Supported by the National Science Foundation

Sharing In LearnSphere
Study Report
September 19, 2017

Una-May O'Reilly and Nicole Hoffman
CSAIL, MIT

Executive Summary

LearnSphere serves as an infrastructure for a community interested in interpreting the student data that is collected from digital learning interactions. It facilitates the web-based demonstration, authoring and dissemination of software workflows that analyze the digital learning data through modeling. Additionally, it acts as a both a repository and portal for software and data related to digital learning interactions.

One critical question with an infrastructure like LearnSphere is whether its users will donate their software and datasets to each other. Without donation, LearnSphere's growth would rest solely on the commitment of its founders and their disciples. What are the specific challenges that software and data donors face? What are the reasons for them? What can be done to foster or encourage data and software contribution?

We have examined these questions. Our first main finding is that, **with respect to users donating software, we can expect with high confidence that users will donate software**. Current attitudes, incentives, requirements and support for contributing software are all ideally aligned to foster donation. We believe that there are constructive ways to encourage and support this donation and while they are necessary, they are not onerous or challenging. LearnSphere largely has efforts and plans in place to provide them.

Our second main finding is that **the situation is dichotomous when it comes to donating data**, prompting two models. One, **direct donation**, involves donation of de-identified or anonymous data that has been pre-cleared for release at time of data collection. LearnSphere continues to successfully exploit this model. The other model, **data access**, which LearnSphere is adopting, respects a research institution's desire to store their own data and approve requests for its use (exclusively by the requestor) by LearnSphere directing requesters to the approval process at appropriate engagement points. Section 4.4 of this report provides a list of ways both models can be further encouraged and considers risks to their success.

Given the merit of data access model and importance of software donation, LearnSphere should incorporate metrics for software and data access facilitation into its self-assessment of use.

Contents

1 Introduction

1.1 Purpose of Study and Document

A goal of the LearnSphere Project is to set up a research community infrastructure on a national scale that supports modeling and data analysis tools centered on digital learning interactions. This will accelerate data-driven discovery and innovations in teaching and learning fields that generate student interaction data.

LearnSphere facilitates workflow authoring for student modeling and data analysis. It provides the TIGRIS workflow that users can run with data LearnSphere provides or with data of their own. It allows users to edit existing workflows or create their own. Most importantly, it facilitates users in becoming community contributors by donating both software and data.

This study considers the topic of data and software donation in LearnSphere. It is motivated by the recognition that donation is challenging. A goal of the study is to identify what factors within and external to LearnSphere impact donation. Based upon this information, we provide analysis on what will foster donation.

1.2 Study Process

We planned when we proposed the study to conduct a large survey that would ask whether the respondents expected to donate data and software. If not, the survey would ask why. At the start however we made a course correction when we realized respondents did not know enough about LearnSphere. Instead, we first surveyed the LearnSphere founders then, we surveyed attendees of a conference tutorial where LearnSphere's workflow was taught. This meant our respondents had appropriate context. Throughout the study we have collected information about data and software sharing beyond LearnSphere.

1.3 Roadmap to Report

We proceed by first, in Section 2, providing an overview of LearnSphere and its stakeholders. We frame it with respect to other efforts. We describe the data and software in question. In Section 3 we review software sharing in general and how it is motivated and facilitated. We describe how LearnSphere facilitates software donation survey then analyze and make recommendations. In Section 4 we proceed similarly. Section 4 presents actions LearnSphere can take.

2 LearnSphere Description

2.1 Overview: Infrastructure = Repository, Portal, Workflow

The LearnSphere Project is envisioned as a key national repository of digital learning interaction data sponsored by the National Science Foundation (NSF). LearnSphere will serve as an infrastructure for the community of researchers and practitioners interested in analyzing the data that is collected from digital learner interactions. Facilitating the web-based demonstration, design, construction and dissemination of software workflows that analyze digital learning data through modeling across a wide variety of educational data. It acts as a repository and a portal for data and software related to workflows and models.

2.2 Elements

LearnSphere's founders are connecting their pre-existing or concurrent efforts on the following projects to LearnSphere by making their data and or methods available through it:

- DataShop Founders K. Koedinger, J. Stamper, Stamper et al. [2011].
- MOOCdb - Founders O'Reilly and Veeramachaneni, Veeramachaneni et al. [2014]
- DiscourseDB - Founder: Carolyn Rose, Jo et al. [EDM 2016], Rose [2017].
- DataStage - Founder Candace Thille
- OpenDataShop at Memphis - founders Pavlik, Graesser, Hu, Cai, and Nye. First clone of DataShop. Pavlik Jr et al. [2016], Stamper et al. [2016]

These founders have created a workflow named Tigris that will eventually demonstrate a combination of their analytic techniques and data sources. We next describe these projects.

2.2.1 DataShop@CMU

DataShop Stamper et al. [2011] is a data repository and web application for learning science researchers, funded by a National Science Foundation. DataShop is the work of the combined effort of CMU, Memphis U., Stanford U. and TutorGen; a company geared towards building the next-generation tools for adding adaptive and personalized capabilities to educational software. It provides secure data storage as well as an array of analysis and visualization tools available through a web-based interface. DataShop can be stood up locally (e.g. `https://datashop.memphis.edu/`, `https://datashop.stanford.edu/`, `https://datashop.tutorgen.com/`). Users do not currently contribute to or have access to the DataShop system software. We defer how users access and donate data to DataShop to Section 4.

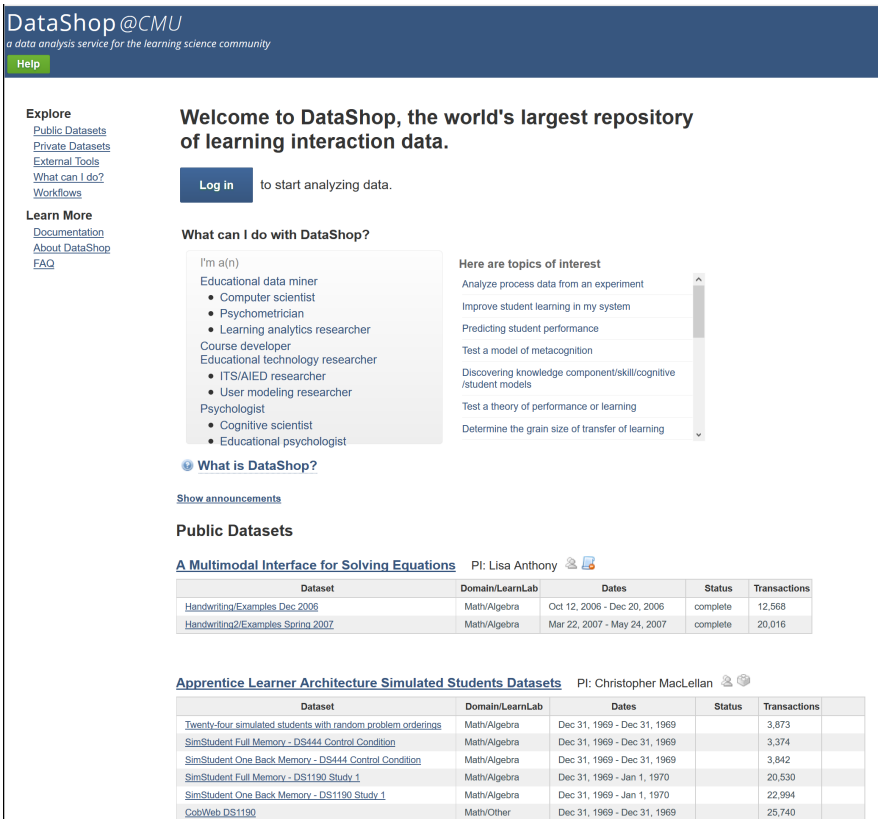

Figure 1: DataShop@CMU landing page

2.2.2 MOOCdb

The MOOCdb project Veeramachaneni et al. [2014, 2015a,b] aims to support education researchers, computer science researchers and machine learning researchers in developing software technology for MOOC data science. The project provides a platform-agnostic functional data model (also called MOOCdb). The lowest layer of its open source software curates different data streams from the edX platform [1] into a set of tables according to the data model. Higher layers transform the data for machine learning, i.e. prepare features. The software is currently at `https://github.com/MOOCdb`.

MOOCdb does NOT provide or collect MOOC data donations. See Section 4 for further explanation and discussion. Users cannot yet access MOOCdb software repository or website through LearnSphere. To date, the open release of extensions to MOOCdb software (or the data model) by third parties has not occurred.

[1] Also Coursera, first version.

2.2.3 DiscourseDB

DiscourseDB Jo et al. [EDM 2016], Rose [2017] is a data infrastructure project that aims to provide a common data model to accommodate diverse sources including but not limited to chat, threaded discussions, blogs, Twitter, wikis and text messaging. The project supports analytics which will facilitate research questions related to the mediating and moderating effects of role taking, help exchange, collaborative knowledge construction and others.

DiscourseDB will be hosted independently from LearnSphere but will be well-integrated with LearnSphere from the perspective of a researcher using it. LearnSphere will allow researchers to see a list of DiscourseDB datasets are available to them presently, as well datasets they may request access to along with a way to request access. Researchers may access DiscourseDB data in two ways: (1) on the DiscourseDB website, they may browse, query, and analyze data using DiscourseDB-native tools, using the same login credentials as the LearnSphere site, or (2) on the LearnSphere website, Tigris components will be available that remotely query DiscourseDB data and introduce it into a workflow; this allows them to apply other LearnSphere members' analyses to DiscourseDB data, or do combined analyses of DiscourseDB data with data from other LearnSphere datasets.

Currently there are two MOOC datasets that will be sharable once DiscourseDB is set up to connect to the LearnSphere portal: the discussion forum from the psych MOOC (a Coursera Discussion forum) and the discussion forum from DALMOOC (an edX discussion forum). In both cases users must ask for permission to use the data. To access datasets through DiscourseDB, a researcher would have to complete training in human subjects research ethics and complete an online data request form before receiving any data.

A documentation wiki can for DiscourseDB can be viewed at `https://discoursedb.github.io/`

2.2.4 DataStage

DataStage holds learning research data derived from courses offered on three platforms: NovoEd, Coursera, and Lagunita, a Stanford University instance of the OpenEdX platform. The platforms are instrumented to collect a variety of data around participants' interaction with the study material. Examples are participants manipulating video players as they view portions of a class, solution submissions to problem sets, uses of the online forum available for some classes, peer grading activities, and some demographic data.

VPOL makes some of this data available for research on learning processes, and for explorations into improving instruction through DataStage. We explain this process in Section 4.

2.2.5 OpenDataShop at Memphis

This instance of DataShop at Memphis Pavlik Jr et al. [2016], Stamper et al. [2016] is interoperable with LearnSphere by way of sharing data ID services with the main DataShop@CMU. Having a remote instance of DataShop gives greater flexibility for local data providers that need local storage or enhanced security. Having the data locally stored motivates local data providers to integrate with the entire LearnSphere project through interaction with the local node. For example, several systems investigated at Memphis, including AutoTutor, ALEKS, and MoFaCTS are now logging in DataShop format natively, or with efficient converters. Having a full-fledged instance of DataShop running locally also allows development of the workflow components independently from the LearnSphere hub. As part of the Memphis development several workflow components are being produced for data analysis, import and transformation.

2.2.6 TIGRIS: LearnSphere's Analytics Workflow

Tigris is a workflow authoring tool. The workflow is a component-based process model that can be used to analyze, manipulate, or visualize data. A generic workflow might consist of components for each of the following steps:

- Import a tab-delimited file and build a TIGRIS input dataset.
- Analyze the file contents with a psychometric modeling technique.
- Visualize the results of the analysis component.

Each component acts as a standalone (black box) program with its own inputs, options, and outputs. The inputs to each component can be data or files, and the output of each component is made available after the workflow has been run. Tigris workflow supports one input data format (tab-eliminated text) and Tigris provides transformers to change between data formats in case input data (or intermediate component output data) is a different format. Each component has a set of required inputs and user-defined options. Each is defined by an XML schema definition (XSD) defining its structure. Components can be flexibly connected into a flow (e.g. import, model then visualize). The newly generated data of one component is passed to the component's successor.

Tigris allows users to start with an example, make changes and then save locally or share. It supports DataShop models. Efforts are underway to extend a provided workflow to include components from Discourse DB and MOOCdb.

2.3 Stakeholders

LearnSphere has a number of stakeholders. They are:

- NSF

- Founders (CMU: JS, KK, CPR, MIT: UMOR, KV, Stanford: CT, UM: PP)
- Instructional teams
- Instructional design teams
- Researchers in learning science and education
- Students

Stakeholders have roles:

- NSF funds the project.
- Founders are developing the infrastructure including authoring tools, portal, access mechanisms and sharing mechanisms. They populate it with their research methods and both publicize and teach about it.
- Students act as subjects in studies.
- Users: The remaining stakeholders (instructional teams, instructional design teams, researchers) are users. Users can also be data and software contributors and have data control granting power.

2.4 Data Properties

LearnSphere data can be described with multiple axes:

- Learner: K-12 or college.
- Source: intelligent tutoring system, problem bank, MOOC, online course.
- Who controls granting sharing rights: researcher, institution, (maker and platform provider).
- How the interactions are logged: clickstream, text (emails, forums and inline comments, written responses, reports), video watching, solutions to problems, speech, digital tool outcomes (circuit designs), programs.
- Collected with IRB that allows donation or extension of use after collection vs collected for a specific study.

2.5 Related Digital Learning Efforts

There are several efforts that investigate learner behavioral data and develop open source analytics for research purposes. They can be similar in goals to LearnSphere or have overlapping issues. They could act as sources of data and software that LearnSphere acts as a portal to eventually. Some are supported by small teams of researchers who provide their outputs on GitHub and a website, e.g. Pardos et al. [2016]. LearnSphere would offer them visibility. Others by larger organizations, e.g. SOLARs effort to create an open software data analytics platform and LACEs effort to brings together evidences about the effects of learning analytics from across the world are uniquely different from LearnSphere and complement it. Yet others by Teaching and Learning Labs/Groups of big Universities. These units mostly care about different data uses but some conduct research similar in scope to LearnSphere. Some efforts revolve around a platform, like Future Learn which is a for-profit platform launched from the Open University. The community and scope of data and software control and sharing around Future Learn is similar to edX.

2.5.1 Science of Learning

Science of Learning developed by Johns Hopkins Science of Learning Institute supports interdisciplinary research that will generate scientific discoveries and build meaningful connections between research, practice, and policy. The vision of Science of Learning is to solve learning issues with innovative research and practices generating scientific discoveries. Their mission is to understand and optimize the ability to learn and to build meaningful connections between research practice and policy through reaching their 3 goals:

- Supporting cutting-edge science of learning research.
- Training future leaders in the science of learning.
- Connecting science to practice.

Science of Learning's stakeholders are students, educators, learning science researchers. Their stakeholders also include the policy makers that will institute their findings in the classroom.

2.5.2 Open Learning Initiative (OLI)

Open Learning Initiative (OLI) offers open and free courses across a number of academic disciplines to colleges, universities and individuals across the US and many other countries. OLI does not grant accreditation for the completion of any course and does not provide any certification of completion.

OLI has 3 key goals:

- Supporting better online learning and instruction.

- Share courses openly and freely.
- Develop a community of use, research and development.

Courses are designed using learning science research to support classroom instruction while supporting the individual learner. OLIs data-driven design captures real-time learner data from everyday instructional activity allowing for immediate and continuous evaluation and feedback so students can assess their own learning effectively. The data driven design also enables the refinement of courses and course materials by informing course designers how students perform in courses while also contributing data to learning science research.

OLI stakeholders, similarly to LearnSphere stakeholders are students, educators, academic institutions, course designers, learning science researchers, founders, and sponsors. OLI and LearnSphere are similar in their goals of collecting and interpreting student data from digital learning interaction to improve online learning and in the roles of the stakeholders. Differing from LearnSphere, OLIs software is not open source and does not share data openly.

OLI is funded by grants from the National Science Foundation (NSF) and industry endowments, is an open-access MOOC platform developed by Carnegie Mellon University and Stanford University.

2.5.3 LearnLab

LearnLab, is a distinctive feature of the Carnegie Mellon Universitys Simon Initiative to improve student learning outcomes. LearnLab is a research facility to support field-based experimentation, data collection, and data mining which aims to enhances scientific understanding of robust learning in educational settings by designing educational technologies to produce an increase in student achievement.

LearnLabs stakeholders are instructors, course designers, learning science researchers, founders, and sponsors. LearnLab and LearnSphere are similar in their goals of collecting and interpreting student data from digital learning interaction to improve online learning. LearnLab does not openly share software, its datasets are available through DataStage.

LearnLab is a Science of Learning Center funded by the National Science Foundation (NSF) and operated by Carnegie Mellon University (CMU).

2.5.4 FutureLearn

FutureLearn is a private for-profit international distance-learning MOOC platform developed by The Open University in the United Kingdom, a public distance learning and research university that offers online classes from over 80 academic institutions from around the world, including the US. FutureLearn offers interdisciplinary courses, programs for professional accreditation, or post-graduate degrees. It has a user base among schools who use it.

Future Learn Stakeholders are Open U, Centre for Education Tech and Interoperability Standards U of Bolton, Infinity Tech solutions, Swedish National Agency for Ed, AtiT (Audiovisual Technologies, Informatics and Telecommunications), Kennisnet, Oslo and Akerhus U College of Applied Sciences, outside institutions, learning researchers, course designers/developers.

Subject interactions are primarily through discussion comments. Unlike many MOOC platforms, discussion can happen on any of the content pages, as opposed to going to different forums. FutureLearn discourages the use of usernames and insists subjects use their real names as identifiers, encouraging users to share personal identifiable information such as their location in discussion comments to help other learners get to know you.

No sharing of software or data by this effort specifically, more towards exploring learning analytics and accumulating evidence to support the use of learning analytics. It is not clear how the data they collect is curated and maintained or whether or not identifying data is scrubbed before sharing with their partners. As outlined in the user terms and conditions the data it collects may be used by FutureLearn, their partners or course and content providers and FutureLearn cannot guarantee the the security or safety implications of the sharing of subjects data.

FutureLearn is associated with Learning Analytics Community Exchange (LACE), led by Doug Clow and Rebecca Ferguson at The Open University, funded by the European Union. LACE connects experts and researchers in Learning Analytics and educational data mining to promote the creation and exchange of knowledge, increase the evidence base for Learning Analytics, and build consensus on interoperability and data sharing.

2.5.5 LACE

LACE (Learning Analytics Community Exchange) is being led by Doug Clow and Rebecca Ferguson at The Open University UK. The LACE Evidence Hub does not share or donate learning data, but brings together evidences about the effects of learning analytics from across the world, and has the capacity to inform policy making for learning analytics by recording, organizing and searching evidence relating to the theory, research and practice of learning analytics and associated educational data mining.

LACE works to support or contest the effectiveness of learning analytics with evidence-based decisions about learning and teaching to improve learning outcomes and improve learning support by remonstrating the validity of the predictive models used by learning analytics systems.

Everyone is welcome to add evidence for or against these propositions to the Hub site. LACE stakeholders are similar to LearnSphere in that they include learners, researchers, course designers and teachers. An additional stakeholder LACE mentions that is not in LearnSphere are policy makers as they strive to influence online learning policy.

2.5.6 SOLAR

The Society for Learning Analytics Research (SoLAR) is an inter-disciplinary network of leading researchers from institutions around the world who are exploring the role and impact of analytics on teaching, learning, training and development. In 2011, SoLAR proposed the creation of an integratable and modularized platform with an extensible toolset to assist in the evaluation of learner activity for educational researchers and professionals. Complete with a user-friendly dashboard with visualized data for learners and educators, institutions, the system would support researchers, administrators, educators, and learners integrating heterogeneous learning analytics techniques. Siemens et al. [2012]

SoLAR lists several aspirations in their proposal for the analytics platform in their 2012 white paper:

- The development of common language for data exchange.
- Analytics engine: transparency in algorithms so learners and educators are aware data is being gathered, researchers can customize analytic methods to reflect the needs of different contexts.
- Dashboards and reporting tools to visualize information and provide real-time information to learners, educators, administrators, and researchers.
- Open repository of anonymized data for training and research development.

SoLARs goal for the project is that all universities in Next Generation Analytics Consortium share anonymized data with each other to facilitate benchmarking and promote ethical and effective practices between institutions. Though this concept is still theoretical, SoLAR concludes integrated data sharing with an education analytics platform will have a profound effect on Learning Science. The technical strategy is that this infrastructure would be an open-platform that would allow software plug-ins that would enable comparison tools and datasets.

The open-platform infrastructure and open sharing of data between institutions would allow researchers to develop and deploy prototypes to improve analytics. Evaluate analytics tools, alter processes based on new insights from analytics and deploy custom feature requests while enabling contextual insights into patterns in analytics on multiple datasets. The development of detailed analysis of user activities could derive features users want and need. Administrators would have clear representations of learner activities and would be able to track learner progress at the institutional level. The open-sharing of data across SoLAR members would permit administrators to analyze the effectiveness and efficiency of programs across institutions. Educators can compare current analytics with other faculty's anonymized datasets, other faculty data, data from similar courses in other universities. The content delivered to the dashboard would be customizable and personalized and track progress for a class or an individual student. The dashboard would have intervention triggers for educators to directly contact students when needed. The learner facing dashboard would provide feedback on progress, basic statistics, and tips on how to become a more

effective learner. The user-friendly toolset would allow access to social web tools like Google Scholar or Wikipedia. Siemens et al. [2012]

The goals of SoLAR are similar to that of LearnSphere in that they aim to make data and software sharing easier and commonplace, but differ slightly in that they are only looking to share between integrated institutions. A key difference between SoLAR and LearnSphere is the types of people they are looking to support. LearnSphere mostly supports researchers, while SoLARs intention is to support researchers right down to individual learners with their platform. SoLARs proposed platform is much more user-friendly and refined, though is still only theoretical while LearnSphere has a working infrastructure.

2.5.7 moocRP

moocRP is an open-source, web-based data repository and analysis tool with an integrated instructor-oriented dashboard. The tool integrates data requests and authorization and distribution workflow features while providing a straightforward format for uploading analytics modules enables reuse and replication of analytics results. The moocRP platform developed by UC Berkeley, and its built-in tools provide answers to the practical questions of how to prepare data, granting users secure access, and examination of the data for actionable information for MOOC instruction. Goals of the project are to address problem areas left unresolved by previous endeavors such as: duplication of analytics work across platforms, the lack of replication of data intensive research, and the divide between published studies and practical application Pardos et al. [2016] and to expand the impact of learning analytics by enabling its use.

With moocRP, the analytics can be brought to the data instead of the other way around. Individuals possessing data do not need to upload it to a centralized location to have analyses run but can instead spin up their own local instance of moocRP and import their desired analytic modules to run locally. The tool is open source, as is the technical design of the pipeline that serves analytics to its users. New analytics and visualizations that are uploaded must also be open-source to function on the system. Pardos et al. [2016]

moocRP is very similar to the MOOCdb model in LearnSphere in the way that software is shared but data is kept locally moocRP software is slightly different in the application of the dashboard which is aimed at course instructors, MOOCdb in LearnSphere is aimed more towards learning science researchers.

Stakeholder roles in moocRP are comparable to that LearnSphere and in its overall goals for the project. Pardos lists user privacy in data sharing as an area of concerns and believes this software sharing model of bringing analytics to the data will alleviate that concern for education researchers and institutions who want to conduct analytics research on data they curate.

2.6 Related Infrastructure and Repositories

For comparison we describe similar projects.

2.6.1 DataVerse

DataVerse is an open source research data repository for researchers, institutions, and developers to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows replication of others' work more easily. Researchers, data authors, publishers, data distributors, and affiliated institutions all receive academic credit and web visibility branding data to coincide with the associated research. DataVerse supports institutional sharing of data by establishing a research data management solution for their individual community. Institutions can participate to set norms for sharing, preserving, citing, exploring, and analyzing research data.

Administratively, DataVerse is funded by Harvard with additional support from the Alfred P. Sloan Foundation, National Science Foundation, National Institutes of Health, Helmsley Charitable Trust, IQSS's Henry A. Murray Research Archive, among many others. A collaboration of The Institute for Quantitative Social Science (IQSS) and the Harvard University Library and Harvard University Information Technology organization makes it openly available. IQSS leads the development of the open source DataVerse software with the Open Data Assistance Program at Harvard.

2.6.2 CKAN

The Comprehensive Knowledge Archive Network (CKAN) is a web-based open source management system for the storage and distribution of open data used by public institutions. CKAN is an international endeavor containing data across a variety of subjects from US, UK, AUS, and Netherlands.

The system is used as a public platform on Datahub, a data management system that makes data accessible by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers. CKAN presents a streamlined way to make data discoverable and presentable. Each dataset is given its own page for the listing of data resources and a rich collection of metadata, making it a valuable and easily searchable data catalogue.

CKAN mostly supports data sharing but also provides over 200 community extensions features and powerful features for data visualization. The web interface allows publishers and curators to easily register, update and refine datasets.

CKANs efforts are aimed at data publishers, institutions that are interested in publishing openly shared data. CKAN would also be a good resource for data scientists and researchers to find data from across disciples. CKAN does not note any privacy issues with sharing, data shared on the site has no identifying features as it is not linked to an individual.

CKAN is community supported and receives commercial support from Link Digital, Viderum, Opengov.

2.6.3 IMPACT

IMPACT is an

> Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program [that] supports the global cyber risk research community by coordinating and developing real world data and information sharing capabilities tools, models, and methodologies.

The Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT) program works as a portal serving as a community resource and marketing platform supported by the U.S. Department of Homeland Security, Science & Technology Directorate, Cyber Security Division. IMPACT coordinates, enhances and develops real world data, analytics and information sharing capabilities, tools, models, and methodologies by making data sharing components broadly available as national and international resources to support the partnership among cyber security researchers, technology developers and policymakers in academia, industry and the government.

The primary goal of IMPACT is to bridge the gap between producers and consumers of cyber-risk relevant data, tools and analytics to inform policy and analysis of cyber-risk and trust. The platform serves as a lab for testing various data sharing models including: batch transfers, data as a service, and visualization techniques enabling empirical data and info sharing between global cyber sec research and dev in academia, industry, and the government.

The distributed research data repo supported by streamlined legal framework, centralized coordination of controlled distribution of datasets, brokering and provisioning between data providers, data hosts, and researchers address the operation, trust and admin costs and challenges that impede data sharing. This model could serve to remove pressure about privacy concerns from institutions with government legal framework and methodologies.

IMPACT currently does not support data analytics features, but plans to support these functions in the future.

3 Software Donation

3.1 Software Donation in the Realm Beyond LearnSphere

3.1.1 Motivations for Software Donation

There are several incentives for a party to openly donate[2] software:

- Developer participation in a software base increases its user base.
- Development costs are shared by the community of users.
- Maintaining a pace of ongoing development is easier and keeps the code from "rotting".
- Students and professionals can hone and exhibit their skills publicly by contributing to the open source project which helps them build their reputations.
- It allows replication (when input data is included) and reproducibility with input data from others. Academic researchers are increasingly being encouraged or required to release their software with their publication because of these reasons.
- In some cases, software can be released with a Digital Software Identifier which allows citation.

3.1.2 Software Donation Control: Licenses

The most common way to share software is through licensing the terms of the softwares use. In some licensed software, users are not permitted to make changes to the software code and must use it in the way that is outlined in the licensing terms and conditions. Other licensing terms explicitly allow changes. Open source licenses are widely used and promoted by the Open Source Initiative (OSI). For OLI to approve open-source licenses, they must meet the strict standard of OSI's definition of open-source. Open-source licensing allows software to be freely used, modified and shared.

Some popular OSI approved open-source software licenses are:

- Apache License 2.0
- GNU General Public License (GPL)
- MIT license
- Mozilla Public License 2.0

[2] We use donate rather than share to make the direction of the contribution clear.

- Common Development and Distribution License
- Eclipse Public License

3.1.3 How to Share: Open-source Hosting Services

The common means of making open source software available is through a code repository platform such as GitHub, Bitbucket, or CodePlex, to only name a few. Any user can then make a copy of the code locally and make changes to the softwares code. The user can then share this new version of the software or keep it private.

Features	Github	Bitbucket	Sourceforge	Gitlab	Kiln	Codeplane	Code Plex	Beanstalk
Pricing*	Free	Free	Free	Free	$18/mo	$9/mo	Free	$15/mo
Private Repo	Paid	Unlimited, Free	Yes	Unlimited, Free	Paid	Unlimited, Paid	Unlimited, Upto 30 Days	10
Public Repo	Unlimited, Free	Unlimited, Free	Yes	Unlimited, Free	Paid	Unlimited, Paid	Unlimited	10
Storage Limit	1GB per repo	2GB	None	None	None	2GB	None	3GB
Users	Unlimited	5 & Unlimited if public	Collaboration not possible	Unlimited	5	Unlimited	Unlimited	5
VCS	Git, SVN	Git, Hg	Git, SVN, Hg	Git	Git, Hg	Git	Git, SVN, TFS, Hg	Git, SVN
Graphs	Yes	No	No	Yes	No	No	No	No
Web Hosting	Static sites. Page generator	Static sites	Dynamic Sites, CMS	Static	Yes	No	No	No
Code Review	Yes	Yes	Yes	Yes	No	No	No	Yes
Wiki	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Bug Tracking	Yes (Login Required)	Yes	Yes	Yes	Yes	Yes	Yes	No
Discussion Forum	No	No	Yes	No	No	No	Yes	No

***Free versions considered for this table. For services which have only paid versions, lowest priced versions have been considered.**

Figure 2: Open-source Hosting Services

Code repository platforms are free or paid by subscription (usually on a monthly or annual basis). Institutions purchase enterprise versions of platform software to develop, store, and share their own software and code with project members internally. The free services are not guaranteed to stay free, many have a free trial period for a determined period of time. It is not uncommon for these services to begin charging a fee after they gain a solid user-base. In one model the enterprise platform revenue cross-subsidizes the free platform. Commercialization of code repository platforms

for open sourced code represents some risk for a project that has no funds for supporting software donation.

Generally, platforms function similarly, but have differences in pricing and structure. For example, Github allows free unlimited public code repositories for all users. A monthly subscription is required to maintain private repositories. BitBucket is second to Github is its usage. BitBucket is a free service for individuals and organizations with 5 users or less. Teams can have unlimited users as long as their repositories remain public. CodePlex is Microsoft's free open-source hosting service. All CodePlex projects are private for a maximum of 30 days. After the 30 days the projects will be removed if they are not made public. There is no free to use any CodePlex features. CodePlex does not set a cap on the amount of users contributing to a project.

3.2 LearnSphere's Software Donation Model

To incentivize software donation the LearnSphere development team will highlight it. They are also investigating software digital identifiers which will standardize software citation by giving software a unique identifier. This would structure how users would cite th software in a publication.

LearnSphere has multiple software donation models. TIGRIS components can be saved on the LearnSphere server. Non-TIGRIS donations depend on the founder team:

DataShop has an External Tools repository. During the upload process the user is asked to provide the tool name, description, language and homepage. Once this is done the author can add files to be made available for download, if they choose. The tools that are uploaded are stored on the DataShop server, typically as a zip file, though in some cases we merely provide a link to the authors GitHub repository. Authors are responsible for updating their tools.

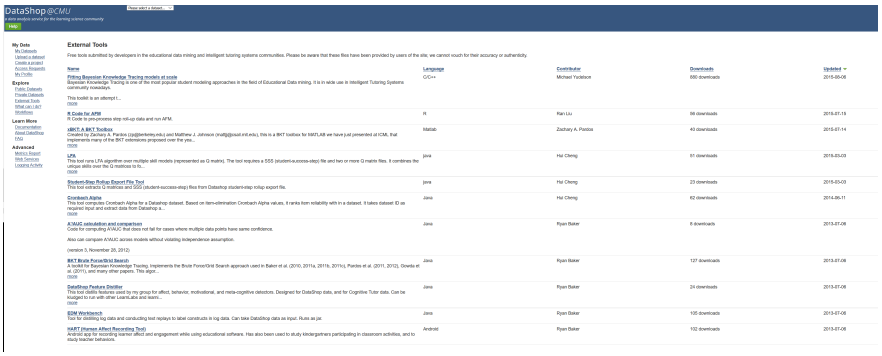

Figure 3: DataShop External Tools

DataShop provides a page with a list of available tools as well as a link for uploading a new tool. The tools themselves do not run within DataShop. Tigris is a computational environment in contrast. Components run within Tigris. In some ways, however, DataShop's model is a precursor to the workflow tool in LearnSphere in that we hope users will create and donate workflow components that integrate their tools/software.

DiscourseDB offers analytics components related to constructs including role taking, help exchange, collaborative knowledge construction, showing openness, taking an authoritative stance, attitudes, confusion, alliance and opposition. In enabling application of such metrics across datasets from multiple platforms, research questions related to the mediating and moderating effect of these process and state measures on information transfer, learning, and attrition can be conducted, building on the earlier research of the team.

MOOCdb and DiscourseDB currently function within LearnSphere via a portal model, i.e. LearnSphere points to their external resources. Both donate software via GitHub repositories and these will eventually be linked from LearnSphere. Work is in progress to integrate project capabilities within a demonstration of TIGRIS. For example, the MOOCdb curation pipeline within the project's Translation_software repository outputs tables or a populated database instance. Feature engineering through software in the DigitalLearnerQuantified repository creates features for modeling. These and/or extractions from the database can be exported in CSV format and then imported to TIGRIS as data.

Open Analytics and Research Service (OARS), a current project in development by Stanford may be eventually be linked to the LearnSphere platform. OARS is an extension of the Stanford open-education system that collects and models student learning data, interprets and presents information to instructors in a dashboard to guide instruction and class activities. The project proposes a human-centered design process to discover instructor needs for using the OARS dashboard to support instructors pedagogical decision-making for the classroom.

3.3 Survey Results on Software Donation

In the survey conducted with LearnSphere PIs and potential LearnSphere software donors, we asked for their individual perspective and experience in software sharing inside and outside of LearnSphere, where applicable, in an attempt to identify possible incentives and impediments software sharing. To clarify sharing issues, we asked participants to characterize the software they work with and outline their sharing experiences, challenges, privacy concerns, and requirements to clarify the software we hope to be shared though LearnSphere. We gauged the participants interest in accessing software shared by others and assessed the roles of those who would be interested in accessing software they shared. We asked participants about their specific circumstances in the possibility of sharing software with LearnSphere, to identify who controls this software, if anyone, to understand better ways to encourage sharing and their expectations of accreditation and licensing of software if they were to share with LearnSphere. This has allowed us to better understand privacy and sharing related impediments are faced around software sharing and how these translate into what they share and with whom.

Regarding software donation, we observed that researchers are open to the possibility of developing TIGRIS components. Only a select few will have the time and resources to develop highly reusable component so but these "super-user donors" are typical in open source systems. LearnSphere can use its portal model to publish links to donor software outside TIGRIS.

4 Data Donation

This section first covers what broadly motivates data donation and obstacles to donation. Next it describes data donation in LearnSphere, survey findings, action LearnSphere can take directly and indirectly.

4.1 Data Sharing: General Motivations and Obstacles

There are multiple motivations to donate scientific data. The government is motivated to make researchers who they fund donate data to one another because the data is a *public good*. Additionally, donation allows reproduction of scientific results. Reproducibility supports scientific integrity and ensures that scientific findings are objective. It supports the repeated demonstration of a scientific conclusion or leads to the revision and improvement of a method or conclusion. However, an obstacle to motivating donation by citing reproducibility is that there is little academic reward or opportunity to publish a replication or reproduction study King [1995] because novel methodology is valued more highly.

On a community level, another motivation is efficiency because data collection and curation takes a tremendous amount of time and effort. On the other hand, sharing doesn't help the research who expended these efforts, it requires altruism.

Motivations can be bolstered by incentives to donate. One incentive is to reward researchers with academic citation. Citation would come through replication studies or use of the dataset for another purpose: "there is a direct effect of third party data re-use that persists for years beyond the time when researchers have published most of the papers reusing their own data." Piwowar and Vision [2013]. Unfortunately, *currently there is no well defined way to cite donated datasets.* Ideally digital object identifiers (DOI) will solve this problem [3] but currently no citation mechanism has been formally standardized though different, not quite perfect, approaches are evolving. Open Researcher and Contributor ID (ORCID) provides a unique identifier for individual researchers to use with their name as they engage in research, scholarship, and innovation activities for use in citation. Alternatively, the FundRef initiative from CrossRef[4] helps funding institutions to report on research outcomes by collecting standardized research funding data from published works and making it available for anyone to search and analyze. Neither FundRef nor DOI provide citation attribution to a particular dataset. Goodman et al. [2014]. A common work-around is to cite the publication in which the dataset was used. This seldom occurs, however when the publication does not relate to the scientific question of the paper using the dataset. LearnSphere tracks this issue and will incorporate new practices around DOIs as they occur.

Risk of a privacy breach is a highly cited obstacle to e-learning data donation. Potential donors and data controllers are concerned that a learner will be identified. The consequences would range

[3] The implementation of a standard DOI is being considered by libraries to support data sharing and dataset citation. We contacted Dr. Micah Altman of MIT Libraries who we will stay in contact with on this topic.

[4] https://www.crossref.org/services/content-registration/

from discomfort to the learner, to legal suits, to loss of public confidence in an effort that includes data collection. There are examples of clear public discomfort with data collection and ethics surrounding data sharing. Reuters published a story in 2013 on parents' concern for privacy regarding a student database created inBloom as a joint project of the Gates Foundation and Carnegie Corporation. While local education officials had legal control over the student's data, parents did not need to consent for the data to be shared. inBloom's privacy policy stated that it "cannot guarantee the security of the information stored," causing fear of abuse of the data if it somehow leaked or stolen.

We did not find a specific example of privacy breach that brought discomfort to learners who were identified or a law suit. A useful follow to this report would be for LearnSphere to develop some examples of privacy protected data derived from privacy protection software.

Another useful follow to this report would be to develop one or more studies that create a synthetic dataset and demonstrate how it is useful. Synthetic datasets try to model the multi-variate distribution properties of the original data. Synthetic students are drawn from this distribution. While synthetic students won't help when instruction is going to be targeted, it will suffice for aggregate analysis. It is however a challenge is express the multivariate distribution with adequate accuracy. See "The Synthetic Data Vault : generative modeling for relational databases", a MIT M.Eng thesis by Neha Patki at `http://hdl.handle.net/1721.1/109616` for an example.

4.2 LearnSphere Facilitation of Data Donation

How or whether data is donated in LearnSphere depends on the founder project:

4.2.1 DataShop

Datasets in DataShop are categorized as *public* or *private* and donors choose a category for their dataset. Anyone logged in can access public datasets. Private datasets can only be accessed with permission from the Principal Investigator (PI) of the project. If a researcher wishes for the data to be public, DataShop verifies it as shareable, in keeping with the IRB and any IRB relevant to the collection of their data. Before a project has been determined to be "sharable," there is no ability to make it public, nor can it be shared outside of a research team. When importing data, all student identifiers are anonymized. Data is stored on the LearnSphere server or linked from LearnSphere.

This approach has been very successful. Many datasets are available with public designation and others that are private are available on a by-request basis. The success is due, in part because in many circumstances, the data has been collected as part of a research study so proper permissions are in place. The researcher collecting or working with the data controlled its donation policy when the circumstances of collection were established with an IRB. De-identification or breach of privacy is not a risk. It is this extra level of security that is also responsible for why researchers are willing to put their data into DataShop.

4.2.2 DiscourseDB and MOOCdb: Data Access but not Data Donation

The DiscourseDB work builds on a foundation of research in analytics applied to text that demonstrate the potential to automatically detect learning relevant discussion behaviors and states such as Help exchange Cui et al. [2009], Gweon et al. [2007], Collaborative Knowledge Construction Ai et al. [2010], Gweon et al. [2013], Mu et al. [2012], Rosé et al. [2008], Openness and Authoritativeness Howley et al. [2012], Attitude Wen et al. [2014b], Cognitive Engagement and Expressed Motivation Wen et al. [2014a], Coordinated Activity Kumar and Rosé [2014], Yang et al. [2014], and Confusion Yang et al. [2015]. Most of these measures have been developed using a publicly available text mining tool bench called LightSIDE Mayfield and Rosé [2013], which has been set up to work along side DiscourseDB. That tool bench is open source and has been used over nearly a decade by over ten thousand users internationally. The DiscourseDB pipeline, consisting of the database itself, a data browser, and annotation tool, and LigthSIDE are meant to streamline the data-to-analytics process.

The founders of DiscourseDB demonstrate its software and generate its research results with data from two MOOC courses. The MOOC students were required to sign an online release form allowing DiscourseDB to use their data for research purposes.[5] The data is controlled by CMU, not the DiscourseDB researchers, and stored in a *private* instance on GitHub. For other researchers to use this data, they must get IRB approval through a process at CMU. This process generalizes to a model of university-controlled data access where, unless data is completely public and deemed to be exempt from IRB approval, the university (i.e. its IRB) from which the data comes must approve donation. So, for example, if a researcher wanted to use DiscourseDB software with MITx data, they would go through the process of obtaining it with MIT and they would not be able to donate the data (or a derivative dataset) to others unless MIT gave them permission.

The same situation holds for MOOCdb where the founders have demonstrated the software framework with MOOC data from 2 different platforms (Coursera and edX) and multiple MOOCS, e.g. Stanford and MITx. If, for example, a researcher wanted to use MITx data, they would go through The MIT Institutional Research group in the Office of the Provost who coordinates an application. The application and ensuring data use agreement demands compliance with student privacy regulations and storage requirements. The application process is stated here: `http://web.mit.edu/ir/mitx/`. Requests for both de-identified and identified data (subject to FERPA) are accepted. A researcher must be accredited and have a IRB approved study plan. Researchers are expected to be conducting research to improve teaching and curriculum or contributing to scholarship on teaching and learning. Appropriate IRB and data handling training and project documentation are required.

It is somewhat because of this indirect chain of control that DiscourseDB and MOOCdb support a data access model rather than a data donation model, pioneered by MOOCdb, where any researcher can standup their own private instance of the project software, enabling them to run the methods in the frameworks on their own data privately creating sequestered datasets. Both DiscourseDB and MOOCdb have curation software that converts data to expected input formats to further enable this. It is unclear at this time if the datasets created within these instances can or will ever be

[5] Whether they gave their consent with much thought is unknown and likely debatable.

shared. If the researchers are not publishing results (e.g. they are investigating for institutional purposes), none of the incentives and motivations apply. It would take the university or researcher being instructed by the government to make the data more accessible for donation to occur. If the researcher wants to donate the data but the institution is not compelled to donate it, donation depends on the institution's whim.

In addition to side-stepping data donation, this model of data access very strongly protects learner privacy because data never has to be transferred out of its home institution. To gain access to data, a researcher from another institution can pass software, e.g. from LearnSphere, to the data controller and the controller then executes the software over their institution's data and shares only the results. The controller may inspect the results and the software to ensure the data providers privacy remains protected. The data access model effectively promotes cross-institution access without donation and is enabled by the curation software allowing translation to a common data schema. One weakness of the data access model is that requests to run software are granted at the will of the data controller. Another is that curation software has to be updated whenever platforms introduce new types of data capture.

The model has a historical explanation. When data donation was not forthcoming from big institutions like MIT, this access model exploited a software solution (a common data schema for all the datasets) to work around this apparently intransigent policy. Whether it enabled further intransigence once the technology work-around became available is debatable. The impact of the access model being supported however may not be as powerful as the institutional will that was at play. As well, it is debatable whether this access model is problematic. If the data controllers, i.e. academic institutions with MOOC data, feel the data is too sensitive to donate but have set up a reasonable means by which researchers from accredited institutions can be granted access, it furthers a practical compromise. One weakness is that institutions potentially have the power to grant access but may be slow or reluctant to do so slowdown may be due to legitimate around verification and administrative costs and no incidents of reluctance are known to us. MIT, for example, has expressed openness and willingness, if a researcher has an IRB approved project, to transfer data under controlled use (via its data use agreement). Nonetheless, it is possible that institutions will experience a host of concerns around granting data access including potential embarrassment to instructors and instructor dislike for it.

One modest follow up action to this report would be to set up software that allows LearnSphere (or the projects themselves) to track how many private instances of DiscourseDB or MOOCdb have been cloned and to provide a short simple question set on what private datasets have been examined, what access policy controls them and what institution they come from so that an inventory is available.

Another helpful follow up action is to provide clear, easy portal-like access to the CMU and MITx data use request procedures from LearnSphere (or the projects themselves). When a researcher publishes software through the LearnSphere portal, they could also be encouraged to provide a link to the data use request procedure from the institution who supplied their data when they demonstrated the software or published on the project.

4.2.3 DataStage

DataStage (http://datastage.stanford.edu/) is a portal hosted by Stanford University outlining different datasets of Stanford platforms. Data obtainable through the portal is controlled by the Office of the Vice Provost for Online Learning of Stanford University and/or Stanford Center for Advanced Research through Online Learning (CAROL), see https://iriss.stanford.edu/carol/resources-researchers. The website documents protocols for accessing learner data, describes technical details about table schemas and other metadata useful for data analysis, and provides practical guidance on using analysis tools with the system.
To request access to learner data all researchers must demonstrate completion of training in human subjects research ethics and complete and online data request form [6]. The form accommodates requests for data with personally identifiable information which Stanford only provides such access in special cases, e.g. instructors of a class. Stanford requests published results to be credited for the use of their data to help them justify the cost of data collection. There is also a data use agreement, see https://stanford.box.com/datauseagreement.

4.2.4 LearnSphere

For LearnSphere, a data donation vs data access model depends on whether the data has been initially collected as part of a research study (easier to donate) or whether it was collected collaterally in some teaching activity and the research is being done retrospectively (up to the teaching institution to donate).

Learnsphere provides 3 categories: *private, shareable, public*, of data access control at time of data import. A human managed, software assisted check of the IRB and other permissions determines whether a dataset can be shareable or public. If the data is categorized as *private*, it is only accessible to the owner and those they specify as having rights to access. If it is *public*, it is directly accessible. If it is *shareable*, a click on the dataset initiates an email to the owner who may decide whether or not to grant access.

To support the access model for MOOC data (see 4.2.2) LearnSphere is considering how to design a means for an access-donor to provide access routing to others when the donor creates a TIGRIS import module. The (new) category of this data would be *shareable*. A click on the dataset name (when the import module is used) would set up and dispatch a request for the data. Granting the request entails a manual check of the IRB and other permissions.

LearnSphere is not intended to provide generic tools for data privacy protection. Founders are considering whether tools or demonstrations of privacy protection specific to its domain should be added.

LearnSphere offers no tools to create a synthetic dataset (see 4.2.2).

Given the merit of data access model and importance of methodology access, LearnSphere should

[6] Web form here: http://carol.stanford.edu/research and PDF here: https://stanford.app.box.com/s/w3eqcikxje05hm2b9ovn366psjyef3t5

also incorporate metrics for software and data access facilitation into its self-assessment of use.

4.3 Survey on Data Donation

We conducted a survey of LearnSphere PIs and, later, a small number of potential LearnSphere data donors at LAK 2017. Our surveys asked for participants' specific perspective and experience in data and software sharing inside and outside of LearnSphere. In an attempt to identify possible incentives and impediments, we asked participants to characterize the data they work with and outline their data sharing experiences, challenges, privacy concerns, and requirements. We asked about any specific circumstances around the possibility of sharing data with LearnSphere. We asked who controls the data they work with. We asked about accreditation and licensing of the data and software. This has allowed us to better understand privacy and sharing related impediments are faced around data and software sharing and how these translate into what they share and with whom.

The survey has a small number of participants because outreach efforts for LearnSphere use are still ramping up. A modest follow up would be to add more data to the survey once more researchers have had hands on experience with LearnSphere projects.

We asked those surveyed the reasons why others may not want to donate their data. Almost all those surveyed cited "Concerns for privacy", "Original conditions of data collection", and "institutional control" as reasons why others do not share their data. We can consider each of these reasons in turn.

- *Concerns for Privacy:* Embarrassment, trust issues, and damaging are just a few examples of the risks for loss of privacy. All but 2 participants noted a concern for privacy but, despite this concern, not a single participant listed any concrete examples of the risks for loss of privacy. It is plausible that participants (and institutions that control the data) perceive risks that are much higher than the actual risks. We won't know this for sure unless there are means in place for those who have lost privacy related to their shared data to report it. As well, participants may have been responding on behalf of their institutions' interest, rather than their own. Since privacy loss is perceived to be more harmful than the advantages of data donation without explicit authorization, it may be that this balance will encourage access control policies to stay as they are.

- *Original conditions of data collection* Here the researchers' hands are tied by original agreements. LearnSphere can't directly control new terms of data collection but a modest follow up action would be to make visible the need for new terms and to suggest some templates for researchers to use when they initiate their study and data collect.

- *Institutional control* It would appear the power to openly donate data is in the hands of the institutions who are collecting data, not the researchers engaging LearnSphere.

4.4 Follow Ups for LearnSphere

There are enough motivations and incentives that LearnSphere is going to be successful in securing data donation from any researcher who has authority to donate. LearnSphere can further subtly encourage it by:

- Tracking the private instances of DiscourseDB or MOOCdb that have been cloned and to provide a short simple question set on what private datasets have been examined, what access policy controls them and what institution they come from so that it can publish an inventory.
- Making citation easy: continually update itself with latest DOI technology.
- Releasing privacy protection software and demonstration.

When the researcher wants to donate but does NOT have authority, LearnSphere can:

- Provide easy to find URL links to data use request sites of the institution controlling the data. Support these being embedded in TIGRIS import components.[7]
- Offer tools the researcher can use to generate a synthetic dataset from the withheld data.

When the data controller does NOT wish to donate, LearnSphere can:

- Offer tools that generate a synthetic dataset from the withheld data.
- Offer tools for privacy protection software.

There are also indirect or broad actions LearnSphere can take:

LearnSphere should support building a dialectic relationship around policies and technology methods related to data privacy protection with stakeholders who donate, provide, and control behavioral data. It should encourage data controllers to become more informed about the central practicalities and risks in assuring online data privacy. It should empower them; and through them, their institutions, to more comfortable extend access to online data to researchers and teachers who can use it to further data driven discovery and innovation in education.

It should encourage *Data Sharing by Design*. It is also arguable that promising students and junior researchers who have matriculated in a more technology oriented culture will be eager to contribute new research methods and modeling software with demonstration data from projects they have initiated right back to deciding what data to collect.

It should encourage the government to oblige the researchers to donate.

[7] Efforts are underway already using a Georgia Tech Psychology.

4.4.1 Possible Risks Ahead

Going forward, there are risks:

- Some data won't be from MOOCs. It will be from blended courses. This data is more at risk for privacy breach.
- Some data won't be from MOOCs. It will be from private courses only open to a university's students and the university may weigh the sharing benefits to be less than the discouragements: time, effort, loss of face, internal pressures not to do so.
- Other universities may not follow MIT and Stanford's lead. They may rate privacy breach risk too high. They may not want to pay for the effort to provide access. They may have internal pressures not to share (in case they look bad).
- Because incentives are in the wrong place.The entity that can grant data control (dub "primary controller") has little incentive to make it easy for the researcher to release a dataset into the infrastructure. It costs the data controller resources in time and money and only their public-spirited side is rewarded. Therefore, even if a researcher is required or incentivized to share data, say by a commitment to a funder or desire to be cited, they will face institutional hurdles and bureaucracy which place a sharing burden on them.

5 Summary

We have examined whether LearnSphere's users will donate their software to each other. **We can expect with high confidence that users will donate software**. Current attitudes, incentives, requirements and support for contributing software are ideally aligned to foster this. **A similar singular conclusion about data donation does not exist.** While we have identified motivation and incentives for it, **donation will only be explicitly possible when a researcher controls the data they want to donate**. Nonetheless **LearnSphere can facilitate a researcher gaining access to the data**. We call this **a data access model** rather than donation. For either case, there are explicit actions LearnSphere is taking or can take up.

References

Hua Ai, Marietta Sionti, Yi-Chia Wang, and Carolyn Penstein Rosé. Finding transactive contributions in whole group classroom discussions. In *Proceedings of the 9th International Conference of the Learning Sciences-Volume 1*, pages 976–983. International Society of the Learning Sciences, 2010.

Yue Cui, Rohit Kumar, Sourish Chaudhuri, Gahgene Gweon, and Carolyn Penstein Rosé. Helping agents in vmt. In *Studying virtual math teams*, pages 335–354. Springer, 2009.

Alyssa Goodman, Alberto Pepe, Alexander W. Blocker, Christine L. Borgman, Kyle Cranmer, Mercè Crosas, Rosanne Di Stefano, Yolanda Gil, Paul T. Groth, Margaret Hedstrom, David W. Hogg, Vinay Kashyap, Ashish Mahabal, Aneta Siemiginowska, and Aleksandra Slavkovic. 10 simple rules for the care and feeding of scientific data. *CoRR*, abs/1401.2134, 2014. URL http://arxiv.org/abs/1401.2134.

Gahgene Gweon, Carolyn P Rosé, Emil Albright, and Yue Cui. Evaluating the effect of feedback from a cscl problem solving environment on learning, interaction, and perceived interdependence. In *Proceedings of the 8th iternational conference on Computer supported collaborative learning*, pages 234–243. International Society of the Learning Sciences, 2007.

Gahgene Gweon, Mahaveer Jain, John McDonough, Bhiksha Raj, and Carolyn P Rosé. Measuring prevalence of other-oriented transactive contributions using an automated measure of speech style accommodation. *International Journal of Computer-Supported Collaborative Learning*, 8 (2):245–265, 2013.

Iris Howley, Elijah Mayfield, and Carolyn Penstein Rosé. Linguistic analysis methods for studying small groups. *The international handbook of collaborative learning*, 2012.

Y. Jo, G. S. Tomar, O. Ferschke, C. P Rose, and D. Gaesevic. Pipeline for expediting learning analytics and student support from data in social learning. *Proceedings of Educational Data Mining*, EDM 2016.

Gary King. Replication, replication. *PS: Political Science and Politics*, 28:444–452, 1995. URL http://j.mp/2oSOXJL.

Rohit Kumar and Carolyn P Rosé. Triggering effective social support for online groups. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(4):24, 2014.

E Mayfield and CP Rosé. Lightside: Open source machine learning for text accessible to non-experts. invited chapter in the handbook of automated essay grading, 2013.

Jin Mu, Karsten Stegmann, Elijah Mayfield, Carolyn Rosé, and Frank Fischer. The acodea framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning*, 7(2):285–305, 2012.

Zachary A Pardos, Anthony Whyte, and Kevin Kao. moocrp: Enabling open learning analytics with an open source platform for data distribution, analysis, and visualization. *Technology, Knowledge and Learning*, 21(1):75–98, 2016. ISSN 2211-1670. doi: 10.1007/s10758-015-9268-2. URL http://dx.doi.org/10.1007/s10758-015-9268-2.

P I Pavlik Jr, C Kelly, and J K Maass. The mobile fact and concept training system mofacts. In A Micarelli and J Stamper, editors, *Proceedings of the 13th International Conference on Intelligent Tutoring Systems*, pages 247–253, Switzerland, 2016. Springer.

Heather A. Piwowar and Todd J. Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, 2013. doi: 10.7717/peerj.175. URL https://peerj.com/articles/175/.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271, 2008.

Carolyn P Rose. Expediting the cycle of data to intervention. *Learning: Research and Practice*, 3.1:59–62, 2017.

G. Siemens, D. Gasevic, C. Haythornwaite, S. Dawsone, S.B. Shum, R. Ferguson, E. Duval, and R.S.J.D. Verbert, K.and Baker. Open learning analytics: an integrated and modularized platform. *White Paper: Society for Learning Analytics Research*, 2012. URL http://www.solaresearch.org/wp-content/uploads/2011/12/OpenLearningAnalytics.pdf.

J Stamper, K Koedinger, PI Pavlik Jr, C Rose, R Liu, M Eagle, and K Veeramachaneni. Educational data analysis using learnsphere workshop. In J. Rowe and E. Snow, editors, *Proceedings of the EDM 2016 Workshops and Tutorials co-located with the 9th International Conference on Educational Data Mining*, Raleigh, NC, 2016.

John C. Stamper, Kenneth R. Koedinger, Ryan S. J. d. Baker, Alida Skogsholm, Brett Leber, Sandy Demi, Shawnwen Yu, and Duncan Spencer. *DataShop: A Data Repository and Analysis Service for the Learning Science Community (Interactive Event)*, pages 628–628. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-21869-9. doi: 10.1007/978-3-642-21869-9_129. URL http://dx.doi.org/10.1007/978-3-642-21869-9_129.

Kalyan Veeramachaneni, Sherif Halawa, Franck Dernoncourt, Una-May O'Reilly, Colin Taylor, and Chuong Do. Moocdb: Standards and systems to support mooc data science. *arXiv preprint arXiv:1406.2015*, 2014.

Kalyan Veeramachaneni, Franck Dernoncourt, Colin Taylor, Zachary A. Pardos, and Una-May O'Reilly. Moocdb: Developing data standards for mooc datascience. *MOOCShop at Artificial Intelligence in Education*, 2015a. URL http://ceur-ws.org/Vol-1009/0104.pdf.

Kalyan Veeramachaneni, Sherif Halawa, Franck Dernoncourt, Una-May O'Reilly, Colin Taylor, and Chuong Do. Moocdb: Developing standards and systems to support mooc data science. *arXiv #1406*, 2015b. URL https://arxiv.org/abs/1406.2015.

Miaomiao Wen, Diyi Yang, and Carolyn Rose. Sentiment analysis in mooc discussion forums: What does it tell us? In *Educational data mining 2014*, 2014a.

Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. Linguistic reflections of student engagement in massive open online courses. In *ICWSM*, 2014b.

Diyi Yang, Miaomiao Wen, and Carolyn Rose. Peer influence on attrition in massively open online courses. In *Educational Data Mining 2014*, 2014.

Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 121–130. ACM, 2015.