifacts: e.g., Models, Executables, etc.
- ☐ Source Code: e.g., in Java, C.
- ☐ Other (please specify) (open question)
- ☐ None of these

Q3 **Research data format(s) to be shared**

Which of the following research data formats you have indicated that you used or created as part of your last computing education research project are important to be made available to others for the purpose of validation?

- ☐ Plain text: e.g., unstructured text, essays, etc.
- ☐ Structured text: e.g., XML, JSON, YAML, etc.
- ☐ (Proprietary) general purpose formats: e.g., Office products (Word, Excel, etc.), SPSS data, SQL dumps, etc.
- ☐ (Proprietary) domain-specific formats: e.g., SCORM, Sensor data, etc.
- ☐ Multimedia: e.g., Image, Video, Audio, etc.
- ☐ Compiled binary artifacts: e.g., Models, Executables, etc.
- ☐ Source Code: e.g., in Java, C.
- ☐ Other (please specify) (open question)
- ☐ None of these

Q4 **File size**

Approximately, what is the volume of the data you created and intend to share as part of your last computing education research project? If you have none, please insert zero.

- megabyte (with input field)
- gigabyte (with input field)
- terabytes (with input field)
- Don't know

Q5 **Number of data files**

Approximately how many files of data did you produce as part of your last computing education research project? If you have none, please insert zero.

- ○ (Input field)
- ○ Don't know

Q6 **Research data ownership**

Who do you believe 'owns' the research data that you have made or will make available to others as part of your last computing education research project?

Funder	○ before publication	○ after publication
State/Governm.	○ before publication	○ after publication
Publisher	○ before publication	○ after publication
Institution	○ before publication	○ after publication
Project Collabor.	○ before publication	○ after publication
Myself	○ before publication	○ after publication

- ○ Other, please specify (open question)
- ○ Don't know

Q7 **Publishing research data**

Have you published the research data that you used or created as part of your last computing education research project in any of the following ways?

- ☐ As appendix to a (peer-reviewed) research publication (e.g., journal article or book chapter). This is in addition to any tables or figures that appeared in the publication itself.
- ☐ As a stand-alone (peer-reviewed) data publication (e.g., a data article in a dedicated data journal or within a data track at a conference).
- ☐ In a research data center.
- ☐ In a data repository provided by my funder.
- ☐ In a data repository provided by a publisher.
- ☐ In a data repository provided by my institution.
- ☐ In a software repository, e.g. GitHub, GitLab, DockerHub.
- ☐ On a personal website.
- ☐ On an institutional website.
- ☐ On another website.
- ☐ Other, please specify (open question)
- ☐ I haven't published the research data.
- ☐ None of the above.

Q8 **Why research data was not published**

Why haven't you published your research data?

- ☐ Didn't know it was possible to publish research data.
- ☐ There was no obligation to publish my research data.
- ☐ There was not enough recognition in my research community.
- ☐ Too many concerns or barriers to overcome.
- ☐ The research data was lost during the research process.
- ☐ Research data is not documented sufficiently.
- ☐ Unsure how to make my research data anonymous.
- ☐ Lack of professional advice or contact persons.
- ☐ Publication of research data will negatively impact my future projects (e.g. participant refusal).
- ☐ It is commercially sensitive.
- ☐ Just haven't had time to publish the research data yet.
- ☐ Other, please specify (open question)

Q9 **Barriers to publishing data**

Thinking about your computing education project still, has any of the following limited your ability to publish research data?

- ☐ Legal concerns (e.g. ownership, privacy).
- ☐ Loss of control of data (e.g., reuse by competitors, misinterpretation).
- ☐ Technical constraints (e.g., data set too big, complicated).
- ☐ Authority or practice considerations (e.g., not peer-reviewed, not done in my field).
- ☐ Resource constraints (e.g., cost, time-consuming).
- ☐ None of the above.

Q10 **Barriers and concerns regarding the publication of research data**

What were the legal concerns that you had regarding publishing research data? (depends on Q9)
- ☐ I'm unfamiliar with the legal regulations.
- ☐ I'm not sure who owns the research data.
- ☐ I'm not allowed to publish the research data.
- ☐ It is difficult to balance privacy and openness.
- ☐ Unclear who will be accountable for published research data.
- ☐ I'm not sure how to deal with personal or sensitive data.
- ☐ Anonymization process is not sufficiently secure to guarantee protection of my research subjects.
- ☐ I'm not sure which license to choose.
- ☐ The research data is subject to patents or could be patented.
- ☐ Other, please specify (open question)

Q11 **Barriers and concerns regarding the publication of research data**

What were your concerns regarding loss of control when publishing research data? (depends on Q9)
- ☐ I'm not confident in the quality of my research data.
- ☐ Anonymization leads to a loss of information, which is significant for research.
- ☐ I'm afraid somebody could misinterpret my research data.
- ☐ Publishing data exposes any errors I made.
- ☐ Somebody could answer one of my research questions before me.
- ☐ Someone may alter my data and republish it.
- ☐ The data is viewed/re-used/distributed/copied for commercial advantage.
- ☐ Other, please specify (open question)

Q12 **Barriers and concerns regarding the publication of research data**

Was one of these authority or practice considerations a reason for not publishing research data? (depends on Q9)
- ☐ Data is not peer-reviewed.
- ☐ Analysis is easier to understand than the raw data.
- ☐ I am under no obligation to publish research data.
- ☐ Little or no value for me as a researcher (not assessed on it).
- ☐ Publishing research data not common in my discipline.
- ☐ Other, please specify (open question)

Q13 **Barriers and concerns regarding the publication of research data**

Did you face one of these technical/processing barriers when publishing research data? (depends on Q9)
- ☐ I'm unfamiliar with the publication process.
- ☐ I require an embargo period for publication.
- ☐ Data files are too large.
- ☐ Data is too complex.
- ☐ Overall publication process is too complex.
- ☐ Unclear where I should publish my research data.
- ☐ Initial anonymization process is not comprehensive.
- ☐ Data curation (e.g. metadata) is too complex.
- ☐ Service (e.g. repo.) I published in before doesn't exist anymore.
- ☐ Other, please specify (open question)
- ☐ I can't say.

Q14 **Barriers and concerns regarding the publication of research data**

Did you face one of these resource barriers when publishing research data? (depends on Q9)
- ☐ Publication of research data is too expensive.
- ☐ Publication of research data requires too much time and/or work effort.
- ☐ I did not find a suitable service (e.g. repository) to publish my research data.
- ☐ Other (please specify) (open question)
- ☐ I can't say.

Q15 **Data/study demographics**

The computing education research project you reported in this study ...
- ☐ ... has been published at an CER conference already (e.g., SIGCSE, ITiCSE, ICER, etc.).
- ☐ … is going to be published at an CER conference (e.g., SIGCSE, ITiCSE, ICER, etc.).
- ☐ … does not fit within the scope of CER (please specify why below).
- ☐ Other reason (please specify) (open question)

Q16 **Publication formats**

What could be a possible way for publishing research data and/or research software at CER conferences (e.g., SIGCSE, ITiCSE, ICER, Koli Calling, etc.)?
- ☐ Existing Formats: Extension of demo/tools/poster formats to include additional materials and guidelines describing the data and technique (approx. 4 pages, double column)
- ☐ Existing Format: Extension of full papers to include research data and software (approx. 10 pages, double column)
- ☐ New Format: Tools track including scientific discussion of software as research method (approx. 6 pages, double column)
- ☐ New Format: Tools track including scientific discussion of software as research method (approx.,10 pages double column)
- ☐ New Format: Data paper format focusing on the description of the research data (approx. 6 pages, double column)
- ☐ Other (please specify) (open question)

Q17 **Mandatory upload of research data and software for publication**

Should each of these formats be associated with the mandatory delivery of a repository, supplementary resources, etc., and if so, when?
- ○ Yes, upon submission of the paper
- ○ Yes, upon acceptance of the paper for publication
- ○ No

Q18 **Review of research data and software**

What are your expectations for the review of delivered resources, repos, supplementary materials? (open question)