Managing the Educational Dataset Lifecycle with DataShop

John C. Stamper[1], Kenneth R. Koedinger[1], Ryan S.J.d. Baker[2], Alida Skogsholm[1], Brett Leber[1], Sandy Demi[1], Shawnwen Yu[1], Duncan Spencer[1]
[1] Carnegie Mellon University, Human-Computer Interaction Institute
[2] Worcester Polytechnic Institute, Department of Social Science and Policy Studies
[1]{ jstamper, krk, alida, bleber, sdemi, shanwen, dspencer}@cs.cmu.edu, [2]rsbaker@wpi.edu

1 Introduction

An ideal scenario for educational research is to perform an experiment, report and publish results, make the results and data available for verification, and finally allow the data to be used in follow up experiments or for secondary analyses. Unfortunately, this scenario often fails after the results are published. Researchers move on to new data and the old data may linger on a legacy server for a short while before disappearing or becoming impossible to comprehend. Managing the dataset lifecycle is a way to address this problem. DataShop (http://pslcdatashop.org) is a central hub for data management of educational data, and in this paper we show how DataShop fits into the dataset lifecycle.

DataShop is an open data repository and set of associated visualization and analysis tools accessible on the web[2]. DataShop has data comprised of millions of student interactions with on-line course materials and intelligent tutoring systems. The data is fine-grained, with student actions recorded roughly every 20 seconds, and it is longitudinal, spanning semester or year-long courses. As of April 8, 2011, over 270 datasets are stored including over 58 million student actions and over 165,000 student hours of data. Most student actions are "coded" meaning they are not only graded as correct or incorrect, but are categorized in terms of the hypothesized competencies or "knowledge components" (KCs) needed to perform that action. Visualizations and statistical analysis tools in DataShop are designed to help model builders and analysts find potential flaws in an existing student model. In the hands of trained users these tools provide a method for discovery of KCs that better match student learning data. As the developers of DataShop, we often overlook the repository features in favor of the tools, but the DataShop open repository is rich in features and provides a strong foundation to follow the steps of the educational dataset lifecycle.

2 The Educational Dataset Lifecycle

We define six steps in the educational dataset lifecycle that are illustrated in Figure 1.

Data Design is the most important step in the lifecycle. As part of a research design, the data design should identify what data will be necessary for analysis to confirm research hypotheses, but should also be forward thinking about what data

could be used in additional analyses. Any data that is easy to collect, whether or not it will impact the initial research, should be considered. Although this step is not specifically linked to DataShop in Figure 1, DataShop can inform the data design process by providing detailed documentation on our tutor message format[1], which will not only make data easy to import into the database, but also provide an excellent reference for data that researchers should strive to collect. Also, by accessing publicly available projects, researchers can explore the data design of other studies that have used DataShop.

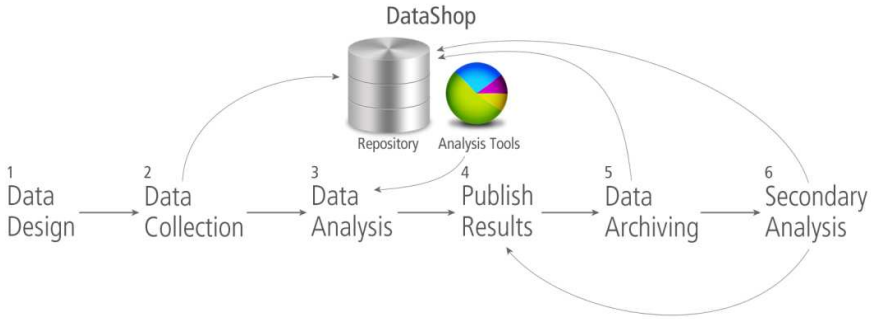

Fig. 1. The six steps of the Educational Dataset Lifecycle showing interactions with DataShop.

Data Collection is the actual logging and storing of data. DataShop offers a number of ways to enable data collection. Direct data logging via our logging API allows for data to be automatically imported into DataShop. The CTAT authoring tool [1], which allows non-programmers to create adaptive tutors, has built-in logging tools for DataShop based on the API, and others have used this functionality as well. Data import via text file or the more robust XML format are also available. For transactional data, such as student logs, custom fields are available in the database. For more unusual types of data that may not fit the structure of DataShop's internal database, there is a facility to attach files directly to a project. There are no restrictions on what these files contain (the exception being legal and copyright issues), and the files might be text logs, spreadsheet information, or even video or audio data files.

Data Analysis is the fundamental part of a research project. A strong data design will make setting up the analysis easier. Some analysis can be performed using the tools available in DataShop. The current tools focus on knowledge component (KC) analysis. Learning curve and error analysis are also provided at the student, problem, or problem step level. Understanding that many researchers may not find these tools useful for their research activities, DataShop provides several different export views of the data into a text file format readily accepted into most analytical tools.

Publish Results is an important step in the lifecycle. In addition to just presenting the results, publications should include information about the structure of the data or at least suggest where this information can be found. DataShop provides the ability to

[1] http://pslcdatashop.org/dtd

link research papers to a dataset. Linking papers not only provides a background on the dataset; it also increases the visibility of the linked papers to other researchers.

Data Archival differs from data collection in that the archival process must focus on making the data accessible and understandable to future researchers. Archival should include background information that clearly explains the structure of the data. Including additional data files, such as pre and post test materials, as part of a dataset is also important. Once the initial research is completed on a dataset, it should be archived in such a way that others could recreate the experiment, and others can clearly understand the data to allow for secondary analysis.

Secondary Analysis can provide tremendous value to the community but is rarely done. The main obstacle with secondary analysis is that the dataset is often missing critical metadata needed to make sense of the data. This is especially the case when the researcher performing the secondary analysis was not part of the original research. If a project is archived properly, any researcher should be able take the data and recreate the original analysis. It is important that the data is adequately described so that the data is not misunderstood or taken out of context in secondary analysis. If the data is structured in a defined format, such as the tutor message format in DataShop, analyses setup to run on one dataset can be applied to many datasets in the same format. This opens up opportunities for educational data mining studies to cover a large number of domains in an efficient manner. To date, over 75 secondary analyses studies have used datasets in DataShop.

DataShop is focused on becoming the premier repository for educational data. We recognize our current data model does not meet every educational researcher's needs and are working to expand the data model to be more inclusive. We are also working to improve the meta tagging available to allow researchers to better document their datasets, and to make the metadata easier to search.

As the cost of collecting and storing data continues to decrease, researchers will become increasingly inundated with larger and more robust data. This is a good thing, but without a sound data management plan, the data could become worthless or, even worse, become misinterpreted and lead to incorrect conclusions. The US National Science Foundation has recognized the importance of data management, and is now requiring a data management plan to be included in every research proposal submitted. We believe that following the steps of the educational dataset lifecycle and incorporating the DataShop repository will enhance data management and allow for better research in the future. DataShop is supported through NSF award 0836012.

References

1. Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. (2006). The Cognitive Tutor Authoring Tools (CTAT): Preliminary evaluation of efficiency gains. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), Proceedings of the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), (pp. 61-70). Berlin: Springer Verlag.
2. Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2010) A Data Repository for the EDM community: The PSLC DataShop. In *Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining.* Boca Raton, FL: CRC Press. pp. 43-56.