Learning from data reuse: successful and failed experiences in a large public research university library

Jung Mi Scoulas[1], Sandra L. De Groote[2], Paula R. Dempsey[3]

Abstract

This paper illustrates a large research university library's experience in reusing data for research collected both within and outside of the library. The purpose of the paper is 1) to demonstrate when, why and how data are reused in a large public research university library, 2) to share tips on what to consider when reusing and reproducing data for research data, including issues of replicability and research ethics, and 3) to share challenges and lessons learned from data reuse and reproducibility experiences. This paper presents five proposed opportunities for data reuse conducted by three researchers at the institution's library, which resulted in three successful instances of data reuse and two failed data reuses. Learning from successful and failed experiences is critical to understand what works and what does not work in order to identify best practices for data reuse. This paper will be helpful for librarians who intend to reuse data for research and publication.

Keywords

Data reuse, surveys, internal and external data, data practices, reproducibility

1. Introduction

Every day, academic libraries collect a wealth of data such as gate counts, book circulations, study room reservations, and use of online library resources including website and chat logs, in order to capture users' activities and behaviors and often to use them for decision making (e.g., staffing). Furthermore, these data are often reused by researchers in order to demonstrate the library's value and impact on students' academic success (Allison 2015; LeMaistre, Shi & Thanki 2018; Soria, Fransen & Nackerud 2013; Soria, Fransen & Nackerud 2017) by examining the relationships between library use, students' academic performance and retention rate, etc. However, the types of challenges and issues researchers encounter, and how to handle the process of data reuse and ensure reproducibility, are understudied. In addition, when reusing data, it was discovered that many published research projects in the library field did not follow data protection practices (e.g., informed consent, anonymization), which may potentially result in violating students' privacy (Briney 2019).

In addition to using internally generated data, library researchers have used data collected by outside entities to demonstrate the library's impact. Examples of large datasets that are widely used in the library field include the Association of College and Research Libraries (ACRL) Library Trends & Statistics survey[4] containing information on staffing, teaching and collections; Integrated Postsecondary Education Data System [5](IPEDS) Academic Libraries Survey containing information on library resources, services and expenditures; and the National Center for Education Statistics[6] (NCES) Academic Libraries Survey. These three datasets can be accessed via ACRL Metrics,[7] an online subscription service. Another large dataset is from the Association of Research Libraries (ARL) statistics,[8] a series of annual surveys containing information on library collections, expenditures, staffing and service activities for ARL member libraries. Many researchers have used the large

datasets described above to demonstrate the library's impact. Mezick (2007) used data from ACRL, ARL and IPEDS in order to examine the correlations between library expenditure, staffing and student retention. Haddow and Joseph (2010) also analyzed data from ARL, IPEDS and NCES to investigate the relationships between student retention and library use (workstation use and logging into library resources. Stewart (2012) analyzed data from NCES and IPEDS to compare graduation rate and library expenditure per student from 2004 to 2010. Crawford (2015) used data from Academic Libraries Survey and IPEDS to measure the relationships between institutional expenditures, library expenditures, library use and students' graduation and retention rates.

However, there are challenges in using large scale data in terms of data reuse and research reproducibility (Yan et al. 2019). Accessing and analyzing these large data sets requires effort, knowledge, skills and expenses such as online subscriptions to access some datasets (e.g., ACRL Metrics[7] and ARL statistics[8]). In particular, it is critical to understand the research design, codes, and data analysis related to statistical software in order to reuse the data or reproduce the results of studies that were conducted by other researchers. Academic librarians are aware of the importance of using evidence-based data for decision making and attempting to determine the library's impact. Scholarship in the field would benefit from more comparative studies that would require sharing data across institutions. However, little is known about questions such as "What types of data can be reused in the library field?" "Are there any challenges to consider when reusing data or reproducing research?" "What issues need to be considered when reusing data or reproducing research?" The answers to these questions are critical for librarians to gain awareness of what data is available to them, to learn how to use evidence-based data, to make the results reproducible, and to increase research productivity.

The purpose of the paper is to: 1) demonstrate when, why and how data are reused in a large public research university library; 2) share tips on what to consider when reusing and reproducing data for research; and 3) share lessons learned from data reuse and reproducibility experiences from a research perspective. This paper will be useful for librarians who are not familiar with reusing existing data to understand what types of data are available for them in their own institutions, where to begin addressing new problems, and how to transform a failed experience for data reuse and reproducibility to a successful experience. This paper provides practical implications for promoting data reuse and reproducibility practices for librarians. It will be helpful for librarians who intend to reuse data and reproduce it in research for publication.

2. Literature review

2.1. Data reuse and reproducibility

In the data life cycle, there is a sequence of the stages of data life, from data creation to data reuse, with data reuse identified as the last stage of its useful life (Briney 2015). In Elsevier's website where it describes research data, "reproducible" and "reusable" data are displayed as the highest stages of its life cycle (Elsevier 2019). Data reuse refers to using secondary or existing data to examine new problems that were not considered in the original study and generate new findings (Yoon 2017; Zimmerman 2008). The National Science Foundation's (NSF) Social Behavioral and Economic (SBE) division subcommittee on reproducible science defined reproducibility as "the ability of a researcher to duplicate the results of a prior study using the same materials and procedures as were used by the original investigator" (Bollen, Cacioppo, Kaplan, Krosnick & Olds 2015, p. 3). In the report on how to promote research practices, Bollen and colleagues considered it "a minimum necessary condition for a finding to be believable and informative" (p. 4). The benefits of data reuse include validating results and potentially increasing research productivity and effectiveness (Yoon & Kim 2017). Research reproducibility requires accuracy of results, transparency of data collection and analysis (Bollen et al. 2015), in order to reproduce the findings. In fact, due to the high percentage of results that cannot be reproduced, published articles in various disciplines from the sciences (e.g.,

neuroscience; Gilmore, Diaz, Wyble & Yarkoni 2017) to the social sciences (e.g., psychology; Open Science Collaboration 2015) have entered a "reproducibility crisis" (Sayre & Riegelman 2018). Under some federal (e.g., the National Science Foundation[9]) and private funding agencies (e.g., the Bill & Melinda Gates Foundation[10]) data sharing policies, researchers who receive grants are obligated to share their data if it is not sensitive data (e.g., GPAs and patrons' IDs) (Briney 2015). Reproducibility is one of the primary reasons of why data sharing is needed (Briney 2015). With the increase in data sharing, data is now frequently accessible and is easier to find for either data reuse or validating results (Briney 2015).

2.2. Data reuse and reproducibility: behaviors and challenges

In a study of survey researchers regarding their perceptions and perspectives of data reuse and reproducibility, they were asked about the main reasons for reusing data (Yan, Huang & Palmer 2019). They found that the top two reasons were to "conduct new analysis" (87%) and to "compare results" (70.5%), and the lowest reason was to "reproduce published articles" (18.5%). Yan and colleagues further reported that the top problem related to reproducibility was "not enough detail in the published paper on how study was conducted" (85.2%).

Increased opportunities for sharing and reuse of research and academic data have raised issues around the ethics of data sharing as well as practical barriers to data reuse and reproducibility. In spite of the increase in data sharing, however, finding datasets remains difficult. Briney (2015) shared strategies on how to find data for data reuse: look for published articles, search for a "subject-specific index" in your specialty (p. 164), look for discipline-specific data repositories in your field, and check other resources such as re3data.[11]

Yoon (2016) conducted interviews with 23 researchers who reused social science data and found that barriers to data reuse involved: inaccurate descriptive information about the data, difficulty accessing data, difficulties with data format, software, special analytic programs, problems with samples (i.e., too many missing values), issues with original data analysis, and data cleaning. Faniel, Kriesberg and Yakel (2012) found that novice researchers in the process of matching and merging data from multiple sources had difficulty dealing with different time periods and creating unique identifiers.

Yoon and Kim (2017) further studied what factors influenced researchers' behaviors in data reuse by conducting a survey of 1,528 participants and found that researchers' perceived usefulness (e.g., increase in research productivity) was the strongest predictor that affected data reuse intentions. Additional factors influencing the reuse of data found in their study included concerns about misinterpreting the data, copyright infringement, availability of internal resources, and availability of data repositories.

Among the types of data capture and reuse in academic libraries is participation in learning analytics initiatives. Perry and colleagues (2018) examined how 54 ARL member libraries participated in practices, policies and ethical issues about learning analytics within member libraries. The results related to behavior policies for learning analytics revealed that although 70% of respondents obtained approval from the university's Institutional Review Board (IRB) for their learning analytics project, not all respondents informed students about their learning analytics initiative (Perry et al. 2018).

Libraries' practices to link data pertaining to students' identifying information associated with their library use (e.g., database logins from library website, book check outs and login from library computer workstation) and their academic success (e.g., GPA and retention rate) draw attention to ethical issues such as maintaining patron privacy, appropriate data handling and de-identification. As Perry, et al show, it is not clear whether or not researchers inform students about reusing data from

their library use information such as book check outs and database logins, or whether the researchers obtained IRB approval for the research project related to measuring the correlations between students' library usage and their academic achievement and learning outcomes. For example, in both studies conducted by Soria and colleagues (2013; 2017), they used students' library usage data including book check outs, interlibrary loans, online chats with reference librarians, and database logins from various data sources. However, the authors did not clearly explain the security of the data system, and so it was not clear whether they informed students when reusing the data for their project and how they handled students' personal information (i.e., identification numbers) in the articles (Soria et al. 2013; Soria et al. 2017). Allison (2015) also examined whether students' library use (checkouts and off-campus access to library resources) had an impact on students' GPA. Similar to Soria and colleagues, Allison also reused data collected from the library to address the research question. Unlike Soria and colleagues, Allison explained in her article that "the data were then made anonymous by removing the ID number that could be linked back to individual student' records" (p. 33). However, whether she obtained IRB approval or whether she informed the students whose data she used was not addressed.

Some researchers demonstrated in their data reuse practices that they practiced adequate data protection in their published research projects. A good example of data reuse can be seen in the study by De Jager, Nassimbeni, Daniels and D'Angelo (2017). De Jager and colleagues studied the correlations between undergraduate students' library use and their GPA using the data obtained from the institution's data warehouse in the University of Cape Town, South Africa. In the article, they described each step in obtaining the anonymized data (e.g., library visits and checks out of the library materials) so there was no possibility of identifying students. They also addressed the challenges of obtaining library data from the data warehouse like delays in obtaining the assurance of anonymized data, the process of securing data from various sources, and ensuring the "integrity and completeness of the reported data" (2017, p. 5). The data reuse project by LeMaistre, Shi and Thanki (2018) is another good example. LeMaistre and colleagues (2018) at the Nevada Stage College investigated whether or not students' use of online library resources was correlated with their GPA using the login data on EZProxy. Nevada Stage College included a data privacy policy and indicated that students have the option to opt out of their private information being saved by contacting the library via the EZProxy log-in page. In their study, LeMaistre and colleagues clearly stated the importance of students' data privacy to protect students' confidentiality (2018).

While data reuse practices vary by discipline, researchers in the library field tend to reuse data collected directly by their institution and library, focusing on measuring the library's impact on students' academic success and learning. However, few studies have examined data reuse practices on other types of data such as survey data. Due to data privacy policies, not all institutions and libraries have access to students' data. When researchers encounter data privacy and policy issues, are there alternatives to measuring the library's value or address reusing existing data? This paper will demonstrate five cases in data reuse practices and how three researchers in a university library experienced successful and failed data reuse practices by focusing on five points: original data description, purpose of the data reuse, level of accessibility, challenges and lessons learned, and outcomes.

3. Case specific examples of library data reuse practices

3.1. Case 1: Student surveys

3.1.1. Background and original data description: Beginning in Spring 2016, the University Library has conducted locally developed biannual surveys for students to: 1) assess current student behavior and satisfaction associated with the use of online resources, library services, and the physical library; 2) examine students' needs related to library resources and services for improvement or expansion of physical library spaces; and 3) determine if there is any correlation between library use and

students' academic achievements. Both the 2016 and 2018 surveys were approved by the institution's IRB. The university library obtained data about students' demographic information and their Grade Point Average (GPA) from the Office of Institutional Research. The survey results were used to information decision making about which areas of services and resources are needed for improvement. The data was shared with various stakeholders inside and outside of the university library. Also, the results were used to determine which areas of services and resources are needed for improvement.

3.1.2. Purpose of data reuse: The proposed need for data reuse was 1) to reproduce the results that were reported by the head of assessment and scholarly communications; 2) to measure the value of the library in terms of students' success from the 2018 student survey data by examining both quantitative and qualitative data; and 3) to use the 2016 and 2018 student survey data to examine any differences between students' library website use and satisfaction and students' use of library spaces and satisfaction.

3.1.3. Level of accessibility: The survey data was captured in Qualtrics (2018 version). Following the completion of the survey it was exported as SPSS and Excel files and stored in the university library Box folders, a web-based cloud file sharing management service accessible only to the Assessment Coordinator Advisory Committee (AC2). The analyzed data and results are stored in the university library's data warehouse (also a Box folder) where it is available to all university library staff excluding student employees.

3.1.4. Challenges and lessons learned: Data was saved in both SPSS and Excel files. Most of the descriptive information about variables in the data dictionary created by the head of assessment and scholarly communications was clear and accurate. Several issues occurred when cleaning and analyzing the original data for reuse. First of all, students' demographic information was coded differently in the 2016 and 2018 survey data. For example, in the 2016 survey the gender code for female was (1) and the code for male was (0), whereas in the 2018 survey the female code was (1) and the code for male was (2). To eliminate any confusion in the data interpretation, all of the codes in the 2016 data were converted to match those in the 2018 data. Second, the survey response scales in the 2016 and 2018 survey data were different. For instance, a 6-point Likert scale [e.g., from very difficult (1) to I have not used this (6), including (3) neutral] was used in the 2016 surveys, whereas a 5 point-Likert scale [e.g., from I have not used this (0) to very easy (4)] was used in the 2018 survey. These different codes and scales were adjusted by converting the ordinal scales to continuous data based on the previous study (Preston & Colman 2000). The last challenge was to select which questions overlap in both surveys. Based on the issues described here, library faculty will maintain the same response scale and codes used in the 2018 student survey for future student surveys. This will allow us to more accurately capture and compare data for making better decisions and improvements in library services. In order for other researchers to replicate and reproduce the published studies, survey instruments and the procedures of data collection and analysis were included in the publications (Scoulas & De Groote 2019; Scoulas & De Groote, under review).

3.1.5. Outcomes: Re-using data from locally developed student surveys expands the existing literature on academic libraries efforts to demonstrate the library impact on students' academic success. If user surveys are carefully designed at the beginning, there are a great number of potential benefits of data reuse. Using locally developed surveys, academic libraries can make decisions using the evidence-based findings for improvement, demonstrate the library value on students' academic success and learning outcomes, and examine user's behaviors and attitudes over time to monitor trends.

3.2 Case 2: Chat with a Librarian

3.2.1. Background and original data description: Transcripts of chat reference interactions between patrons and the University Library reference providers are created in the course of ordinary work

serving patron information needs of the UIC community. The transcript data are produced by the LibChat platform (Springshare[12]), which is commonly used by academic libraries to provide virtual reference services. Other platforms that produce similar data include LibraryH3lp[13] and QuestionPoint[14] (acquired by Springshare from OCLC in May 2019). The University Library retains the transcripts for 5 years for purposes of follow-up with patrons, training, and quality control.

3.2.2. Purpose of data reuse: Chat transcripts provide a solid basis for understanding what library users need, how library staff work with patrons, and how reference services might be improved. There is extensive research using chat transcripts. A review of such studies from 1995-2010 found the researchers were most concerned with what level of service was provided, who used the service, what questions were asked, and the ways in which providers responded (Matteson, Salamon & Brewster 2011). Most of the studies are single-institution case studies. Other than a few studies of consortial data (Kwon 2006; Meert 2009), it is unusual for studies of chat transcripts to provide cross-institutional comparison (Dempsey 2017). A recent study of UIC Library chat transcripts analyzed the extent to which patrons were referred to subject specialists and how chat providers framed the referral (Dempsey 2019). On ongoing study examines in-depth reference interactions to gauge whether librarians from a range of institutions who provide virtual reference believe they should have been referred to a subject specialist.

3.2.3. Level of access or reuse data: Any University Library employee with credentials for the LibChat system has access to transcripts for the past 5 years. In order to harvest them for research, however, IRB approval is needed. The privacy rights of the patrons and the well-being of the chat providers employed by the library must both be considered as risks in balance with potential benefits. Patrons have a right to privacy outlined in a UIC Library policy (UIC Library 2017). All identifying information is scrubbed from the data, and therefore researchers and readers of the published reports will have no way to identify individuals represented in the data. However, it is possible, though highly unlikely, that a patron's topic of research could serve as an identifier. Investigators using transcripts must consider carefully the details presented in published quotations from the data. In the case of chat reference providers, they might recognize their own work if quoted in a study. This recognition could cause distress if the analysis identifies shortcomings in the service provided, but no library employee should be judged on the basis of one chat interaction. Thus, the analysis must take into account the dignity of people doing challenging work in a fast-paced environment and handle critiques in a way that emphasizes the aggregate and does not harm either of the individuals participating in any one chat interaction.

3.2.4. Challenges and Lesson learned: Privacy. De-identifying data is a significant investment of time, because even though names entered by patrons can be scrubbed automatically in LibAnswers, patrons and providers frequently use one another's names and provide e-mails and other identifiers that must be deleted individually. *Data ownership.* Who owns these data, and under what circumstances can they be shared? Transcripts are created as part of ordinary work flows and most likely represent work for hire by the University Library. The effort that goes into de-identifying and cleaning the data in other ways does not confer ownership on the researchers; rather it is done as part of one's professional responsibilities with the purpose of benefitting patrons and the knowledge base of the profession. These efforts provide more benefits the more widely the data are shared, because multi-institutional data are likely to generate more generalizable findings. As noted above, cross-institutional studies are unusual – sharing data to promote replicating studies across institutions would improve knowledge in the field. If possible, establish and document answers to questions of ownership before embarking on data collection and cleaning and share data in a repository such as the Qualitative Data Repository at Syracuse University.[15]

3.2.5. Outcomes: Re-using naturally occurring data in the form of chat transcripts has contributed to the scholarly conversation as an empirical basis for establishing best practices for navigating the

reference interview, teaching information literacy in the virtual context, providing patrons relevant resources, and referring patrons to subject specialists. If used thoughtfully to design training tools, these findings have the potential to improve virtual reference across academic libraries. Moving toward wider availability of data for replication in cross-institutional studies would have even more impact.

3.3 Case 3: Library collections and research productivity

3.3.1. Background and original data description: The data for this study came from various sources with different purposes. The Scopus[16] database is the largest source of abstracts and citations of peer reviewed research literature. This database is typically used to find literature on specific topics but for the case being discussed here, it was used to identify publications by institution and the references used in them. The Higher Education Research and Development (HERD) survey[17] contains information of research and development expenditures at U.S academic institutions. ARL Statistics[8] data contains annually conducted survey data from the ARL member libraries about collections, expenditure, staffing and service activities.

3.3.2. Purpose of data reuse: Demonstrating the value and potential impact of the academic library on research at academic institutions can help support arguments for maintaining or increasing funding to the library. Recent studies have not explored funding, collection size, and collection use and their relationship with research output (publications) using existing data, nor explored if new library metrics (database searches, journal article downloads) can predict research output at academic institutions. The purpose of this study was to explore the impact of the research library on faculty productivity by using ARL statistics, HERD expenditure data, and Scopus publication information.

3.3.3. Level of access or reuse data: The data from the various sources was merged together into one data set. As noted above, the Scopus database, which requires an institutional subscription, was used to identify the number of publications produced at specified institutions, and the number of refences used in them. Synthesized data on faculty publications and references is not directly available in Scopus and the data needed to be searched for and recorded. For example, to find the number of publications for institution Y, a search by institution Y was conducted and the results limited by year. The publication data for institution Y needed to be displayed in a specific way to capture the number of references included in the publications for a given year. Both the number of publications and number of references included in the publications needed to be manually recorded in a spreadsheet. HERD data, which indicates the research and development expenditures of an institution, is synthesized annually and shared in a downloaded spreadsheet where each row lists an institution and its data. Each institution was searched for in a summary table for a specific year, and the needed data was entered into to a spreadsheet. Finally, ARL data was also retrieved to provide information about the expenditures and resource use of the libraries included in the study. To retrieve the data, pull-down menus are available for institution and data variable. For each year in the study, the institutions included in the study and variables of interest were selected. These data were then exported in spreadsheet format, that was then merged with the data collected from other sources. A subscription is also required to access ARL data.

3.3.4. Challenges and lesson learned: Because the data for this study was collected from three different resources, it was challenging to ensure the three data sets lined up for each institution. For example, some institutions have separate budgets and administrative lines for their health sciences colleges and libraries. It was not always clear from the collected data sets if it was all locations of an institution, or just certain disciplines or cities where data would be captured. Familiarity, particularly with large state institutions that have multiple locations was helpful to understand what academic locations would be included under the title of an institution. Close inspection of data coverage was needed to ensure that data sets were representative of the same population. One aspect that we

wanted to explore with this study was how productive faculty were. We were able to obtain the number of publications for an institution through searches in Scopus, and ARL provides data on the number of full and part-time faculty. But this data was not sufficient to approximate average faculty productivity at an institution because it was not clear how many publications may have been written by individuals other their faculty, including students, fellows, post docs and staff, nor was it clear that faculty would be defined similarly at all institutions. Use of the data was also hindered by the ability to readily find data dictionaries that clearly defined and described the data. Knowing who to ask to get a definition of a data variables was important. Finally, the data collected by the ARL was greatly changed between2014 and 2015. While adding new data measures meant exploring the ability of new measures to assess productivity, because the collection of some data points ceased, it meant limiting the number of years that could be retrospectively be studied, as well as possible changes over time.

3.3.5. Outcomes: If data from different sources is collected thoughtfully and carefully, data can be merged and reused to explore new relationships. Due to uncertainty with the alignment and comprehensiveness of some of the institutional data between data sources, several institutions were excluded from this study. Given that it was not possible to determine the average number of publications per faculty by institution using the re-used data, partial correlation analyses were done holding number of students and faculty constant to explore the impact of libraries.

3.4 Case 4: University Library's Undergraduate Engagement Program (UEP) data

3.4.1. Background and original data description: "Finals Week Relaxation Station" is a successful UEP program that targets undergraduate students and helps them to manage their stress, which is thought to influence their academic success. Since Fall 2016, an outreach coordinator has collected data from students who participated in this program by asking them to swipe their ID cards when visiting the relaxation station. The data contains the date the program occurred and students' identification numbers.

3.4.2. Purpose of data reuse: It was not clear whether this program is providing beneficial services for the targeted audience. Further, there is interest in measuring any correlation of students using the Finals Week Relaxation Station on their GPA by comparing groups (i.e., one time use vs. more than one time). The Outreach Coordinators and the Assessment Coordinator wanted to use this data not only for their internal use (identifying the users' characteristics) but also to investigate the impact of the program on students' academic success.

3.4.3. Level of access or reuse data: Originally, the data was accessible only to one of the outreach program coordinators. After discussing data reuse with the other outreach coordinator for the purposes described above, the first outreach program coordinator shared the raw data via Box folder. The data was saved in Excel format and each event per semester was saved in a separate Excel file. A total of 6 Excel files were created. Given that the raw data contained only students' identification numbers, this information was sent to the Office of Institutional Research (OIR) to retrieve detailed information about the students: program, class level, GPA for the beginning of the semester and at the end of the semester. The Assessment Coordinator combined and organized the data and sent it securely to the OIR. The process of merging data into one file and retrieving the data from OIR took about a month. The Assessment Coordinator analyzed the data and shared the descriptive statistics with the outreach coordinators.

3.4.4. Challenges and lessons learned: The findings were unexpected and interesting. There was interest in publishing the findings by demonstrating the background of the UEP and how program impact was measured. However, a couple of critical issues for data reuse were discovered. First, when collecting data, students were not informed of how their data would be used. Second, while this project aimed to establish a sustainable and welcoming culture in the library for undergraduate students' learning and academic success, this project has not received IRB approval for data capture

as a research project. The university library is committed to protecting students' privacy; we could not overlook the ethical issues that may harm students' privacy and autonomy. Given that the findings from the data indicate that students' use of the Finals Week Relaxation Station increased over time, and many students used it more than one time, this information will be valuable for increasing buy-ins inside and outside of the library in order to demonstrate the impact of the program. To proceed with publication of the results, we need to address the ethics of data reuse. As the historical data cannot be used, the outreach coordinators and assessment librarian will need to explore other methods of data capture if there is further interest in publishing impact studies.

3.4.5. Outcomes: While this project may not proceed for publication at this time, the project allowed the outreach coordinators to consider how and when data needs to be collected to better understand the users and outcomes of the program. To this end, the collected data can be utilized to demonstrate not only whether the desired outcomes were met, but also if the UEP program has an impact on students' learning outcomes.

3.5 Case 5: Faculty survey

3.5.1. Background and original data description: Since Spring 2017, the University Library has conducted locally developed biannual surveys for faculty to: 1) assess how faculty members utilize library resources (both online and in print) for their teaching, research or scholarship and 2) examine the university faculty's level of satisfaction with the library's programs and services. The University Library obtained data about faculty's demographic information from the Office of Institutional Research (email address, faculty status, the highest FTE department, etc.). Prior to conducting each survey, all of the documents associated with these proposals were submitted to the IRB for approval. The 2017 survey was approved by the IRB as a research project, whereas the 2019 survey was determined by the IRB as a quality improvement project, stating that the 2019 survey project is considered as having "no intent to produce or contribute to generalizable knowledge," meaning that "this initiative was deemed not human subjects research and was therefore not received by the Institutional Review Board" based on the objectives that the University Library proposed.

3.5.2. Purpose of data reuse: The proposed need for data reuse were to: 1) to compare the differences in the university faculty's library use in 2017 and 2019; and 2) to measure the impact of the faculty's library use and satisfaction on their research productivity.

3.5.3. Level of access or reuse data: Similar to the student survey data, the faculty survey data was stored in Box folders with access restricted to AC2. After analyzing the data, the summarized data was stored in the secure university library data warehouse where it is available to all university library staff excluding student employees.

3.5.4. Challenges and lesson learned: As noted above, when submitting the proposal to the university IRB, the researchers indicated that the objectives of the projects were to "identify the library resources and services used by the University faculty for teaching, research or scholarship and examine faculty's perceived importance and level of satisfaction with library support." Based on the objectives stated in the IRB application, the university IRB determined "this project as a Quality improvement project with no intent to produce or contribute to generalizable knowledge." In other words, this project is no longer considered as a human research project and was not reviewed by the IRB. In our effort to reduce the number of questions asked in the survey, we may have reduced the usefulness of the data collected. It is not clear how this determination may impact the use of this data set with previous and future data sets where the data is considered "generalizable". Additionally, as an afterthought, we realized that as part of the data embedded in the survey, we could have also included information about the number of publications of each faculty member. This would have made the results of the survey more useful to us, and we would also have had data that would have made the results more generalizable. Based on the issues that the authors encountered,

we learned the importance of original data collection: whether key variables were included at the beginning. Once the data is collected, it is done. We cannot go back to collect the data again. That is, if the original data is not sufficient and does not contain key outcome variables (in this case, publications), this will prevent us from reusing the data to draw a meaningful finding.

3.5.5. Outcomes: From the lessons that were learned above, we will be mindful not only about focusing on the immediate questions we want to answer, but also obtaining meaningful data that can be used for future decision-making and trends to demonstrate the value of the library.

Table 1. Summary of case studies conducted by three researchers in a large public research university

	Types of Data	Location of Data Preservation	Purpose of Data Reuse	Outcomes
Case 1: Student survey	Survey (numeric and text)	University Box folder, University Library Data Warehouse	1) To reproduce the results 2) To conduct new analysis 3) To compare the results	Conference presentations (Scoulas and De Groote, June 17 2019) (Scoulas and De Groote, June 18 2019) Publication (Scoulas and De Groote 2019) (Scoulas and De Groote under review)
Case 2: Chat with a Librarian	Transcripts (text)	LibAnswers, Springshare servers	To conduct a new analysis	Publication (Dempsey 2019)
Case 3: Library collections and research productivity	Surveys (numeric)	ARL Statistics, Scopus and HERD	To conduct a new analysis	Publication In progress
Case 4: Undergraduate Engagement Program	Students' identification information (numeric)	University Box folder, University Library Data Warehouse	To conduct a new analysis	Unable to publish
Case 5: Faculty survey	Survey (numeric and text)	Qualtrics Server, University Box folder, University	To conduct a new analysis and compare the results	Unable to publish

		Library Data Warehouse		

4. Lessons learned from data reuse experiences

As shown in the five case examples described above, our goal for this paper is to share what three researchers in a large public research university library experienced throughout the process of data reuse practices for research: what worked, what did not work, and what to improve for the next research project. Below is the summary of the key lessons the authors learned during our data reuse practices.

Key lessons that can be derived from these five case examples:

- *Instrument.* Maintain key questions and response scales to compare trends over time. The core questions can be a great asset for taking a longitudinal approach.
- *Ethical issues*. Be sure to obtain informed consent from users when gathering data, even if you are not sure whether research will be conducted at that time or in the future.
- *Privacy*: Consider the implications of the research for possible violations of user privacy. This should extend beyond financial or legal ramifications and also take into account the participants' dignity and overall well-being.
- *The Importance of including key variables in the original data collection*. Carefully design the original research project. If key variables are missing, such as the number of publications per each faculty member in the faculty survey, the data is less likely to be reused for demonstrating the library's value in the faculty's research productivity.
- *Documentation of data, data collection and analysis*: Record every procedure of data analysis and the codes used. In addition, save all of the data analysis output. This information will be critical for data reuse, replication and reproducibility in research.
- *Data ownership*. When starting a project with existing data, think ahead about rights and responsibilities surrounding those data to establish whether you can share the data, take them with you to a new employer, etc. (CDL UC3[18]). Consider making data available in a repository when possible to promote cross-institutional research.
- *Data coverage and definitions*. Understand the coverage of the data in order to merge like data sets together. If it is not clear if each data set is using the same source to produce the data or if the definitions of a data variable are not specified, data reuse may not be possible.

5. Implications for other librarians

While most of the literature highlighted the rigorous research practices of journals (e.g., Elsevier), funding agencies (e.g., NSF), and libraries with published articles, few studies discuss and share the individual researchers' successful and failed experiences in data reuse and reproducibility in the library field. If every researcher is committed to learning and following the full process of data management (e.g., proper data storage, recording all the steps of data analysis, documenting any changes in data analysis, data output, and depositing data in archives) within their organizations, they will share their data confidently. Further, other researchers can easily reuse data that is available within their organizations and reproduce the results that were conducted within or outside of their organizations. If a researcher is concerned about the issues of sharing raw data due to confidentiality, at least a summary of statistics (e.g., a matrix of correlations) needs to be presented

in their publications, which will enable other researchers to reproduce the results using those statistics (Bollen et al. 2015). This will be useful for librarians who are interested in becoming involved in research and scholarship activities by reusing data that already exists in their organizations and outside of the library, or by practicing reproducibility. To this end, librarians can have empirical evidence for establishing best practices for navigating various projects that will be beneficial for other librarians across academic libraries.

References

Bollen, K, Cacioppo, J, Kaplan, R, Krosnick, JA & Olds, JL 2015, *Social, behavioral, and economic sciences perspectives on robust and reliable science*. Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences. Available from: http://web.stanford.edu/group/bps/cgi-bin/wordpress/wp-content/uploads/2015/09/NSF-Robust-Research-Workshop-Report.pdf. [23 December 2019].

Briney, KA 2015, *Data management for researchers: Organize, maintain and share your data for research success*. Pelagic, Exeter, UK.

Briney, KA 2019, 'Data management practices in academic library learning analytics: A critical review', *Journal of Librarianship and Scholarly Communication*, vol. 7, no. 1. https://doi.org/10.7710/2162-3309.2268

Crawford, GA 2015, 'The academic library and student retention and graduation: An exploratory study', *portal: Libraries and the Academy,* vol. 15, no. 1, pp. 41–57. https://doi.org/10.1353/pla.2015.0003

De Jager, K, Nassimbeni, M, Daniels, W & D'Angelo, A 2017, 'The use of academic libraries in turbulent times', *Performance Measurement and Metrics*, vol. 19, no. 1, pp. 40–52. https://doi.org/10.1108/pmm-09-2017-0037

Dempsey, PR 2017, 'Resource delivery and teaching in live chat reference: comparing two libraries.' *College & Research Libraries,* vol. 78, no. 7, pp. 898–919. https://doi.org/10.5860/crl.78.7.898

Dempsey, PR 2019, 'Chat reference referral strategies: Making a connection, or dropping the ball?', *College & Research Libraries,* vol. 80, no. 5, pp. 674-93. https://doi.org/10.5860/crl.80.5.674

Elsevier 2019, 'Research data'. Available from: https://www.elsevier.com/about/open-science/research-data. [23 December 2019].

Faniel, IM, Kriesberg, A & Yakel, E 2012, 'Data reuse and sensemaking among novice social scientists', *Proceedings of the American Society for Information Science and Technology,* vol. 49, no. 1, pp. 1-10. https://doi.org/10.1002/meet.14504901068

Gilmore, RO, Diaz MT, Wyble, BA & Yarkoni T 2017, 'Progress toward openness, transparency, and reproducibility in cognitive neuroscience', *Physiology & Behavior,* vol. 176, no. 1, pp. 139–48. https://doi.org/10.1016/j.physbeh.2017.03.040

Haddow, G & Joseph, J 2010, 'Loans, logins, and lasting the course: Academic library use and student retention', *Australian Academic and Research Libraries*, vol. 41, no. 4, pp. 233–244. https://doi.org/10.1080/00048623.2010.10721478

Kwon, N 2006, 'User satisfaction with referrals at a collaborative virtual reference service', *Information Research: An International Electronic Journal*, vol. 11, no. 2, p. n2. Available from: http://www.informationr.net/ir/11-2/paper246.html. [23 December 2019].

LeMaistre, T, Shi, Q & Thanki, S 2018, 'Connecting library use to student success', *portal: Libraries and the Academy*, vol. 18, no. 1, pp. 117–40. https://doi.org/10.1353/pla.2018.0006

Matteson, ML, Salamon, J & Brewster, L 2011, 'A Systematic review of research on live chat service', *Reference & User Services Quarterly,* vol. 51, no. 2, pp. 172-89. https://doi.org/10.5860/rusq.51n2.172

Meert, DL & Given, LM 2009, 'Measuring quality in chat reference consortia: a comparative analysis of responses to users' queries', *College & Research Libraries.* vol. 70, no. 1, pp. 71–84. https://doi.org/10.5860/0700071

Mezick, EM 2007, 'Return on Investment: Libraries and Student Retention', *Journal of Academic Librarianship*, vol. 33, no. 5, pp. 561–566. https://doi.org/10.1016/j.acalib.2007.05.002

Open Science Collaboration 2015, 'Estimating the reproducibility of psychological science', *Science,* vol. 349, no. 6251. https://doi.org/10.1126/science.aac4716

Perry, MR, Briney, KA, Goben, A, Asher, A, Jones, KML, Robertshaw, MB & Salo, D 2018, '*Learning analytics*', SPEC Kit 360. Washington, DC: Association of Research Libraries. Available from: https://publications.arl.org/Learning-Analytics-SPEC-Kit-360/. [23 December 2019].

Preston, CC & Colman, AM 2000, 'Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences', *Acta Psychologica,* vol. 104, no. 1, pp. 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Sayre, F & Riegelman, A 2018, 'The reproducibility crisis and academic libraries', *College & Research Libraries*, vol. 79, no. 1, pp. 2–9. https://doi.org/10.5860/crl.79.1.2

Scoulas, JM & De Groote, SL 2019, 'Factors affecting university students' library visits in person and online using a multiple regression approach', paper presentation, *Evidence Based Library and Information Practice 10*, Glasgow, United Kingdom, June 17. Available from: https://eblip10.org/abstracts/tabid/8487/Default.aspx#scoulas. [23 December 2019].

Scoulas, JM & De Groote, SL 2019, 'Assessing the university library's impact on students' academic performance', poster presentation, *Evidence Based Library and Information Practice 10*, Glasgow, United Kingdom, June 18. Available from: https://eblip10.org/Abstracts/tabid/8487/Default.aspx#scoulas2. [23 December 2019].

Scoulas, JM & De Groote, SL 2019, 'The Library's impact on university students' academic success and learning', *Evidence Based Library and Information Practice,* vol. 14, no. 3, pp. 2–27. https://doi.org/10.18438/eblip29547

Stewart, C 2012, 'An overview of ACRL metrics, Part II: Using NCES and IPEDs data', J*ournal of Academic Librarianship*, vol. 38, no. 6, pp. 342–45. https://doi.org/10.1016/j.acalib.2012.09.018

UIC University Library 2017, 'User privacy policy.' Available from: https://library.uic.edu/about/policies#privacy. [23 December 2019].

Yan, A, Huang, C & Palmer, CL 2019, 'Data reuse and reproducibility in Earth System Science: a survey of current practices, barriers, and expectations'. Available from: https://www.essoar.org/doi/pdf/10.1002/essoar.10500464.1. [23 December 2019].

Yoon, A 2016, 'Red flags in data: Learning from failed data reuse experiences', *Proceedings of the Association for Information Science and Technology,* vol. 53, no. 1, pp. 1–6. https://doi.org/10.1002/pra2.2016.14505301126

Yoon, A 2017, 'Data reuse trust development', *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 946–56. https://doi.org/10.1002/asi.23730

Yoon, A & Kim, Y 2017, 'Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories', *Library and Information Science Research*, vol. 39, no. 3, pp. 224–33. https://doi.org/10.1016/j.lisr.2017.07.008

Zimmerman, AS 2008, 'New knowledge from old data: The role of standards in the sharing and reuse of ecological data', *Science, Technology, & Human Values,* vol. 33, no. 5, pp. 631–52. https://doi.org/10.1177%2F0162243907306704

Endnotes

[1] Jung Mi Scoulas is a Clinical Assistant Professor and Assessment Coordinator in the Richard J. Daley Library, University of Illinois at Chicago. Correspondence should be addressed to Jung Mi Scoulas and can be reached by Email: jscoul2@uic.edu
[2] Sandra L. De Groote is a Professor and the Head of Assessment and Scholarly Communications in the Richard J. Daley Library, University of Illinois at Chicago. Email: sgroote@uic.edu
[3] Paula R. Dempsey is an Assistant Professor and the Head of Research Services & Resources in the Richard J. Daley Library, University of Illinois at Chicago. Email: dempseyp@uic.edu
[4] https://acrl.countingopinions.com/
[5] https://nces.ed.gov/ipeds/
[6] https://nces.ed.gov/surveys/libraries/
[7] https://www.acrlmetrics.com
[8] https://www.arlstatistics.org/home
[9] https://www.nsf.gov/bfa/dias/policy/dmp.jsp
[10] https://www.gatesfoundation.org/How-We-Work/General-Information/Information-Sharing-Approach
[11] https://www.re3data.org
[12] https://springshare.com
[13] https://libraryh3lp.com/

[14] https://www.oclc.org/en/questionpoint.html
[15] https://qdr.syr.edu/
[16] https://www.scopus.com/search/form.uri?display=basic
[17] https://www.nsf.gov/statistics/srvyherd/
[18] https://uc3.cdlib.org/2016/09/08/who-owns-your-data