A Specification For CS Education Dataset Documentation

Samiha Marwan[1], Austin C. Bart[2] and Thomas W. Price[1]

[1]North Carolina State University, NC, USA
[2]University of Delaware, DE, USA

Abstract

Sharing datasets has many benefits, such as enabling study replication, and supporting secondary analysis. However, many of the publicly available datasets in computing education lack comprehensive documentation or omit key contextual information. For example, missing classroom contextual information (such classroom demographics) or instructional interventions used makes datasets difficult to interpret, and therefore, inhibits the usefulness of shared data. While there is work in standardizing data formats for Computing Ed research (e.g. ProgSnap2 format), there is no standard for describing contextual metadata which is vital to provide high-quality scientific insights. In this paper, we propose a documentation practice to support CS Ed researchers in creating a comprehensive, consistent dataset documentation. We also call for collaboration among researchers to adopt the proposed documentation approach to allow convenient, and reliable use of educational datasets.

1. Introduction & Related Work

Data forms the foundation of any scientific empirical evidence. One way to ensure validity and robustness of such evidence is through data sharing, replication studies, and secondary analysis. In Computing Education Research (CER), secondary data analysis has several benefits: it (1) supports meta-analyses and cross-study comparisons, (2) enables new researchers to test hypotheses without the need to run a new study, and (3) allows for exploring new research questions using existing data. However, even when data is shared in standardized formats, the absence of clear and comprehensive documentation limits the effective reuse of datasets and the potential for secondary analysis.

Researchers often struggle to describe their datasets and study contexts in a way that is complete and clear for others to interpret and reuse data appropriately. Without comprehensive documentation of classroom contextual information or the instructional interventions used, datasets would be difficult to interpret, limiting their usability and raising questions about the validity of their results. For example, a secondary analysis of a programming assignment dataset might observe a low completion rate on a particular task. Without contextual information (such as knowing that this task was offered as an optional extra credit, not a required one), researchers might incorrectly interpret the low completion rate as evidence of poor performance, leading to false conclusions.

Prior research emphasized the importance of data sharing [1], standardization of data formats [2], and standardization of study designs in CER. However, there is much less attention on *standardizing the process of documenting datasets*. The lack of a comprehensive documentation might be due to: (1) researchers find it time intensive to decide which details to include about their data, and classroom context, (2) crafting documentation from scratch can feel overwhelming, and (3) there is often uncertainty about which features and resources are essential to interpret and reuse the data appropriately.

Sandres et al. outlined the advantages and challenges of developing a data repository, and introduced a folder hierarchy design with metadata files [3]. However, their design does not support researchers in *actually writing* clear and complete dataset documentation. A working group in the 2023 CompEd conference examines the available datasets from repositories like DataShop, and GitHub [1]. Their findings revealed that most datasets have inconsistent documentation and are missing key contextual information (e.g. table descriptions, assessment questions, population demographics) which limits the reuse of this data. In the Machine Learning (ML) field, Gebru et al. proposed the concept of a

CSEDM'25: 9th Educational Data Mining in Computer Science Education (CSEDM) Workshop
samarwan@ncsu.edu (S. Marwan); acbart@udel.edu (A. C. Bart); twprice@ncsu.edu (T. W. Price)

'datasheet' design that explains the dataset motivation, usage, and collection process [4]. Pushkarna et al. presented *'Data Cards'* which is a proposed structured summary of essential information in ML datasets that could not be inferred directly from data (such as data training methods, and intended use cases) [5]. While the 'datasheet' and *'Data Cards'* ideas are relevant to our proposed work, they do not align with the specific documentation needs of CS Ed datasets (such as description of instructional strategies and classroom context). **In this preliminary work**, we propose a documentation design embedded within a single, guided file, that prompts researchers with targeted questions to describe key elements of their dataset (e.g., classroom setting, assessment structure, table summaries). Such documentation design can reduce cognitive load by eliminating the need to decide where or how to begin the documentation process.

2. Methods

For space limitations, we provide in Appendix A a template for our proposed documentation design where future users can use. This design has been revised by several CS Ed researchers. For each element in documentation design, we provide a question designed to illustrate the information needed (discussed in detail below). We use the following design principles to create the proposed documentation design: (1) *Balance Effort with Completeness:* The proposed design guides users to include the most critical elements of documentation, while allowing for more comprehensive input when time and resources permit. (2) *Ease of Creation without Redundancy:* The documentation design provides one place to document any useful resources either in the form of adding links, or uploading files without requiring duplicate content. (3) *Guidance for newcomers:* The proposed design can be usable by those unfamiliar with dataset documentation norms, but also flexible and useful for experienced users.

The proposed documentation workflow is implemented as a structured set of questions (see Appendix A), where researchers can complete it by simply answering each question, resulting in a standardized file that serves as a documentation for their dataset. The documentation design begins with a **'Documentation Objective'** which acts as a motivating introduction highlighting the value of a clear and complete documentation. This is followed by a section on **'Author Contact Information'**, which is essential when future users have questions or need access to other resources. The documentation is then organized into four categories of dataset information: **'Dataset Overview'**: This section collects general information about the dataset, such as its name, and relevant published papers to this dataset. The **'Dataset Log Data Attributes'** section collects details about the dataset structure and content, such as files format, tables and their attributes. The **'Dataset Contextual Questions'** section collects metadata information which is any information about the context of the classroom or lab study such as programming environments used, participants' demographics, or any relevant contextual variables that helps in interpreting data accurately. **'Dataset Resources'** section documents any instructional or assessment materials that would help in understanding any evidence resulting from secondary data analysis. For example, dataset resources include the problems used, the assessment questions, assessment policies, final exams, ..., etc. These resources can be directly added to the documentation, or just add a link for this data resource. In case any of the above information cannot be shared publicly, the Author Contact information sections can enable users to request access permission to this data.

In our **future work**, we plan to release example documentation for two datasets, and to implement this documentation design where some fields can be autocompleted, such as the tables' names, or column names that could be automatically extracted from the uploaded data. Additionally, we want to call the CER community to incentivize datasets' documentation to encourage dataset owners to document their data, which is crucial to enable reuse and secondary analysis on their datasets.

Acknowledgments

This material is based upon work supported by the National Science Foundation under grant #2213792.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check, and Paraphrase. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] N. Kiesler, J. Impagliazzo, K. Biernacka, A. Kapoor, Z. Kazmi, S. G. Ramagoni, A. Sane, K. Tran, S. Taneja, Z. Wu, Where's the data? exploring datasets in computing education, in: Proceedings of the ACM Conference on Global Computing Education Vol 2, 2023, pp. 209–210.

[2] T. W. Price, D. Hovemeyer, K. Rivers, G. Gao, A. C. Bart, A. M. Kazerouni, B. A. Becker, A. Petersen, L. Gusukuma, S. H. Edwards, et al., Progsnap2: A flexible format for programming process data, in: Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education, 2020, pp. 356–362.

[3] K. Sanders, B. Richards, J. E. Moström, V. Almstrum, S. Edwards, S. Fincher, K. Gunion, M. Hall, B. Hanks, S. Lonergan, et al., Dcer: sharing empirical computer science education data, in: Proceedings of the Fourth international Workshop on Computing Education Research, 2008, pp. 137–148.

[4] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. III, K. Crawford, Datasheets for datasets 64 (2021) 86–92. doi:10.1145/3458723.

[5] M. Pushkarna, A. Zaldivar, O. Kjartansson, Data cards: Purposeful and transparent dataset documentation for responsible ai, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 1776–1826.

A. Appendix - Dataset Documentation Template

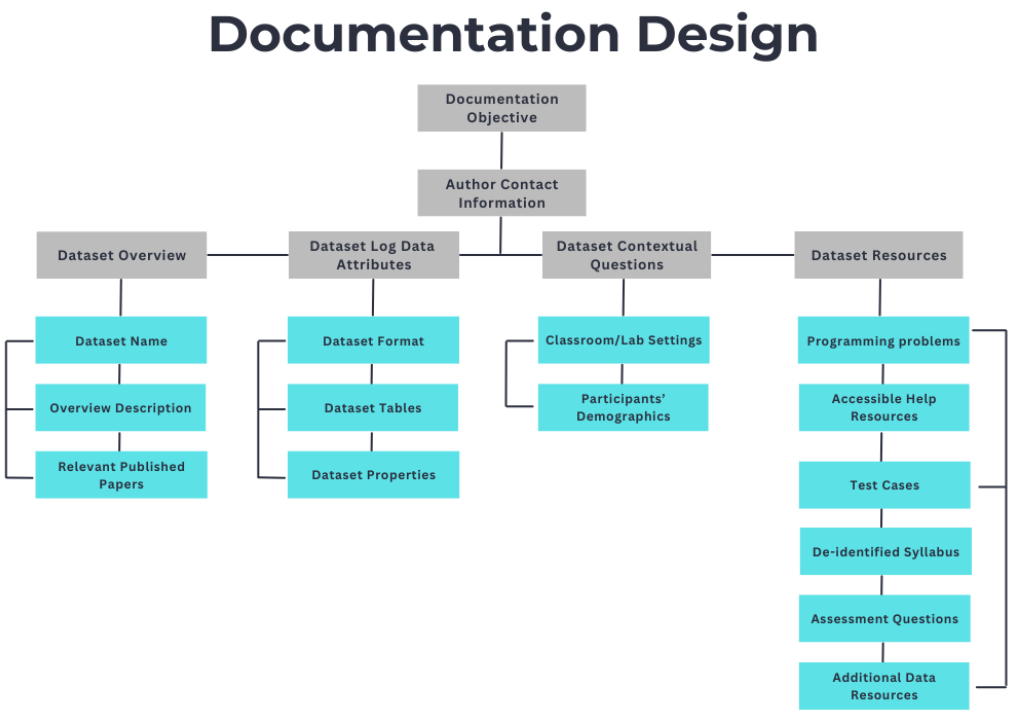

Figure 1: Documentation Design Chart

A.1. Dataset Documentation Objective

There are multiple benefits to documenting data. For example, thoughtful dataset communication can facilitate and enable data reuse, support data analysis replication, allow collaboration between researchers and educators, and mitigate risks and incorrect results caused by inaccurate data.

Please answer as many questions as you can. If a question has no answer, write 'N/A'.

A.2. Author Contact Information:

1. Name:
2. Email address [Preferable if the email is not linked to an institution]:
3. Backup Contact [if applicable]:

A.3. Dataset Overview

This section provides an overview of the dataset.

1. What is the Dataset Name? [answer]
2. Please Provide an Overview Description of the Dataset: [answer]
3. If you have, Please Add Relevant Published Papers: [answer]

A.4. Dataset Log Data Attributes

This section includes all information about the dataset tables.

A.4.1. Dataset Format

1. Provide an overview of the dataset format. If you have, add a reference of this format: The data is stored in the [answer] format, ….

A.4.2. Dataset Tables

For each table, write its name, short description, its columns, and short description of each.

TABLE 1

1. Name:
2. Description: [answer]
3. No. of Columns:
4. Columns: [write down each column name, a description for it, and, *if applicable*, the possible values of this column]

TABLE 2

1. Name:
2. Description: [answer]
3. No. of Columns:
4. Columns: [write down each column name, a description for it, and, *if applicable*, the possible values of this column]

A.4.3. Dataset Properties

1. Are there any unique properties for this dataset? [answer]

A.5. Dataset Contextual Questions

This section contains questions designed to provide any relevant details that help in understanding and interpreting the data.

A.5.1. Classroom/Lab Settings

As applicable, please provide:

1. Programming Language Used:
2. Programming Environments Used:
3. Number of Instructors:
4. Duration of Activities, or Lab Work:

A.5.2. Participants' Demographics

As applicable, please provide a paper that includes the student/participants demographics, or type the population's:

1. Population Size:
2. Age:
3. Grade Level:
4. Gender:
5. Race/Ethnicity of Students:
6. Prior CS Experience:

A.5.3. Assessment Settings

As applicable, please provide:

1. What kind of assignments did the student have? (e.g. quizzes, projects, multiple choice questions, ...)? [answer]
2. What is the Grading Policies (e.g. : late policies, penalties, ..., etc)? [answer]
3. What are the topics covered before the data is collected?
4. What is the grade percentage of each assessment? [answer]
5. How the assessments are assessed (if applicable, can you provide the rubric?) [answer]

A.6. Dataset Resources

This section collects additional file resources to give more context to the log data. For each question, you can either type your answer or upload a file. If not applicable, write N/A.

1. What are the Programming Problems Text Given to Students? The assignments for the course, their descriptions, and solutions, can be found in ...
2. What are the Help Resources Accessible to Students during Practices (e.g. feedback type, or other interventions)? [answer]
3. If possible, Please Include a De-Identified Syllabus: [answer]
4. If possible, Please Provide the Assessments Questions (e.g. Midterms, Final Exams, Pretest, Posttest): [answer]
5. If you have any Additional Data Resources, Please Upload it Here: [answer]