Abstract

We propose a project centered at the University of Michigan (U-M) to study how libraries impact learning. We define learning as the process of acquiring knowledge and/or skills through formal study or instruction (course instruction), or experientially through laboratory research or clinical activities (research). However, learning is also the outcome of this knowledge acquisition process, including the ways or means by which this knowledge is disseminated to society (publishing). We use analytics – the discovery, interpretation, and communication of meaningful patterns in data (https://en.wikipedia.org/wiki/Analytics) – to address the research question: *How does the library impact learning, specifically in the areas of course instruction, research, and publications?* Libraries should strive to improve learning outcomes for the communities they serve, just like how healthcare institutions should ideally improve individuals' health outcomes. This type of work is best described as library learning analytics (LLA), which entails embedding library data within institutional learning analytics ecosystems, and is guided by a more encompassing definition of learning that extends beyond the classroom. Our goal is to establish the groundwork for a group of diverse institutions to use a common analytics framework. The three-year project, which will commence in June 2018, has two main goals as follows:

Goal 1: **To understand how the U-M library impacts learning, specifically in the areas of course instruction, research, and publications.** We plan to achieve this goal through modeling and analyzing both deidentified and identifiable library use data including: website server logs, library catalog server logs, proxy server logs, circulation history and related data, and campus status & affiliation data. These datasets are linkable through both strong and unambiguous identifiers that are unique to each individual user, as well as IP addresses and timestamps. First, using deidentified data, we aim to apply clustering algorithms to identify typologies of library users on the basis of library interactions e.g. user type A (high degree of use), user type B (low degree of use), etc. Second, we aim to replicate the clustering analysis using the identifiable data, using the deidentifiable clustering findings for robustness checks. Third, our goal is to use clusters from identifiable data in the sequence mining, and predictive and prescriptive modeling of the links between learning outcomes and library user types. For privacy reasons, results will be shared using only aggregated and anonymized data.

***Goal 2*: To develop tools, scripts, and protocols that serve the library analytics needs of our community of project advisory group (PAG) institutions.** We plan to achieve this goal through two modes of engaging with the PAG community. First, we plan to develop and share with PAG institutions *data dictionaries* (names, definitions, and attributes about data elements that are in a database) of the datasets we will be using in the project. This will facilitate data harmonizing across PAG institutions (and with U-M data) and thus smooth the exchange of ideas and best practices over the course of the project. We will maintain a repository of tools and scripts developed for data cleaning and database construction and grant PAG members early access to this repository. We also plan to regularly update PAG members via email on the latest status of data management and analysis activities with a view to making it easier for them to voluntarily replicate our work or perform their own LLA studies. Thrice a year, we will hold virtual research meetings with PAG members for more thorough debriefs of project activities. Lastly, throughout this process we will stay in regular phone and email contact with PAG members in order to help them resolve replication issues as they arise. Second, we will give PAG members access to aggregated and anonymized findings (from *Goal 1*, using U-M data) through a web-based dashboard running on data that will be located on a secure virtual data enclave. Further, in addition to the dashboards, we will hold annual PAG Workshops where member representatives will be tutored and immersed in training tasks to ease voluntary replication of the project, but using their own data.

Impact: This project will provide guidance to PAG institutions and other libraries on how to best design and implement empirical, holistic LLA studies of the links from library usage, to learning outcomes such as course instruction, research, and publications. The project will produce a set of tools, scripts, and protocols that will be freely available to all libraries. The study will serve as a template for other libraries with respect to: 1) studies that collect and store library use data with individual identifiers while maintaining the privacy of individuals; 2) designing and implementing a holistic LLA study of the link from library use to multiple learning outcomes, and; 3) creating a secure cyberinfrastructure (data repository, virtual enclave, and dashboard) for LLA research that facilitates collaboration in a community of diverse institutions in the US and Canada.

A. Statement of National Need

We propose a project centered at the University of Michigan (U-M) to study how libraries impact learning. Our project aligns with the NLG program's focus on libraries as community anchors by studying how libraries impact learning in two types of communities. ***First***, the project focuses on the local community served by a library. For example, the U-M library anchors learning for a diverse community of: a) U-M's faculty, students, and fellows; b) members of the public and academic visitors (who are welcome to use resources in-building); c) reciprocal borrowing with Ann Arbor area academic institutions; d) peer institutions in the U.S. e.g. the Big Ten Academic Alliance; and e) professions e.g. the U-M Law Library is open to all Michigan Bar members. ***Second***, the project is intended to use U-M as a "training" case study to help other libraries empirically assess and predict their impact on learning outcomes. The project is expected to catalyze the analytics capacity of a community and network of library institutions. At the core of this community is a diverse set of faculty and professionals from 15 libraries (higher education and public) and professional associations that collectively constitute the project advisory group (PAG). The project team will be guided by the PAG community in developing *computational tools* (calculations and algorithms for database and data management, presentation, and extracting meaning from data), *scripts* (source codes and programs), and *protocols* (written, predefined, and standardized procedures) that are usable across different institutional settings. The PAG will guide the project to ensure the study's replicability across institutional settings. In turn, PAG members will have first access to all tools, scripts, and protocols developed for the project including the dashboard and other data visualizations such that they should be able to carry out their own analytics on how their libraries impact learning. Our goal is to establish the groundwork for a group of diverse institutions to use a common analytics framework, which we expect would even facilitate voluntary replication at and benchmarking across PAG institutions. We use analytics – the discovery, interpretation, and communication of meaningful patterns in data (https://en.wikipedia.org/wiki/Analytics) – to address the research question: *How does the library impact learning, specifically in the areas of course instruction, research, and publications?*

Beneficiaries

The library will be better placed to have a fuller sense of how it is impacting learning outcomes for the community of individuals that it serves, enabling it to improve service provision including the addition or removal of specific services as needed. For example, the establishment of a link between specific library activities and learning outcomes, for example research, would help the library decide whether it would be worthwhile to offer instruction or coaching with respect to the lifecycle of research projects.

A survey of academic libraries in the UK revealed that 94.6% of them were interested in anonymized benchmarking with other institutions, and 87.7% were interested in analytics that assess how libraries impact learning outcomes (Showers, 2014, 2015). The PAG community has disparities with respect to the resources that are available for analytics. This project would address this problem by developing computational tools, scripts, and protocols that will be made available to PAG institutions (and others) at zero cost. Further, the project team plans to provide free consultation to any PAG institution interested in doing their own analytics (replicating any or all of the analytics employed in the study).

Theory

Learning analytics (LA) is an emerging discipline that uses approaches such as data mining, machine learning, natural language processing, and visualization in order "to provide educators and learners with insights that might improve learning processes and teaching practice" (Lang, Siemens, Wise, & Gašević, 2017). Drawing on the multiple ways libraries interact with the communities they serve, our definition of learning is richer and extends beyond classroom outcomes. We define learning as the process of acquiring knowledge and/or skills through formal study or instruction (course instruction), or experientially through laboratory research or clinical activities (research). However, learning is also the outcome of this knowledge acquisition process, including the ways or means by which this knowledge is disseminated to society (publishing). Libraries should strive to

improve learning outcomes for the communities they serve, just like how healthcare institutions should ideally improve individuals' health outcomes. Libraries collect rich streams of data that have been used for administrative purposes. *Library analytics* has hitherto meant the use of data to understand, analyze and visualize patterns of library activities such as checkouts (https://osc.hul.harvard.edu/liblab/projects/library-analytics-toolkit). There is a growing recognition of the need to use analytics to assess the impact of the library on the communities it serves (Showers, 2015). To date, most efforts to understand the library's impact have been correlational e.g. with respect to improving learner outcomes. More sophisticated models would enable us to unlock the true potential of the data held by libraries. This type of work is best described as library learning analytics (LLA), which entails embedding library data within institutional learning analytics ecosystems, and is guided by a more encompassing definition of learning that extends beyond the classroom. The best way to tap into diverse resources and perspectives in the development of better explanatory, predictive, and prescriptive models of the library's impact is through networks of organizations that constitute agile LLA communities. The Library Impact Data Project (LIDP; https://library.hud.ac.uk/blogs/lidp/) is an example of a network of LLA organizations, albeit in the UK (Stone & Ramsden, 2013). Critically, LIDP was a step in the right direction with respect to establishing common, shared standards for transactional and usage data, a prerequisite for explanatory, predictive, and prescriptive LLA. Importantly, the type of work performed and promoted in the LIDP project has yet to be done or replicated in the United States. We expect this project to use the better practices identified by the LIDP project, but also incorporate more advanced analytical methodologies that should enable us to construct robust explanatory, predictive, and prescriptive models of the library's impact.

Why Analytics

The potential of data analytics and mining has transformed entire fields such as industry, healthcare, and science, and is an emerging and increasingly important field in the education sector. Tied to this wave of advancements in data science is a cognizance of the need and moral imperative for institutions to better respond to the needs of their clients. For example, in healthcare, advancements in analytics have been one of the key propellants for the rise of precision medicine, an approach to disease prevention and treatment that accounts for individual differences in genetics, environments, and lifestyles. Precision medicine is far superior to the traditional one-size-fits-all approach where disease prevention and treatment strategies are developed with the "average" person in mind, without regard for individual differences with respect to gender, race/ethnicity, age, and physical environment. The moral imperative to better address the needs of the individual, rather than the "average" person, is behind the rise of LA in higher education for uses such as generating predictive models to help instructors identify students at risk of failing a course early enough in an academic term (see for example *Student Explorer* at U-M, which is an early-warning system to help academic advisors identify at-risk students: http://ai.umich.edu/portfolio/student-explorer/).

Over the past two decades, technology has brought significant sustained change to both the availability of information and the way library users seek it. At the same time, the library has been under pressure to articulate and document its value. In higher education, that means the library's impact on learning outcomes like course instruction and research productivity (Oakleaf & Association of College and Research Libraries, 2010). Some empirical analyses have correlated library activities to learning outcomes in academic and school libraries. Correlational studies, however, do not generate causal findings and are therefore limited with respect to shaping library programs and potential interventions. Correlational studies have also narrowly focused on one or a couple of outcomes, such as student attainment. In fact, the library is embedded in different dimensions of learning in communities. We need empirical studies that can provide a holistic look at how the library impacts learning in order to develop robust metrics for the library's true value. For example, there has been a gradual shift from libraries reporting simple use statistics, to more robust studies of how the academic library impacts the university especially following the publication of the report "The Value of Academic Libraries" (Oakleaf & Association of College and Research Libraries, 2010). This report identified several aspects of a comprehensive research agenda on the value of an academic library, including instruction and student retention, faculty research productivity, and faculty grants (Hinchliffe, Oakleaf, & Davis, 2010; Oakleaf & Association of College and Research Libraries, 2010). A key challenge to developing holistic, empirical studies of the library's impact

is the management, protection, and analysis of the small, large, and "Big" datasets that need to be created by linking library data to outcomes and other institutional data. Libraries have been under-resourced with respect to the capacity to conduct research projects that are not just correlational, but that are explanatory, predictive, and prescriptive. This is important as libraries, including school libraries, would benefit from causal studies as they are more likely to result in targeted and effective interventions. Advances with respect to understanding the impact of the library on learning could be catalyzed by exploiting the diverse resources and perspectives of networks of communities in the library field. This project is designed to advance library knowledge and practice by improving our understanding of the library's impacts on course instruction, publications, and research. The project is also intended to lead to the development of tools, scripts, and protocols – for explanatory, predictive, and prescriptive analysis – that could be accessed and used by a broad range of libraries.

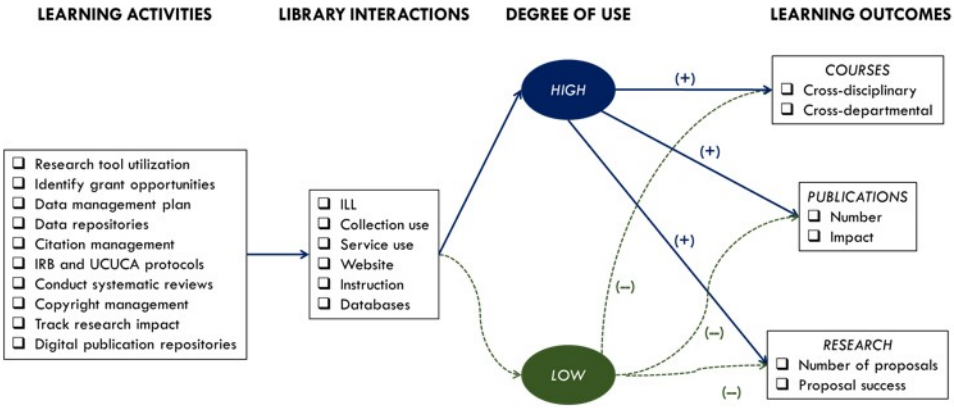

Figure 1. Impact of the library on learning

Study Framework

Our proposed study examines how library services (e.g. ILL, circulation, websites, instruction, databases, etc.) impact learning outcomes (course instruction, research, and publishing). Significantly, the study enables the testing of multiple pathways for associating the use of a specific library service or resource, to learning outcomes (Figure 1). With this focus on paths, we can begin to account for individual differences with respect to usage of a specific type of service, and its potential ramifications for learning outcomes.

Privacy Concerns

We are cognizant that despite its enormous potential to help us gauge how libraries contribute to learning, LLA also raises significant concerns with respect to protecting the confidentiality and privacy of the communities it is intended to help, and data security more broadly. Thus, we plan to employ best practices from research in LA, medical, and social science fields with respect to protecting the privacy and confidentiality of individual data. For example, we used confidential human resource and grant management data under the guidance of rigorous institutional review board (IRB) protocols to study the factors of collaboration in research (Kabo, Cotton-Nessler, Hwang, Levenstein, & Owen-Smith, 2014). In this study, we only published anonymized and/or aggregated findings such that individuals could not be identified. The IRB protocols also governed who had physical and electronic access to the data, and the security of the data storage environment. We expect to copy the best practices analogous to how medical data is used to make key discoveries (e.g. gender and racial health disparities) without violating the Health Insurance Portability and Accountability Act (HIPAA) of 1996. HIPAA regulates data privacy and security issues regarding medical information. Libraries have for good reasons prioritized the privacy and security of user data, removing individual identifiers from library use data in order to protect the privacy of their users (Coombs, 2004). This practice has, however, had the unintended consequence of hampering libraries from using LLA to assess how they impact learning outcomes. Libraries are now poised to take the lead with respect to how privacy and security concerns can be addressed in frameworks that permit LLA. The U-M Library recently revised its privacy policy such that it is able to balance its twin moral obligations of improving learning outcomes for the communities it serves, while at the same time ensuring that data security and privacy concerns are addressed robustly (see Appendix 2). This proactive approach mirrors those in the LA community with respect to using student data in the context of Family Educational Rights and Privacy Act (FERPA) restrictions (Brooks & Thompson, 2017; Zeide, 2017). Lastly, libraries need to realize that one of the biggest drivers for the need for LLA is the community of individuals that they serve. Individuals

have become increasingly accustomed to personalized and data-driven experiences in their day-to-day lives (Showers, 2014). If data could be used to improve learning outcomes, then individuals want to know why institutions are not using the data for this purpose (Showers, 2014). Libraries should take the lead on the key issue of how to deploy analytics to improve learning outcomes, but without giving privacy concerns short shrift.

B. Project Design

Project Goals & Outcomes

The project is committed to studying and measuring the links between how individuals interact with libraries, and how these interactions impact learning – the process of acquiring knowledge and/or skills through formal study or instruction (course instruction), or experientially through laboratory research or clinical activities (research). However, learning is also the outcome of this knowledge acquisition process, including how this knowledge is disseminated to society (academic publishing). The project has two main goals as follows:

Goal 1: **To understand how the U-M library impacts learning, specifically in the areas of course instruction, research, and publications.** We plan to achieve this goal through modeling and analyzing both deidentified and identifiable library use data including: website server logs, library catalog server logs, proxy server logs, circulation history and related data, and campus status & affiliation data. These datasets are linkable through both strong and unambiguous identifiers that are unique to each individual user, as well as IP addresses and timestamps. First, using deidentified data, we aim to apply clustering algorithms to identify typologies of library users on the basis of library interactions e.g. user type A (high degree of use), user type B (low degree of use), etc. Second, we aim to replicate the clustering analysis using the identifiable data, using the deidentifiable clustering findings for robustness checks. Third, our goal is to use clusters from identifiable data in the sequence mining, and predictive and prescriptive modeling (e.g. statistical and computational models) of the links between learning outcomes and library user types. These analyses will control for potentially confounding factors at multiple levels, such as individual- (such as demographics), organizational- (such as departmental affiliation), and spatial-level (such as lab/office distance to library) variables that may explain library use. For privacy reasons, results will be shared using only aggregated and anonymized data.

***Goal 2*: To develop tools, scripts, and protocols that serve the library analytics needs of our community of project advisory group (PAG) institutions.** We plan to achieve this goal through two modes of engaging with the PAG community. First, we plan to develop and share with PAG institutions *data dictionaries* (names, definitions, and attributes about data elements that are in a database) of the datasets we will be using in the project. This will facilitate data harmonizing across PAG institutions (and with U-M data) and thus smooth the exchange of ideas and best practices over the course of the project. We will maintain a repository of tools and scripts developed for data cleaning and database construction and grant PAG members early access to this repository. We intend to visit PAG institutions in order to establish the strong, trust-based working relationships that enhance collaboration. We also plan to regularly update PAG members via email on the latest status of data management and analysis activities with a view to making it easier for them to voluntarily replicate our work or perform their own LLA studies. Thrice a year, we intend to hold BlueJeans[1] Research Meetings with PAG members for more thorough debriefs of project activities. Lastly, throughout the project we plan to stay in regular phone/email contact with PAG members in order to help them resolve replication issues as they arise.

Second, we will give PAG members access to aggregated and anonymized findings (from *Goal 1*, using U-M data) through a web-based dashboard running on data that will be located on a secure virtual data enclave. Further, in addition to the dashboards, we will hold annual PAG Workshops where member representatives will

[1] BlueJeans is a cloud-based conferencing service that offers audio, video, content sharing and is well-suited for content-rich cross-national and international collaborations: (http://its.umich.edu/communication/videoconferencing/blue-jeans). An advantage of BlueJeans is that it is compatible for use with some types of sensitive data, including Protected Health Information (PHI).

be tutored and immersed in training tasks to ease voluntary replication of the project, but using their own data.

Project Assumptions

Our proposed LLA activities are aligned with an approach to library assessment that is grounded in certain assumptions about knowledge and learning practices (Knight & Buckingham Shum, 2017). Specifically, we assume that LLA: a) can contribute to diminishing learning disparities by empowering learners and educators/scholars through the provision of mid-course corrections or prescriptive models based on past behavioral outcomes; b) can expand how we define learning and subsequently extend our conception of what data we should be collecting to address a more encompassing view of learning, and; c) can complement traditional, non-computational ways of assessing how libraries impact learning.

Communities Served

The *first community* is comprised of the individuals that are served by the U-M Library. For fall 2015, U-M had a population or potential library user community of 43,651 students, 7,056 faculty, and 1,174 postdoctoral research fellows (University of Michigan, 2016a, 2016b). All groups are involved in multiple aspects of the learning process at U-M, be it undergraduate participation in learning communities like the Undergraduate Research Opportunity Program (UROP), or researchers co-authoring journal publications. In a survey of several young alumni cohorts, 96% indicated that course instruction, faculty, and research experiences were their most meaningful learning experiences while at U-M (Zaruba, 2009). Our *study population* is the individuals from the three groups above that are involved in the learning process as captured by the three outcomes of interest: course instruction, research, and publications.

The *second community* is comprised of the 15 PAG institutions (see Appendix 3), 12 of which are colleges and universities in the US and Canada. Collectively, the libraries in these higher education institutions directly serve close to 400,000 faculty, students, and staff. The PAG colleges and universities are institutionally diverse with respect to enrollment size (2,700-60,595), region (New England, Great Lakes, Far West, and Canada), and locale (City Large, City Small, Suburb Large, Town Distant, Rural Fringe, and Rural Distant).[2]

The other three PAG institutions include two that are professional associations of higher education libraries, and an award-winning public library. The Big Ten Academic Alliance Library Initiatives (BTAA-LI) focuses on optimizing student and faculty access to the combined resources of 15 top research universities in the US, and creating a collaborative environment where staff across the member libraries can work together to solve mutual problems. The Association of College & Research Libraries (ACRL) is the largest division of the American Library Association (ALA) and currently has more than 11,000 members which is nearly 20% of ALA's total membership. Lastly, the Ann Arbor District Library (AADL) serves the residents of the Ann Arbor area and is a previous recipient of the "National Library of the Year" award by the *Library Journal*. AADL provides a broad range of services at no charge to residents in its service area and to others outside this service area for a fee.

Data Description & Analysis

We plan on retrieving and analyzing both **deidentified** and **identifiable** library data with a focus on post-2015 data. In 2016 the U-M Library revised its privacy policy to make possible the collection of individual identifiers (Alexander, Bradley, & Varnum, 2018). Library data sources include: website server logs, library catalog server logs, proxy server logs, circulation history and related data, and campus status & affiliation data. We also plan to use the Learning Analytics Data Architecture (LARC), a research-focused data set containing information about students who have attended U-M since approximately 1996. These datasets are linkable through both strong and unambiguous identifiers that are unique to each individual user, as well as IP addresses and timestamps. The size of these datasets would be in the order of a few gigabytes a day assuming that most of the

[2] Region and locale data are drawn from *The Integrated Postsecondary Education Data System* (IPEDS).

data are retained on the servers, plus additional database transaction logs. *Identifiable* data are those generated by authenticated users while *deidentified* data are those that are unauthenticated but originate from networks managed by U-M as opposed to the world at large. For example, of the just over 1 million "sessions" on the library catalog for calendar year 2016, about 300,000 sessions were originated from U-M networks, and roughly 100,000 sessions were authenticated (or identifiable). The identifiable data will also be merged with U-M Data Warehouse[3] (administrative e.g. human resources, grants, course instruction, etc.) and publications datasets.

I. Data Types

Deidentified library use data: We plan to retrieve unauthenticated library use data with no individual identifiers including: website server logs, catalog search logs, proxy server logs, and circulation history data. We expect it to take about six months to retrieve, clean, and database the deidentified and identifiable (below) library data.

Identifiable library use data: We plan to collect authenticated library data that has individual identifiers. This will entail working with LIT staff to retrieve data such as: website server logs, catalog search logs, proxy server logs, and circulation history data.

Outcomes & administrative data: Our goal is to pull administrative data from the U-M Data Warehouse on e.g. human resources (such as demographics, departmental affiliation, job type, gender), course instruction, grant proposals, etc. Additionally, we plan to carry out a data pull of publications metadata from various sources such as Scopus and PubMed. We anticipate that it should take no more than three months to pull these data.

II. Analysis Plan

Our analysis is intended to be carried out in three major steps as follows. First, using deidentified data, we plan to apply clustering algorithms to identify typologies of library users on the basis of library interactions (Figure 1) e.g. user type A (high degree of use), user type B (low degree of use), etc. Second, we intend to replicate the clustering analysis using the identifiable data and use clustering findings from the deidentified analysis for robustness checks. Third, we plan to use clusters from identifiable data in analytics involving sequence mining, and predictive and prescriptive modeling. Our proposed study is a holistic, longitudinal, and multi-method (e.g., regression modeling[4]) empirical analysis informed by patterns of library use (Figure 1) and models of information behavior.[5]

III. Types of Analysis

We expect to mainly use LA methods such as sequence mining, clustering, and predictive and prescriptive modeling, each of which allows us to address different issues with respect to how the library impacts learning.

Sequence mining: This method could be used to examine patterns of learning and library activities to ascertain if there is a specific order among them. For example, analysis of the linked library-outcomes data may reveal that some learning activities e.g. enrolling in a STEM course are preceded or followed by a specific order of interactions with the library. Sequence mining enables us to examine issues such as: *How does interaction with the library influence a student's GPA? How does interaction with the library influence receipt of a research grant? How does interaction with the library influence publications in high impact journals?*

Clustering: Clustering refers to a family of unsupervised methods of data analysis that group objects into classes, typically along dimensions like similarity of objects within the same cluster, or dissimilarity of objects in different clusters. Clustering enables us to study patterns in the associations between library interactions and learning outcomes to address issues such as: *Given three categories with respect to course, grant, and publication performance – underachieving, as expected, overachieving – what characterizes the individuals in each category with respect to library interactions?*

[3] "Data Sets in the U-M Data Warehouse." (http://www.mais.umich.edu/reporting/datasets.html)
[4] "An Introduction to Regression Analysis." (http://www.law.uchicago.edu/files/files/20.Sykes_.Regression.pdf)
[5] "Information Behavior." (https://pages.gseis.ucla.edu/faculty/bates/articles/information-behavior.html)

Predictive and prescriptive modeling: These methods use past behavior to reveal insights about the future. Predictive analytics aims to understand the future and address the issue of what will likely happen given the constraints of what happened in the past with respect to library activities and learning outcomes. Prescriptive analytics is intended to address the issue of what we ought to do given the past link between specific library activities and learning outcomes. Both types of analytics employ methods like text mining, statistical modeling and forecasting, machine learning, and computational modelling. These types of analysis allow us to examine issues such as: *Among student cohorts taking courses that entailed library instruction, which library interaction factors predict performance at the end of semester or academic year? Among faculty submitting proposals in a given fiscal year, which library interaction factors predict grant funding success? What types of library instruction should be offered to help student persistence in e.g. STEM courses? What types of library instruction would help investigators improve their chances of submitting successful research proposals?*

Preliminary Analysis

Two preliminary analyses that used U-M library server logs revealed that: a) with respect to enrolling in a course with library instruction within the first year, the odds of year-to-year retention (enrolling at U-M in the second year) are 1.78 ($p < 0.001$) times greater for enrollees (N = 10,051) over non-enrollees (N = 8,749),[6] and; b) with respect to submitting a research proposal in the 4-10 years preceding the library service year, the odds of using/searching library resources are 1.15 ($p < 0.05$) times greater for unsuccessful (N = 5,680) than successful (N = 5,143) investigators.

Project Implementation

I. Project Activities (see *Schedule of Completion* for more details)

a) Year One (June 2018 – May 2019)

Set up project website & infrastructure (data repository, virtual data enclave, dashboard); retrieve /pull data; clean, and database data; data merges; develop data analysis plan; create data dictionaries, tools, scripts, and protocols (& share with PAG); BlueJeans research meetings; site visits to PAG institutions; PAG workshop.

b) Year Two (June 2019 – May 2020)

Site visits to PAG institutions; data analysis; prepare presentations; share preliminary findings with PAG; revise tools, scripts, and protocols; BlueJeans research meetings; 2020 Coalition for Networked Information (CNI) Spring Membership Meeting; PAG workshop.

c) Year Three (June 2020 – May 2021)

Prepare manuscripts; site visit to PAG institution; work with PAG institutions on voluntary replication; BlueJeans research meetings; 2020 Library Assessment Conference; 2020 CNI Fall Membership Meeting; 2021 Association of College & Research Libraries (ACRL) Conference; PAG workshop; public release.

II. Project Resources: Personnel, Time, Budget

We are requesting $318,149 in direct costs and $177,131 in indirect costs ($495,280 in total costs) for the three-year study. The direct costs are: $231,134 of salaries and wages, and $ 61,515 of fringe benefits to support the team of investigators, staff, graduate student, and consultants; $15,000 for travel to PAG sites that are not in the State of Michigan, and to research conferences; and $10,500 to cover the costs of the yearly PAG workshops.

a) Key Personnel

***Felichism "Felix" Kabo,* M.Arch, Ph.D., Principal Investigator.** Felix is research faculty at the Institute for Social Research (ISR) and the Michigan Institute for Clinical and Health Research (MICHR). He is affiliated with the Institute for Healthcare Policy and Innovation (IHPI). He has expertise in network science, statistical

[6] The 2013-2014, 2014-2015, and 2015-2016 cohorts are considered for this analysis because they can be followed for at least a year.

modeling, and in using small, large, and "Big" data for research and evaluation. He will oversee all project activities including setting up the project cyberinfrastructure (data repository, website, dashboard, virtual enclave) data management and storage, data analysis and visualization, statistical modeling, dissemination of results, project management, and communication with the PAG. He is responsible for ensuring that project resources are first available to PAG members, and then to others in the library field and beyond.

***Stephanie Teasley,* Ph.D., Co-Investigator.** Stephanie is a Research Professor, School of Information, and the Director of the Learning, Education & Design (LED) Lab. She is also the President of the Society for Learning Analytics Research (SoLAR), an international and inter-disciplinary network of leading researchers who are examining the role and impact of analytics on teaching, learning, training and development. She will help the project with rich domain expertise on areas on learning, and explanatory, predictive, and prescriptive analytics.

***Laurie Alexander,* Co-Investigator.** Laurie is the Associate University Librarian for Learning and Teaching. She will provide the project with subject domain expertise in issues such as user instruction, learning spaces, information technologies, the undergraduate experience, and reference services.

***Doreen Bradley,* Co-Investigator.** Doreen is the Director of Learning Programs and Initiatives at the U-M Library. She will provide the project with subject domain expertise on areas such collaborations between library instructors and faculty across campus to further information literacy programs, and assessing the impacts of library instruction.

***Bryan Skib,* Co-Investigator.** Bryan is the Associate University Librarian for Collections. He will help the project with domain expertise on issues such as management of library materials, collection development strategy, the preservation of library materials, document delivery activities, and library technical services.

***Maurice York,* Co-Investigator.** Maurice is the Associate University Librarian for Library Information Technology (LIT). He will provide subject domain expertise on the library's technology environment including digital collections and access tools, library website, technologies used for the delivery of library services, and library information sources and resources. All the LIT staff involved in the project (whether funded directly by the project, or via cost-share through the university) will report directly to Maurice.

b) Consultants & Project Staff

***Robert Melendez,* Research Computer Specialist.** Robert is research staff at ISR. He will provide data management and coding expertise, in addition to providing the technical skills necessary for creating databases and computerized programs for reviewing, coding and tabulating the various forms of data that this study will maintain. Robert will also provide his expertise in the creation of the secure data repository, including ensuring that all data have been anonymized before being released to the PAG and/or the public.

***Ken Varnum,* Technical Lead, U-M Library.** Ken is the U-M Library Information Technology (LIT) Senior Program Manager for Discovery, Delivery, and Learning Analytics. His project responsibilities include providing subject domain expertise on technology and project management for the library's discovery interfaces, delivery interfaces, and the library's learning analytics infrastructure. He will coordinate the LIT developers and programmers to retrieve the library use data and make it available to the project team, and bring his library data knowledge and expertise to bear on the project.

***James Hilton,* Ph.D., Consultant.** James is the University Librarian & Dean of Libraries, Vice-Provost for Digital Education & Innovation, Arthur F. Thurnau Professor at the School of Information, and Faculty Associate at ISR. Dr. Hilton is a national leader in technology issues around higher education. For example, he has led, championed and fostered technology initiatives that cross boundaries between institutions, and between academic and information technology units. He will provide expertise on collaboration and metrics development in teaching, learning and research.

***Michael Clark*, Ph.D., Consultant.** Michael is the Statistician Lead at Consulting for Statistics, Computing & Analytics Research (CSCAR). Dr. Clark has expertise in mixed models, additive models, latent variable models, text analysis, structural equations modeling, machine learning, Bayesian inference, reproducible research practices, data visualization, and efficient programming practices. He will provide statistical and computing support including analysis, reproducibility, code optimization, and visualization and reporting.

***Nancy Allee*, MLS, MPH, Consultant.** Nancy is the Deputy Director, Taubman Health Sciences Library & Library Faculty, Department of Learning Health Sciences. She will guide the project on the role of the library in the health sciences with respect to areas such as: academic and clinical engagement; research and informatics; collections and information services; and global health and outreach.

c) Project Finance Personnel

Project finances will be managed by Survey Research Center (SRC) programmatic finance staff with years of experience managing sponsored research funds. They will report regularly to Dr. Kabo on account status. We are not requesting any funds for the SRC programmatic finance staff.

d) Project Advisory Group (PAG)

The PAG is comprised of experts in library practice and research drawn from 15 institutions in the US and Canada (see Appendix 3 for a list of PAG members and their parent institutions). The PAG will advise the project team on how to best capture the pathways between the library's interactions with its users, and the three learning outcomes (research, course instruction, and publications). Another key PAG function will be providing guidance to the project team such that the study outcomes and findings are applicable across the spectrum of small-, medium-, and large-sized colleges and universities. For example, we will create and share data dictionaries with PAG members and design tutorials that help them to voluntarily replicate any or all aspects of our study at their home institutions. Demands on the PAG are minimal. They will be requested to: a) provide input and guidance over the course of the project; b) participate in the BlueJeans Research Meetings; and c) attend the annual PAG Workshop (we will use project and other resources as needed to ensure that all PAG members are able to attend the workshops). PAG members are eligible to collaborate with the project team on publications, reports, and other research products. We are requesting funds for the PAG workshops (see Appendix 3 for a tentative agenda for the 2019 workshop).

Communications Plan

The project is intended to reach the widest possible audience among general and academic libraries. Thus, we are requesting funding to present findings at leading conferences in the library field, including: the Library Assessment Conference, the Coalition for Networked Information (CNI) Membership Meeting, and the Association of College & Research Libraries (ACRL) Conference. We plan to host annual PAG workshops at U-M over the course of the project where we will share aggregated and anonymized results with PAG members and immerse them in training activities geared to ease voluntary replication of the project, but using their own data. Additionally, PAG institutions will be given the opportunity to share progress reports with respect to replication activities. In consultation with PAG institutions, we will explore the option of making the workshops publicly accessible via recordings and/or live streams. We plan to create a project website that will be publicly accessible, including the dashboard for displaying aggregated and anonymized findings. Our goal is to regularly share project materials (e.g. data dictionaries, protocols, tutorials, and findings) with PAG institutions. We plan to work closely with PAG institutions to identify other opportunities for disseminating and promoting the project e.g. via the Big Ten Academic Alliance and at ACRL. Lastly, we also plan to produce at least two peer-reviewed manuscripts starting in Year 3.

C. Diversity Plan

U-M and PAG institutions serve communities that are diverse with respect to gender and race. There is also diversity with respect to learning activities, and even types of learners. For example, two PAG members are

community colleges that serve not only as pipelines to leading research universities (including two that are also PAG members), but also as key centers for continuing education or for individuals looking to make career switches or transitions. The PAG community is also quite diverse in terms of enrolment, region, locale, and institutional type (from small and medium private colleges, to large public research universities). We plan to consult with units from the U-M Office of Diversity, Equity & Inclusion to explore how study outcomes could be used for educational outreach e.g. to schools in the area and region. Lastly, the project team and PAG are a diverse group of individuals with respect to race (Black, White, Latino), gender, disciplinary backgrounds (library & information science, organizational studies, statistics & computation), and institutional affiliations.

D. National Impact

This project will provide guidance to PAG institutions and other libraries on how to best design and implement empirical, holistic LLA studies of the links from library usage, to learning outcomes such as research, course instruction, and publications. The project will produce a set of tools, scripts, and protocols that will be freely available to all libraries. The study will serve as a template for other libraries with respect to: 1) studies that collect and store library use data with individual identifiers while maintaining the privacy of individuals; 2) designing and implementing a holistic LLA study of the link from library use to multiple learning outcomes, and; 3) creating a secure cyberinfrastructure (data repository, virtual enclave, and dashboard) for LLA research that facilitates collaboration in a community of diverse institutions in the US and Canada. This study addresses three major impediments to this type of holistic investigation: a) balancing user privacy needs with access to data with identifiers; b) holistic LLA probes of the library, and; c) how to leverage subject domain, statistical, and computational expertise at U-M and PAG institutions in order to enable even resource-strapped members of the PAG community to conduct the LLA they need to better assess how libraries impact learning.

A critical element of this project is the extensive use of search and server logs for that are authenticated and un-authenticated. While previous studies have worked with authenticated user logs (Nackerud, Fransen, Peterson, & Mastel, 2013; Soria, Fransen, & Nackerud, 2013, 2014, 2016), this is the first study we are aware of that will also analyze un-authenticated user logs, significantly enhancing our capacity for holistic LLA.

This project is intended to advance the capacity of U-M and the PAG community to use LLA to understand how the library impacts learning. The project goals are designed to be achieved through an open and transparent process where the study outcomes (datasets, analysis, tools, scripts, protocols) and findings should be useful and relevant to the PAG and wider library communities. The project is expected to lead to a greater LLA capacity in the PAG community and to enable its members to better articulate the value of their libraries, irrespective of differences in PAG member size or budgetary constraints. Beyond its utility to the PAG community, this project should also significantly impact the broader field of libraries as, in accordance with a commitment to openness and transparency, the study outcomes will be developed using open access and open source frameworks as much as possible. A recent report on UK academic libraries recommended research on trends in user (students and researchers) behaviors, as well as highlighted the need for a library analogue to LA (Pinfield, Cox, & Rutter, 2017). Moreover, over 80% of interviewees within and beyond the library field in the US and UK said measuring the library's impact on students would have a significant or greater impact on their institutions in the coming 10 years (Pinfield et al., 2017). This clarion call for analytics is relevant to a wide range of libraries, and, through its capacity for LLA, this project is well poised to make a positive national impact in that regard.

SCHEDULE OF COMPLETION												
Project Activities	**Year 1; June 18-May 19**				**Year 2; June 19-May 20**				**Year 3; June 20-May 21**			
	Q1	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**
Set up the virtual data enclave; establish privacy and data access protocols												
Set up data repository												
Set up project website; including dashboard that will run on the website												
BlueJeans Research Meeting #1												
Retrieve, clean and database library data (website server logs, catalog search logs, etc.)												
Pull, clean, and database data from U-M Data Warehouse												
Pull, clean, and database publications data from e.g. SCOPUS and PubMed												
BlueJeans Research Meeting #2												
Create and coordinate data dictionaries												
Develop computational tools, scripts, and protocols; share with PAG												
Establish common analytics framework including scripting languages e.g. R, Python, SQL												
Merge/link library, administrative, and outcomes data												
BlueJeans Research Meeting #3												
Commence development of tutorials for database and data management, and data merges												
Commence development of data analysis plan												
PAG Workshop #1												
Complete development of tutorials												
Complete development of data analysis plan												
Commence clustering data analysis												
BlueJeans Research Meeting #4												
Initiate work with PAG institutions on voluntary replication												
Roll out mature data repository, virtual data enclave, and project website												
BlueJeans Research Meeting #5												
Revisions to tools and infrastructure based on PAG feedback												
Commence sequential mining data analysis												
Share preliminary findings with PAG; revise/respond to PAG feedback												
BlueJeans Research Meeting #6												
Prepare presentation(s) based on findings												
Coalition for Networked Information (CNI) Spring Membership Meeting												
Give PAG access to mature dashboard, repository, and virtual data enclave												
PAG Workshop #2												
Commence predictive modeling												
Commence prescriptive modeling												
BlueJeans Research Meeting #7												
Share preliminary findings with PAG; revise/respond to PAG feedback												
Prepare manuscript(s) based on findings												
BlueJeans Research Meeting #8												
Library Assessment Conference												
Coalition for Networked Information (CNI) Fall Membership Meeting												
Work with PAG institutions on voluntary replication												
BlueJeans Research Meeting #9												
Association of College & Research Libraries Conference												
Prepare repository, dashboard, and other resources for final public release												
PAG Workshop #3												

Q1 = June - August
Q2 = September - November
Q3 = December - February
Q4 = March - May

DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

- ☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

We will create a research dataset using: a) library data (website server logs, library catalog server logs, proxy server logs, circulation history and related data, and campus status and affiliation data); b) administrative data (e.g. courses, research activities and funding, human resources, etc.) from the University of Michigan Data Warehouse; c) student data from the Learning Analytics Data Architecture (LARC), and; d) publications data from sources such as SCOPUS and PubMed.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The principal investigator on the project and the University of Michigan (U-M) will hold the intellectual property rights for the research data they generate.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

We do not anticipate any privacy concerns with the dataset. The data will be anonymized and deidentified before being deposited on the secure data repository for sharing.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:
URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

The data are intended to help us understand the value of the library to the university. We plan to retrieve deidentified and identifiable library use data including: website server logs, library catalog server logs, proxy server logs, circulation history and related data, and campus status & affiliation data. First, using deidentified data, we aim to apply clustering algorithms to identify typologies of library users on the basis of library interactions e.g. user type A (high degree of use), user type B (low degree of use), etc. Second, we aim to replicate the clustering analysis using the identifiable data, using the deidentifiable clustering findings for robustness checks. Third, our goal is to use clusters from identifiable data in the sequence mining, and predictive and prescriptive modeling (e.g. statistical and computational models) of the links between learning outcomes and library user types. These analyses will control for potentially confounding factors at multiple levels, such as individual- (such as demographics), organizational- (such as departmental affiliation), and spatial-level (such as lab/office distance to library) variables that may explain library use. For privacy reasons, results will be shared using only aggregated and anonymized data. The size of these datasets would be in the order of a few gigabytes a day assuming that most of the data are retained on the servers, plus additional database transaction logs. We also plan to use the Learning Analytics Data Architecture (LARC), a research-focused data set containing information about students who have attended U-M since approximately 1996, as well as administrative data from the U-M Data Warehouse (grants, courses, human resources).

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

The project will require approval from U-M's Health Sciences and Behavioral Sciences Institutional Review Board (IRB-HSBS). We intend to submit an IRB protocol in June 2018 and expect to have secured approval by July 2018 as we will request an expedited IRB review compliant with procedures established by U-M IRB-HSBS.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

Some of the data will include personal identifiers such as names. The data will be deidentified and anonymized prior to any release to the research community. This work will be done by research staff at the Institute for Social Research (ISR) with extensive experience in handling and preparing sensitive and restricted data for research use and public release.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

n/a

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Library and university servers will be used to collect the server logs. LARC and U-M Data Warehouse data are stored on a centralized repository at the university. Staff who have undergone the requisite training and clearance are authorized to pull data from this repository.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

We will create a codebook and data documentation that will be posted on the same site as the dataset. The documents will contain all the information necessary -- e.g. study design, variable-level detail, etc. -- for others to use the data accurately and effectively. These files will be shared as PDF and text files and will be stored on the project data repository. We will also create data dictionaries that we will share with members of the project advisory group (PAG). These dictionaries will be stored on the project data repository.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Long-term archiving, management, and dissemination of the data will via the project data repository and virtual data enclave that will be created for us by staff from the Inter-University Consortium for Political and Social Research (ICPSR). The repository and data enclave will enable us to maximize access to the data, store the data safely and securely, and protect the confidentiality and privacy of individuals in the data.

A.8 Identify where you will deposit the dataset(s):

Name of repository: Inter-university Consortium for Political and Social Research (ICPSR) Virtual Data Enclave

URL: (https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html)
"The virtual data enclave (VDE) provides access to restricted-use data and is a virtual machine launched from the researcher's own desktop but operating on a remote server, similar to remotely logging into another physical computer. The virtual machine is isolated from the user's physical desktop computer, restricting the user from downloading files or parts of files to their physical computer. The virtual machine is also restricted in its external access, preventing users from emailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally. Available options of the VDE include file sharing among project team members and vetting of results." (https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html)

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?
.
The data management plan will be reviewed annually. At our request, an ICPSR unit manager will conduct a disclosure review of all files that the investigator wants to use after their use of the enclave.