DATA MANAGEMENT PLAN

Writers:
Shreya Goel
Devan Goto
Chuheng Hu

Submitted To: Amy Nurnberger

Teachers College
Columbia University

December 19, 2016

Abstract

When working in interactive learning environments, some students attempt to "game the system." In this paper, we present a "proposal for data management and sharing" in a future research project. The research project is to design and implement a detector that can accurately identify whether a student is gaming the system and is hurt or not, within a Cognitive Tutor mathematics curricula. In this research project we would also explore this detector's generalizability, and find that it transfers to both new students and new tutor lessons. We discuss the goals, recommendations, data to be produced, data management strategies and budget for this future research project.

Section 1: Introduction

We have chosen to manage the data of a published article. The article that we chose is "Developing a Generalizable Detector of When Students Game The System (Baker, Corbett, Roll, & Koedinger, 2008)." This article can be retrieved from: https://link.springer.com/article/10.1007/s11257-007-9045-6, by downloading the pdf. For this assignment we are assuming that this article has not yet been done, and that the research is yet to begin. We are not evaluating how they managed their data; instead we are presenting how we would manage their data.

Our goal within this report is to successfully showcase how we would manage the article's data. Describing, and/or making recommendations in seven different areas will do this. Each area will focus on an aspect of managing educational data. In addition we will also have two more sections, to supplement the seven areas of managing data. Our report will start with one of those two sections. In the first section we will briefly discuss the goals of the project and the assumptions of the operating environment within the article. This will provide some contextual information that

will give a background on the research within the article. Which will lay the groundwork for our first area of data management, description of data to be produced.

Without a section describing your data there is no context, and without context one's data is almost rendered unusable (Carlson, & Anderson, 2007). Simply, knowing the data you're producing is the first step to managing your data. In this section we will discuss and describe all of the data that is involved in the research article. This will include class lesson data, class observational data, and log files, among other data. This area will incorporate skills we used in assignment one, such as describing data interactions and its sources.

Once we have described the data, the next step is describing which data we will keep and how we will curate it. As we discussed in class, there are some data that you collect that you may not want to share. In this study we are dealing with students (human subjects) in middle school. Therefore we must comply with IRB rules and protect the privacy rights of the students (IES, 2016). In this section we will discuss our data collection and selection process, while also providing our curation policy.

After discussing our collecting and selecting process of the data we will go into detail of our documentation process for curation. The ability to re-use and replicate data is often dependent on "enough context" (Carlson, & Anderson, 2007). According to Kervin, Michener, and Cook most errors in re-use are due to a lack of adequate metadata (2015). If we want our study's data to be reused or replicated we must provide metadata. The last thing we want is to provide data of poor quality, as it can be extremely harmful for organizations in analytics (Stiles, 2012). In this section we will discuss metadata in three forms: descriptive, structural, and administrative. This section will provide others the means to either re-use or replicate the data. By doing this we are incorporating a few of assignment three's data quality categories.

Once all of the metadata is complete, in lieu with data collection and selection, we must discuss how we will share our data. As discussed by Cooper, interoperability is important in learning analytics (2014). Which is why we discussed the formats in which we will share our data before we discussed where we will share our data; with the goal of optimizing interoperability in the long term. In this section we will also discuss our choice of repository. Our repository will have "a mission to provide reliable, long-term access to managed digital resources to its designated community, now and into the future (CRL, OCLC, 2007)." The last thing we want is to choose a repository that goes down, and doesn't have our data backed up.

In addition to our repository we will also have other storage locations, and means to protect our data. In this section we will discuss data storage and security requirements. A fair question is, "why would we want to secure data that is public?" Different items/data should require different levels of security (Castillo, 2013). This is true for our data. While what we are making public may not require a high level of security, some of our raw private data will; and this is why we will be discussing security.

The levels of security are imbedded within our ethical and legal considerations. We will recognize student data laws and IRB rules. To assist we will use the recently passed SOPIPA law as a guideline for this section (Leong, Polonetsky, 2016). While ethics is something that is often overlooked it is something we deem invaluable.

Following ethical and legal considerations is our budget section. This will encompass the budget for managing the data. Our budget will fall within the UK Data Service's framework (2013). We will also justify each of our proportional values by providing monetary estimations and explanations of each category.

Our final section will be the IES data plan. This will be a brief compilation of the research proposal. Like assignment two we will concisely and coherently portray our data. However, this

time we will be focusing on data sharing. This will include items such as documentation, data sharing methodology, and ethical/legal considerations.

Section 2: Goals of the project and assumptions of the operating environment

This project aims to develop a gaming behavior detector for the cognitive tutor system and evaluate the generalizability of our detector.

The three goals for our project are: to testify if the gaming detectors can effectively detect gaming; to distinguish between gaming that hurts from that does not hurt; to generalize the gaming detector to different data set.

2.1 Cognitive Tutor: The cognitive tutor is a learning software that assigns students subject questions and gives them response to their answers. This software will break the problems into the steps of the process used to solve the problem in order to help visualize student thinking. And according to the students answer, they will give response messages to indicate if their answers are right or wrong, and it may even provide students a little instruction if a known misconception is detected. A student can also choose hint for extra assistance.

2.2 Gaming the system: The detector is used to detect when students are playing with the hint function and not using it the way they are supposed to. The behaviors that we will record as gaming in the in class observations are: 1. Quickly and repeatedly asking for help until the tutor gives the student the correct answer (cf. Aleven, 2001). 2. Inputting answers quickly and systematically.

2.3 Research Pipeline:

2.3.1 Detector Building: To build the detector for gaming behaviors that hurts we would need to collect data from more than 100 student/lesson pairs in order for statistical significance. We will collect the log files from the Cognitive Tutor and to extract useful information including:

1) Details about the action 2) Knowledge assessment 3) Time 4) Previous interaction. We will also collect data from field observations to categorize behaviors into: gaming the system, off-task behavior, talking on-task, and working in the tutor.

Besides, in order to distinguish gaming that hurt from those doesn't hurt, we will give students pretests and posttests for each lesson, to determine those whose marks drops when gamed.

Then we will combine a Fast Correlation-Based Filtering (Yu & Liu, 2003) 2 and Forward Selection (Ramsey & Schafer, 1997) method to select the predictors for gaming that hurts.

2.3.2 Generalizability testing: After building the detector we will run cross validation to test on the new student group, which we collect using the same way. We will then compute the accuracy of our detector to ensure it's high enough to devise intervention.

Section 3. Description of the data to be produced

In our project we wish to work on designing, testing and then generalizing a gaming detector within Cognitive Tutor, a learning environment that is designed to promote learning by doing.

Defining the Data of our Research:

3.1 Obtaining the Data: To test this gaming detector, we will attempt to get data from XYZ schools, in New York City, who are using cognitive tutors as a learning platform for mathematics. We base our concentration on same tutoring domains and collect lessons on probability and geometry in dataset – I, for testing and generalizing the gaming detector; and lessons on probability, geometry, percentage and algebra for dataset – II, for testing the generalizability of the gaming detector. For dataset – I and dataset – II we intend to collect data from 2 different schools each for a duration of 2 years for case I and a duration of 1 year for case II. Data for both these datasets will be collected and processed together; however, it will be post processed separately at different stages.

Scott et al. (2012) mentioned the various stages of data life cycle and the various other concepts of data category, data volume, data types, formats which we have used to describe our data as follows.

Data Stage	Data Description	Data Source/Category	Data Volume	Format
Raw Data	Collect Data from schools coordinating with Cognitive tutors	Cognitive Tutor, Schools	Combination of small and large files.	.csv .txt .log
Pre-Processed Data	Dataset – I and II will include: Knowledge component model, Class observational data, Pretest & posttest for each lesson, Log files, Metadata	Reference, Observational, Textual and Numerical data	Combination of small and large files	.csv .txt .log
Processed Data	Data is processed and Gaming Detector is designed (Dataset – I and II): Knowledge map, Class Observation Data, Cognitive tutor log Dataset, metadata, algorithms	Derived, Observational, Textual, Numerical data and Code Book	Combination of small and large files	.csv .txt .rmkd
Post processing Data - I	Test and Improve the Gaming Detector if needed using Dataset - I	Derived, Observational, Textual, Numerical data, Code Book	Combination of small and large files	.csv .txt .rmkd
Post processing Data - II	Code books – gaming detector. Test the generalizability of gaming detector using Dataset - II	Derived, Observational, Textual and Numerical data, Code Book	Combination of small and large files	.csv .txt .rmkd
Analyzing Data	Analyze if the gaming detector can be used as Generalized Detector	Derived, Observational, Textual and Numerical data, Code Book	Combination of small and large files	.csv .txt .rmkd
Publish Data	Publish the results of the designing a generalized gaming detector	Textual and Numerical data, Code Book	Combination of small and large files	.csv .txt, .pdf
Curate Data	Upload the data to Data Shop and Columbia Academic Commons,	Textual and Numerical data, Codebook	Combination of small and large files	.csv .txt .rmkd .log

	including metadata: Scripts, Codebooks; Logfiles, Algorithms, Datasets			

3.2 Using the Data: After we have collected the data for our research, we process the data. In this process, we intend to clean these datasets. Knowledge component model is where the components of knowledge are arranged in a computer readable format. This model helps the tutor analyze if the student knows a particular skill or not. It will not require much of processing. Log files will be in a computer readable format. We will attempt to convert the raw log files to raw log dataset after anonymizing these while they are linked with the class observational datasets using RNG (RNG: Random Number Generator - a program which provides a random number to anonymize the datasets). We will then attempt to clean the dataset. The class observational data along with the pretest and posttest scores will also be anonymized using the same software RNG, and to avoid any misinterpretation of anonymized data, we will anonymize the complete dataset at once before the log files are converted to datasets, so that each student gets one and only one random number instead of 2 or 3 in different datasets. Process of cleaning of the observational class data and log files will include processes like: managing NULL values, anonymizing data, removing any undesired variables, including pretest and posttest scores if not included already, personalized progress on various knowledge components. Variables like "student ids, unique steps, total steps, total time, pre-test score, post test scores, number of hints used, responses of cognitive tutor, time intervals between hints, knowledge components in the exercise and more" are expected in these files.

The post processing stage will be divided broadly into 2 parts – I and II. In the Ist stage we will attempt to use dataset 1 to: Firstly, differentiate between students who gamed the behavior from those who did not. Secondly, differentiate between the students who have been hurt due to gaming (students who did gained less than 2 skills from pretest to posttest) from those who have not been

hurt due to gaming (students who learnt at least 2 new skills from pretest to posttest stage; students who had no scope of improvement due to already possessed skill – only 0 or 1 skill could be gained, only in high level performers).

In the IInd stage we will attempt to test the generalizability of the gaming detector using the dataset – II. Same steps will be followed as in Post processing stage – I. After the post processing stage we will analyze the generalizability of the gaming detector using the results so obtained.

3.3 Looking after the Data: File names will be very wisely chosen, which do not hinder the future exploration of the files amongst others (Scott et al. 2012). Metadata will be updated throughout the research project, in order to retain all the important information (Scott et al. 2012). Concerns about selection, collection and curation of data; data storage and security have been addressed separately in section 4 and 7 respectively. We use a variety of backup methods, which are again discussed in detail in section 7. By the end of the research project we estimate to produce about 300 gigabytes of data.

3.4 Summary: In this section we described the data to be collected, processed and analyzed during this future research project. We used the article "Introducing Research Data" by Scott et al. (2012) as a guideline to formulate this section of the proposal.

Section 4: Selection, collection and curation policy

4.1 Purpose of Selection collection and curation policy: Appraisal and selections of data is defined as " *evaluate data and select for long-term curation and preservation* "(Higgins, S., 2008). The role of appraisal and selections of data in a data curation pipeline is important and mainly beneficiary to 2 stakeholders:

1) The research team themselves, since with appropriate selection, money used to be spent on archiving those data sets that are unlikely to be useful in the long-term will be saved. Aside from that, *"they will also benefit from knowing in advance how their own data will be assessed, and what*

they should plan to do in order to increase the chance of their research having an enduring impact." (Whyte, A, et.al,2010) .

2) The other researchers who are interested in the data will also benefit, since discovery process will be made easier when the data is selected. (NISO,2007)

4.2 The aspects to consider: We will adopt the how to guide: How to Appraise and Select Research Data for Curation and rank the 7 quantiles in the order that is specific to our case.

The aspects we will consider are the followings:

The consideration:	Specifications
Relevance to mission	Our research goal is to generate a detector for gaming behavior and test the generalizability of such detector, thus, undoubtedly, the prime data are data we collected from the log file of the cognitive tutor , the in class observation recorded by our researchers, the pre-test and the post-test scores. And the following aspects will give us a standard while considering: what else to collect, what should be eliminated, which version of data to keep.
Full documentation	For the readability of our data, meta-data and code book need to be archived. And different version of datasets may be preserved to serve for the full documentation purpose depend on different cases. For instance, we need to preserve the anonymized raw log file, so that we will be able to trace back and prove the accuracy of our data.
Scientific and historic value	It is not really applicable to our case, the prime datasets are unlikely to have different historic or scientific value, considering the fact that the data we collect only serve for our research goal. Although, we might consider the metadata explaining our method and variables are less likely to be outdated compared to the primary data.
Non-replicability and uniqueness	Technically, all the data could be replicated but it will not be the same datasets and it's extremely time consuming. Especially, the data retrieved from the cognitive tutor must be associate with the in class observation in order for it to be meaningful.
Economic Case (storage cost, and management cost)	We understand that some uncollected data (like those features in the log file and in class recordings that were not transcribed) might be valuable to other studies. However, constrained by the economic case. We will give priority to the data that is most relevance to mission.
Privacy Issue	All names and ids need to be removed from our preserved datasets including the log file and the in-class observation.
Potential for redistribution	The redistribution is largely depend on the data explanations which includes meta-data and code book

4.3 The selection, collection and curation policy: We understand that selection and appraisal is an ongoing process along with the cycle of our research. Thus it requires the data archiver to work closely with our researcher and to make sure that we have selected the appropriate raw data, processed data, and metadata for future selection.

The specific data curation policy for our project is asking the data curator to work with the researchers at the end of each research phase and clarify the version of datasets that is going to be kept in the repository by going through the 7 perspectives with the researcher, and also select the dataset for sharing before posting it to the data shop.

Section 5: Documentation, Reusability and reproducibility

In this section, we will establish the convention and requirements in documentation including file naming, metadata, and citation as well as other rules to follow in order to achieve higher reusability and reproducibility.

5.1 Metadata: There are many standards for meta-data like DDI, DUBLIN CORE, MARC, ISO 11179 etc. and we chose to adopt the DDI 3.0 for their unique features and their compatibility with other standards. We are clear that the two main scientific norms that relies upon having good metadata are data sharing and replication (Vardigan M., Heus P., et.al, 2008). We understand that in order for any secondary analyst and a researcher to understand and work with the data set, the open dataset need to come with a neat documentation which need the cooperation among the different parties in a research team. Thus I will specify the requirements in response to the particular people who will be responsible for that, in most time either the researcher or the data curator.

I adopted the list of important metadata elements posted by ICPSR (2016) and categorize them into descriptive, structural and administrative. All the metadata will be found either in the readMe file, the data code book or the description section on the data sharing platform.

5.1.1 Descriptive Metadata

	Description	**People in Charge**	**Where to find**
Contact	In this part, we will include the contact information (Name, Email Address, Role) of the Principal investigator(PI), the funding sources and data collector/produce	Data curator is responsible for collecting these information from the researchers	in the contact session of the readMe file.
Project description	We will include a brief explanation of the goal of the study, the role of our data	The source of this is already produced by the researcher in the introduction sections. However, the curator need to make it more applicable to data users.	in the readMe file , on the project description section on the data shop which is where we are going to share the data, and in the data code book.
Data collection	We will include a brief summary of the data collection method, and the time when it's collected. Also we will specify how different collection method act in the study. For instance the data collected through in-class observation vs. the log file.	Researchers and research assistants are responsible for providing these information and work with curator about how to clearly express it.	In the readMe file and the data code book.
Data source	We will include a brief summary of the platform where we got the data from, the course subject and content involved. And for the processed data like the code posted, we will provide explanatory files to explain our algorithm.	Researchers and research assistants are responsible for providing these information and work with curator about how to clearly express it.	In the readMe file and the data code book.

5.1.2 Structural metadata

	Description	**People in Charge**	**Where to find**
Data Units	We will include a brief summary of the number of students involved, the # of actions we captured	Data curator is responsible for collecting these information from the researchers	In the readMe file and the data code book.
Variables	We will include a brief summary of the # variable involved , the explanation of each variable and the order of the variables. More	The source of this is already produced in the "data description" section of our paper. The data	In the project description section at the data shop , and in the data code book.

	specifically, we will state the meaning of each column name and the possible entries under each column. The explanation of knowledge components variables will also be included.	curator need to make some refinement for it to be more useful for other researchers And explanation for KC were established by cognitive tutor.	And the KC are already available on data shop.

5.1.3 Administrative metadata

	Description	**People in Charge**	**Where to find**
Technical information on files	We will include the related Information on file formats, and file linking. More specifically it will be a map of the directories of different files and the software we recommend to open different files.	Data curator is responsible for recording and updating these information at the end of each experiment period.	In an additional pdf format of the readMe file, since we think a visual explanation would be better than verbal explanation.
Citations to related publications	We will include any thing cited in the paper or relevant to the data.	Data curator will build a standalone reference document based on the reference section.	In the standalone reference document in .txt form.
Data Collection Instrument	We will establish the rules we followed in making in class observation and a link through which people could access the training for using cognitive tutor which is the website of cognitive tutor.	Data curator is responsible for collecting these information from the researchers	In the data code book and on the cognitive tutor website.
Additional assessment	We will include the person to contact if you need to gain additional assessment to the original log file or the different versions of the data.	Data curator is responsible for collecting these information from the researchers	In the readMe. file.

5.2 File Naming: The file naming for ongoing research process will include the version. In establishing the conventions, I referred to file naming instructions discussed by Nurnberger, A. (2016) and made a few adjustment according to this specific study.

Since the different data files may differ in the version of the file, source of the data (cognitive tutor or in class observation), Researcher's name, the course, the time that data are collected at, and the number of subjects involved. The other information's regarding different files

are consistent. Thus based on these differences, we need to follow a convention that will allow researchers to identify the file they are looking for in the naming. The following convention is discussed and approved among the research team:

Ihe initial of the researcher_"ob"for in-class observation and "ct" for cognitive tutor_YYYYMM_number of students_Version N. xxx,

5.3 Standards for completeness, consistency and structure of the data: Since part the data is derived from the log file. The nature of the data has avoided some problems like miss-recorded data fake data and absent data at the first place. And for the class observed data, we will ask the researcher to follow the strict rules we established to avoid inconsistency. The possible blank entries should be explicitly marked as NA(not applicable). They may appear in some columns where the particular action didn't fall in such column.

In order to ensure better structure, each data file will follow the same column order which is: Id, the contents associated with student's action, the tutor's response, and the corresponded information of the involved knowledge/material.

5.4 File formats: As discussed in the data description section, we will adopt the common formats for our files that are unlikely to be outdated , commonly accepted and easily interpreted. And we will use R to analyze our data, since R is free to download and unlikely to be outdated.

5.5 Conclusion: Reusability and reproducibility: In our case, the primary data we relied on is from the cognitive tutor, thus some of the variables are exclusive for the users of this software. Nevertheless, we will present the data in a neat format as stated in the data description section and we will provide explanatory meta datas with our file. We currently is not expecting a huge amount of reuse coming from bystanding parties. However, we are maintaining high accessibility, high quality meta-data, conventional data forms and high data quality to support possible reuse and reproduction.

We will also post the anonymized log file on our repositories. By doing so, we believe the reuse value of our data will increase since we allow people to 1)check the accuracy of our data 2) extract other features from the log file that we didn't. Last but not least, through posting our code and metadata that explains it , we hope to increase the reproducibility of our data.

Section 6: Sharing/Communication plan for data products

Our data is available as pdf documents, csv, log files , txt files, and our codes of our algorithm in Rmkd formats will be widely and freely disseminated minimally via the DataShop and Columbia academic commons. We chose DataShop to share most of our primary data mainly for 2 reasons: 1) Due to the specific data source (cognitive tutor), the audience of our data might be other fellow researchers in the field of learning analytics or other users of the cognitive tutor system who are already familiar with the datashop. And since the datashop is designed for preserving data retrieved from cognitive tutor, it has a more friendly navigation and system for our data. 2) The datashop, as an established repository, requires users to login before exporting the data, thus, they will monitor the audience of this data for security purpose.

The accessibility of the data will follow the terms of conditions of the DataShop. If any user requires extra help, they could contact us through the information attached within the ReadMe file where explicitly addresses the PI's contact. Despite the advantages of the datashop, we will also have a secondary repository, Columbia academic commons, since " DataShop makes no guarantees about the data in terms of quality or future availability" as mentioned in the IES DP.

Section 7: Storage & Security

The data files will be managed, processed, and stored in a secure environment. We will have firewall systems, lockable computer systems with passwords, virus protection, and surge protection. Access to digital files will be protected with password protections. Our log files will be anonymized, but we will also maintain a file of the raw log files. This will *not* be available for

public consumption. The PI will hold this data on a secure computer and hard drive, and manage/maintain it. Any inquires or requests about this data will be directed to the PI. This data is not to be shared or checked out unless approved by the PI.

A master copy of each data file (such as anonymized data, and metadata) will be stored on 6 secure locations: computer, usb flash drive, external hard-drive, and three online digital repositories. For our repositories we intend to use PSLC DataShop & Academic Commons for sharing and posting our data, but we will also use a third local repository to privately host our data until the research has been completed and published. Until then we will privately upload an updated version of our data into the local repository every 14 days, just in case our other 3 storage locations go down. As another layer of storage we will also be using Mac's time machine. As we have discussed in class, backups are good, but what's really important is that the backups work. We intend to address this by checking every storage location and time machine before use to validate that they all work properly. In addition, we will check all of our storage locations to make sure they are working properly every two weeks.

Section 8: Ethical & Legal Considerations

For this assignment we are tasked with managing data for a research project that involves students in middle schools. This means before we can begin to manage the data produced by our research team, we must first get certified by Teachers College (TC) Institutional Review Board (IRB). We will do this by completing nine modules (seven basic and 2 elective), with 100% accuracy, on the CITI program. "The CITI program is becoming the standard online training program and is widely accepted at institutions, organizations, and research programs around the world (Teachers College, Columbia University, 2016)." After we have successfully completed the models we will file our IRB applications with TC and await approval. We will not begin to manage data until we have become IRB certified. In addition before we begin our study every student

(participant) in the study will have to sign a participation consent form. This will make all parents aware and knowledgeable in our data management practices, and if they are not satisfied they can opt out accordingly.

Since we are managing data of middle school students, a challenge we face is in relation to ethics and legality. We will approach this challenge by imposing SOPIPA (Senate Bill No. 1177 Chapter 839) upon ourselves. Although we do not have to do this we are doing this because we want to set a positive precedent within the learning analytics community. Learning analytics is built on a foundation of trust, accountability, and transparency (Pardo, Siemens, 2014); and this is precisely why we are imposing this upon ourselves even when we do not have to. This is a bill passed in California that protects student data. Its four main points are as follows: Do not advertise based on any student data; Do not use student data to amass a profile about a student; Do not sell student data to a third party; Do not disclose covered information.

To combat this we will take two precautions. The first to fully anonymize all of the identification data we collect, such as names and ID numbers. We will remove all such data, in our shared datasets, to protect the privacy of every student. The second precaution we will take is to create a terms of use document for anyone who want's to use our data. In the terms we will include many terms, but the two main ones are as follows; Data is not for commercial use, only for research purposes; De-anonymizing the data is strictly prohibited.

One other legal consideration is in regards to sharing our knowledge component model (map). DataShop has an agreement with Cognitive Tutors that will allow us to post this information, but we are unsure if Academic Commons does. To combat this we will only post this information on DataShop, and wait for approval or denial from Cognitive Tutors to decide whether or not we'll post it on Academic Commons.

In order to make sure that ethical and legal considerations are always in place, our PI will be responsible for enforcing every issue addressed here. In addition he or she will be in place to validate that we not only abide by the issues above, but also remember to address each issue.

Section 9. Budget: For Managing the Project Data

In this section, we present the Budget for managing the project data, for the entire duration of the project. We have used the Data Management Costing Tool and Checklist, provided by UK Data Services in 2013. This section is divided further into 2 sections, explaining the budget and its justification.

1. Data Cleaning: $10,200 ($10,200 / $15,935 = 64.01%): Once the log files and class observational data have been anonymized they must be sent to a data cleaner. They will need to convert the log files into a datasets with the proper organization that our researchers have requested. In addition they will need to format all of the data from the log files into a dataset. These tasks require a high level of skill and time to accomplish. We are estimating that this will take an expert 102 hours to accomplish. After doing some due-diligence we found that $100 an hour is a reasonable rate for the efficiency (Soleil, 2013). Thus, $100*$102= $10,200.
2. Anonymization: $750 ($750 / $15,935 = 4.71%): Identification indicators need to be eliminated from the log files and class observational data, and arbitrary name ID's must be put into place (via Excel's RNG). This process is will require someone to manually generate arbitrary ID's for all students while matching the names from two separate sources (log files and class observational data). This is estimated to take 50 hours, and we will hire a data manager at a $15 hourly rate. Thus, $15*50=$750
3. Metadata: $2,500 ($2,500 / $15,935 = 15.69%): We will have three types of metadata: descriptive, structural, and administrative. All metadata can be found in a combination of the readme file, data codebook, and the description of data. These items will clearly depict every aspect

of our data management, from methodology to data description. This is not a simple task and we expect this to take up a decent amount of time. We are estimating that this will take a total of 125 hours. For this task we will hire a data manager at a $20 hourly rate. Thus, $20*125 = $2,500

4. File Formatting: $225 ($225 / $15,935 = 1.41%): We will offer an array of file formats overall, but we will not have to switch through many different types. For this reason we are estimating that this will only take 15 hours to complete. This task will be addressed to a data manager at a $15 hourly wage. Thus, $15*15 = $225

5. Data Storage & Sharing: $500 ($500 / $15,935 = 3.14%): Space & maintenance will be the duty of the repository and PI. Academic Commons provided a repository that will maintain our data, but we will also provide a flat fee for the PI of $500 to maintain the raw log files for a minimum of 10 years.

6. Data Transfer & Access: $500 ($500 / $15,935 = 3.14%): This will be the duty of the PI for the first 10 years, and he or she will decide what to do when that duration ends. This will entail managing any applications received to view the raw log files. He or she will be paid a flat fee of $500 for this time span.

7. Data Back-Up: $200 ($200 / $15,935 = 1.25%): The repositories are free and Time Machine is a free application in Mac computers. The usb's and external hard drives are expected to cost an estimated $200.

8. Data Security: $1,060 ($1,060 / $15,935 = 6.65%): For firewall & virus protection we will utilize Avast's Premier service for a $60 cost. For Data Insurance we are estimating this to be $1,000 (UK Data Service, 2016), for a total of $1,060.

Summary: Our budget for managing the data shows us that two sections account for more than 79% of our budget. The reason is because data cleaning and metadata require a lot of effort/time. Data cleaning was not only our most time consuming task, but also our most expensive. Experts in

data scrubbing go for rates upwards of $110+ an hour, but we decided that $102 an hour was a fair rate. The other six sections accounted for around 20% of the budget, and this is because they were things that did not require much time or money. While anonymization was something that was going to take some time to do, we were only intending to pay $15 an hour. Something like security was $1,060 total, but in proportion to our data cleaning cost it was relatively small. Thus we will allocate 79.7 credits to data cleaning and metadata, while we will only provide 20.3 credits to the other six sections of our budget.

Section 10: IES Data Plan

Types of Data To Be Shared & Procedures for managing and for maintaining the confidentiality of the data to be shared; For this project we will be sharing our code (algorithm), knowledge maps, anonymized log files, and cleaned datasets. Everything we share will be fully anonymized by eliminating all student identification indicators (such as name) and will be replaced with a unique arbitrary ID generated by an Excel program. This will protect the privacy of each student in our data. For our non-anonymized data we will be taking an extra precaution by having it stored on a secure computer and hard drive.

Roles and responsibilities of project or institutional staff in the management and retention of research data; For this project we intend to post our data on two repositories. Since DataShop does not guarantee the longevity of the data, we have also decided to incorporate Academic Commons as a repository to host our data. They will maintain the data and protect its longevity.

Expected schedule for data sharing, The data will be made available to the public once the "main findings from the final study dataset are published in a peer-reviewed scholarly publication (IES, 2016)." The data will be made available on two repositories once this comes into fruition.

Format of the final electronic dataset; File formats will be offered in an array of forms: csv, rmd, txt, among other formats. For a more detailed explanation please visit our "File Formats" section.

Documentation to be provided; Documentation will be provided with consistent file names, diverse file formats, metadata, citation, and codebook, among others. This information is in place to ensure that replication of our research team's analysis. For a more detailed description of documentation please visit our "Documentation, Reusability, and Reproducibility" section.

Method of data sharing; The PI will provide the data for public consumption through two repositories once the research has findings have been complete and published.

Whether or not a data sharing agreement that specifies conditions under which the data will be shared is required; Our terms for DataShop are taken directly from DataShop. The terms provided on DataShop are clear and provide a good precedent for data management. These terms can be retrieved at https://pslcdatashop.web.cmu.edu/Terms. Academic Commons is in place to fill-in where DataShop falls short, such as "DataShop makes no guarantees about the data in terms of quality or future availability (DataShop, 2016).

Any circumstances that prevent all or some of the data from being shared: To release our knowledge maps we may need third party consent from Cognitive Tutors. On DataShop it seems that they have an agreement that will enable us to share this information, but we are unsure if Academic Commons will allow us to do the same. To combat this we will only post this information on DataShop, and wait for approval or denial from Cognitive Tutors to decide whether or not we'll post it on Academic Commons.

References

Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008) Developing a Generalizable Detector of When Students Game the System. User Modeling and User-Adapted Interaction, 18, 3, 287-314. https://link.springer.com/article/10.1007/s11257-007-9045-6

Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. Journal of Computer-Mediated Communication, 12(2), 635–651. https://doi.org/10.1111/j.1083-6101.2007.00342.x

Castillo, T. D. (2013, March). Information security explained. https://blogs.ucl.ac.uk/dmp-ss/2013/03/26/information-security-explained/

Cooper, A. (2014, March). Learning Analytics Interoperability – The Big Picture in Brief http://laceproject.eu/publications/briefing-01.pdf

CRL, OCLC. (2007, February). Trustworthy Repositories Audit & Certification: Criteria and Checklist ru). Online Computer Library Center, Inc (OCLC) and The Center for Research

DataShop. (2016). DataShop Terms of Use. Retrieved from https://pslcdatashop.web.cmu.edu/Terms

DCSM Vardigan, M., Heus, P., & Thomas, W. (2008). The Digital Curation Sustainability Model. Data Documentation Initiative: Toward a Standard for the Social Sciences.*International Journal of Digital Curation*, 3(1), 107–113.

Higgins, S. (2008) The DCC Curation Lifecycle Model, The International Journal of Digital Curation, Issue 1, Volume 3 , 2008 . Retrieved from: http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48

ICPSR: A Case Study in Repository Management (2016.)Retrieved from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/lifecycle/metadata.html

IES. (2016) Data Sharing. https://ies.ed.gov/funding/datasharing_implementation.asp

Kervin, Karina E.; Michener, William K.; and Cook, Robert B. (2013). "Common Errors in Ecological Data Sharing." Journal of eScience Librarianship 2(2): Article 1. http://dx.doi.org/10.7191/jeslib.2013.1024

Leong, B. and Polonetsky, J. (2016). Passing the Privacy Test as Student Data Laws Take Effect. EdSurge. https://www.edsurge.com/news/2016-01-12-passing-the-privacy-test-as-student-data-lawstakeeffect?utm_content=bufferc0042&utm_medium=social&utm_source=twitter.cm&utm_campaign=buffer

Libraries (CRL). Retrieved from http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

NISO Framework Working Group. (2007, December). A Framework of Guidance for Building Good Digital Collections - 3rd edition. *National Information Standards Organization (NISO*). Retrieved from http://www.niso.org/publications/rp/framework3.pdf

Nurnberger, A. (2016) On naming your files , Retrieved from https://moodle2.tc.columbia.edu/pluginfile.php/421389/mod_resource/content/1/HUDK4054_Metadata_201609_v01.pdf

Pardo, A. and Siemens, G. (2014), Ethical and privacy principles for learning analytics. Br J Educ Technol, 45: 438–450. https://doi.org/10.1111/bjet.12152

Scott, M., Boardman, R., Reed, P., and Cox, S., (2012). Introducing Research Data. University of Southampton, Faculty of Engineering and the Environment http://eprints.soton.ac.uk/338816/

Soleil, O. (2013). What Does It Cost To Clean My Data? Retrieved from http://datascopic.net/cost-of-data-cleansing/?doing_wp_cron=1481742173.2197239398956298828125

Stiles, R. J. (2012). Understanding and Managing the Risks of Analytics in Higher Education: A Guide. Educause. Retrieved from https://net.educause.edu/ir/library/pdf/epub1201.pdf

Teachers College, Columbia University. (2016). IRB Training and Certification http://www.tc.columbia.edu/institutional-review-board/training-and-certification/

UK Data Service. (2013). UK Data Service – Data management costing tool and checklist. http://data-archive.ac.uk/media/247429/costingtool.pdf

Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. *Edinburgh: Digital Curation Centre*. Retrieved from: http://www.dcc.ac.uk/resources/how-guides