Big Data and Academic Libraries: The Quest for Informed Decision-Making

Tiffini A. Travis and Christian Ramirez

abstract: Libraries remain one of the last places on campus where the purging of usage data is encouraged and "tracking" is a dirty word. While some libraries have demonstrated the usefulness of analytics, opponents bring up issues of privacy and debate the feasibility of student-generated library data for planning and assessment. Using a study conducted at the University Library, California State University, Long Beach, the authors of this article identified practical knowledge of data research that academic librarians will benefit from understanding. Readers will learn about the role campus culture plays in data gathering, be exposed to the complexities of learning analytics and Institutional Review Board (IRB) clearance, and read how the authors weighed the ethical use of big data analysis for assessing students.

Introduction

Collecting and using personal data always present an uneasy balance of ethics, privacy, and potential disaster. Whether it is companies mining data to influence presidential elections or commercial brands using analytics to push their products, there is a fine line between demonstrating value and violating a user's rights. This is particularly the case in an academic setting, where the users are students.

Big data consists of large data sets generated by the compilation of information over time. Companies use big data to calculate the impact of business models or to create user profiles based upon the actions of consumers. It is a proven method of assessing business success and predicting buyer behaviors. For educators, big data has emerged as an opportunity to improve academic performance by drawing together student information over many years. In higher education, a common

For educators, big data has emerged as an opportunity to improve academic performance by drawing together student information over many years.

portal: Libraries and the Academy, Vol. 20, No. 1 (2020), pp. 33–47.

form of big data is learning analytics. The 2011 Horizon Report, a collaborative project of the New Media Consortium and EDUCAUSE Learning Initiative, defines learning analytics as "the interpretation of a wide range of data produced by and gathered on behalf of students to assess academic progress, predict future performance, and spot potential issues."[1] One of the first large-scale applications of big data for academic decision-making was an initiative called integrated planning and advising services (IPAS). Designed to increase student retention, IPAS combined multiple data sources to identify and predict students at risk for failure.[2] As the use of analytics increased exponentially in higher education, libraries also found that data generated from library-controlled systems could help improve services for students and the campus as a whole.

Two Faces of Data: Analytics and Value

Libraries strive to safeguard individual privacy by creating a safe space for users to explore a variety of materials. As a rule, they do not retain long-term data at the user level to avoid sharing records with the government, employers, or any other third party. With the rise of social media and Internet platforms, the control of user information has changed. Third-party vendors already offer metrics derived from users' database searches and browsing habits. The ability of libraries to receive statistically rich reports via vendors and in-house Web analytics has created pressure to consider user information in tailoring library services and acquisitions. Single sign-on systems, which permit students to use one set of log-in credentials to access multiple applications, link student accounts to learning analytics across campus. Such technology has made it easy for academic libraries to cross-reference their data to demonstrate impact on student success. Research employing big data in libraries has evolved from data-driven assessment for financial expenditures, such as serials, interlibrary loan, and online purchases, to other areas of library service. The seminal work *The Value of Academic Libraries: A Comprehensive Research Review and Report* identified the use of analytics as a potential measure of educational impact of libraries. This acknowledgment contributed to an increase in the use of large data sets for examining the significance of libraries, initially focusing on student retention and academic achievement. Jisc, a nonprofit company in the United Kingdom that provides digital and technology services to support higher education, was at the forefront of encouraging the adoption of analytics by libraries. Jisc's Library Impact Data Project was one of the first initiatives to pull together library usage data and graduation rates across multiple universities in the United Kingdom.[4]

Research employing big data in libraries has evolved from data-driven assessment for financial expenditures, such as serials, interlibrary loan, and online purchases, to other areas of library service.

Data sets used by libraries subsequently became more complex by utilizing multiple demographic variables and including data across several years. Melissa Bowles-Terry examined grade point averages (GPAs) and included the demographic variable of biological gender.[5] In 2015, a series of case studies featuring big data analysis of librar-

ian instruction aimed to determine the relationship between instruction and student grades.[6] Several of these studies also investigated the correlation between library teaching and student retention.[7] Another growing area of library analytics research explored the benefits of course-specific library instruction on undergraduates' grades.[8] Multiple studies expanded demographic parameters to include such variables as high school GPA, ethnicity, and socioeconomic status.[9] Many of these studies also grouped students by cohorts and included multiple library access points, such as consulting digital and print materials and using the library as space. The most comprehensive big data study in the United States included over 42,000 students across 12 campuses. In this study, Joni Blake, Bowles-Terry, Shirlene Pearson, and Zoltán Szentkirályi examined gender, ethnicity, class level, and age. They built upon previous studies by adding the variables of librarian teaching methods and course type.[10]

Two Faces of Data: Analytics and Ethics

An equal number of articles have either advocated for or cautioned against the use of large data sets made up of student information. One area of concern is privacy. EDUCAUSE dedicated an entire issue of *EDUCAUSE Review* to examine the topic in higher education.[11] Information science literature also debated concerns regarding library use of data-driven research. The most compelling argument against learning analytics in libraries was that it violates the American Library Association Code of Ethics. Kyle Jones and Dorothea Salo focus on the need to protect vulnerable populations. They coined the term "dataveillance" to describe the tracking of students without their consent, which may do more harm than good.[12] Grant Campbell and Scott Cowan cautioned that big data should be considered from the perspective of LGBTQ privacy due to the integral link between identity formation and exploration of the queer community in library spaces. Violations of privacy in a library setting could lead to fear of being "outed."[13]

In recognition of the potential pitfalls, many universities have set up their own guidelines for the ethical use of big data. Stanford University in Stanford, California, created an initiative to explore and track this issue. The use of big data to analyze user behavior or tailor services has the potential to allow damaging invasions of privacy. Ethics in data analytics research can also become a concern when the students researched are those whom educators wish to help most. As Jacob Metcalf and Kate Crawford note, "Data sets and algorithms have historical, material specificity that is laden with political and ethical values."[14] For example, data derived from tracking use of learning management systems may seem useful for increasing student retention and preventing student failure. However, labeling learners based upon past decisions or artificial profiles can have the unanticipated effect of perpetuating bias and might overlook insights or the educational growth students achieve.[15] Researchers must remain mindful that markers used to flag at risk students can be invasive and a source of stigma.

The use of big data to analyze user behavior or tailor services has the potential to allow damaging invasions of privacy.

Another criticism of this type of research is whether causation is truly a sound methodology for measuring library value. However, correlation analysis in social science research can be useful for creating predictive models of future situations or behaviors and insightful when coupled with additional nonnumerical factors. Correlation studies will never definitively measure the role of the library in student learning because so many factors play a role in academic performance. The variation in published research results is an indication of the difficulty of using correlation analysis to show a relationship between library instruction and student learning.[16] Earlier studies with one or two demographic fields were less definitive than those that incorporated more qualitative measures, such as course level GPA, precollege performance, and cumulative college performance indicators. While adding more qualitative data to analysis can increase the likelihood of demonstrating causation, it can also lead to flawed inference if interpreted incorrectly. With these issues in mind, librarians can still incorporate professional ethics in this type of research by using data-gathering approaches that minimize risk while working toward a better educational experience for students.

Correlation studies will never definitively measure the role of the library in student learning because so many factors play a role in academic performance.

Few published studies addressed the ease or difficulty of gaining access to information or any sort of guidelines for planning a study before the data are collected. This case study will examine the steps initiated at a large urban university for a learning analytics project centered on student information and big data research. It is the goal of the authors to provide practical considerations and lessons learned for any librarian attempting to implement a big data project. While this article focuses on the experience of at one campus, the considerations outlined are applicable to many.

Behind the Numbers: Planning a Big Data Project

When using large data sets, keep in mind several considerations. Think of it as a three-step process: developing a plan for data collection, navigating campus channels to gain access to the data, and finally, devising a plan for analyzing the data. While these steps may seem simple and straightforward, several issues can become stumbling blocks to this type of data-centered research.

Planning for Data Collection

Uses for the Data

Librarians have long advocated for a curriculum that allows students to apply information literacy (IL) skills at multiple points across the disciplines. Big data can enable libraries to assess trends of past learning or predict future learning based upon user profiles. Creating such profiles can help librarians develop a curriculum map of a student's most likely sequence of classes based upon past course pathways. Using predictive profiling, libraries can estimate where IL instruction should occur based upon past needs.

Much of the library literature has advocated the usefulness of evidence-based practice for establishing library value. Comparing data sets from similar institutions using the National Survey of Student Engagement (NSSE) and the National Center for Education Statistics (NCES) has become a popular way to benchmark academic libraries.[17] Analytics is also a useful tool for curriculum mapping and measuring the use of library instruction throughout the university. This type of analysis holds the potential to identify the optimal point in a student's academic career to incorporate or emphasize IL concepts. Comparing data sets from like institutions and working with campus-generated data are viable approaches to programmatic assessment. Big data offers a potential to reveal patterns of student learning and to identify where library intervention can increase IL skills. Additionally, such data can give the library a more accurate idea of how many students attend instruction and how often.

Big data offers a potential to reveal patterns of student learning and to identify where library intervention can increase IL skills.

Developing the Methodology

When constructing a methodology for big data research in higher education, it is vital to set the foundation of the research questions before selecting which fields to include. The data set for this study at California State University, Long Beach (CSULB) included over 1 million records from 2012 to 2016. If librarians do not regularly conduct quantitative research using course-generated data, it is helpful to consult a statistician who works with learning analytics. Since library analytics is primarily correlation analysis, controlling for anomalies, such as grade inflation or instructor bias, is useful. Demographic variables are also important, whether to classify the students or to determine significant differences. Since many universities do not record socioeconomic status or label students by skill level, certain demographic fields, such as Pell Grant eligibility and high school GPA, are useful to ascertain socioeconomic or at-risk status. The authors worked with the assistance of a statistics professor at CSULB to separate data fields into two categories: quantitative and qualitative (categorical data). Within the data, native and transfer students were identified and classified into two groups: those who attended library sessions and those who did not. The goal was to obtain as granular an analysis as possible to reveal a fuller picture of student-library interaction.

The second part of developing a data collection plan was determining which courses to include for analysis. In addition to any course that received face-to-face library instruction, the study looked at courses that focused on basic writing or critical thinking and advanced writing courses in the General Education program. This approach had the dual benefit of examining potential impact and integration into the General Education curriculum as well as illustrating the typical trajectory of students who are at risk and take remedial composition courses. In addition to all research methods courses, General Education courses were selected if they identified information literacy as a primary student learning outcome. Courses that identified assignments or content directly related to information literacy were included as well. It was important to encompass courses

Determining what can be obtained from other departments or offices and which information should come from the library can help automate the gathering of relevant data and ensure efficient collection.

with research components or content that emphasized using IL skills (writing and critical thinking). This approach cast a wide net to track patterns of student performance in a variety of courses and to compare student grades with library instruction versus those without.

In data aggregating projects, once the data sources are determined and the optimal fields selected, automation of the collection process ensures consistency. Determining what can be obtained from other departments or offices and which information should come from the library can help automate the gathering of relevant data and ensure efficient collection.

Takeaway: Your Data Are Only As Good As Your Source

During this process, the team at CSULB encountered several complications that might apply to other libraries. The most notable problem was the quality of the data available from the library. When collecting data, all course information must be entered uniformly. At a minimum, the name of the instructor and course section will be needed to merge library data with campus enrollment data. At CSULB, some librarians included section numbers, while others did not. Still other librarians put the title of the course in the title section but omitted the course number. Small errors such as these required hours of clean up going back to 2012. Compounding the issue, the library instruction system had a glitch that downloaded times and dates of some courses incorrectly. Fixing this bug took more time than anticipated because the only way to fill in missing fields was to manually match courses with librarians' calendars and the university schedule of classes.

To avoid these issues, libraries should regularly aggregate and anonymize library and campus data every semester or academic year to maintain an active database with longitudinal information. Systematic data collection allows libraries to assess changes in patterns or usage that may have implications for library services.

Navigating Campus Channels

Campus Culture, Buy-In, and Access

Campus culture and system interoperability can enable or hinder cross-referencing of data. Assessment of student learning is required at the department and college level, but data are rarely shared across campus or between units. Campus culture influences willingness to share data. Because some departments do not value assessment, they collect little student-level data. If using analytics is not an intrinsic part of the departmental culture, the department may use inconsistent collection methods or none, making it more difficult to obtain the data.

The first step of the research project was to identify units on campus from which to request data for analysis. CSULB has many units that collect various types of student information, but they seldom share data with one another for assessment, outside of the information required by the California State University (CSU) System or for accreditation.

Before starting any research project, a library needs to investigate potential sources of data available on campus. Several divisions would be ideal as sources of complementary data. From exchanges with each unit, it became obvious that unit politics and department culture played a large role in willingness to work with the library and with one another.

Before starting any research project, a library needs to investigate potential sources of data available on campus.

The authors first contacted the Academic Advising Center, which had recently purchased an advising system that collected analytics to track bottleneck courses and student progress. The second office consulted was Institutional Research & Assessment (IR&A), which amasses the most student data on campus. The final unit identified as a potential partner was Academic Technology Services, which had data on usage of the learning management system. Data from all three units would give a fuller picture of what role the library might play in student performance and identify gaps in IL instruction across disciplines.

When Institutional Research & Assessment named a new director and hired additional employees, there was a cultural shift from guarding data to sharing data. After the change in leadership, the unit became more willing to collaborate and more open to innovative approaches for delivering data for department and program assessment. The office now engages in collaborative initiatives not only on campus but also with feeder colleges and high schools. The new hires welcomed the chance to provide large data sets for analysis of student performance. This change in culture helped move the library project forward and created an opportunity for future collaboration.

A lack of interoperability, on the other hand, stalled collaboration with Academic Advising. While the advising unit was willing to collaborate, its use of a proprietary system meant it could not provide raw data.

Departmental culture played a role in the availability of data from Academic Technology Services. The authors hoped that Technology Services could supply data regarding student use of the learning management system, including log-ins, length of time on course sites, and click-throughs to library materials or links. The unit focused, however, on deploying instructional technologies and services to instructors rather than analyzing student learning. In this faculty-focused culture, the priority was not measuring student learning. Consequently, data from the learning management system collected by Academic Technology Services was not available for cross referencing.

The culture of the library also affected data collection. Like most library administrators, the dean of the library at CSULB was sensitive to the privacy of user data. For this reason, there had never been a large-scale effort to collect or keep data longer than necessary. The library kept only a running tally of circulating items connected to individual records. It regularly purged detailed loan information, interlibrary loan requests, and overdue fine history. At most, the library maintained five to six years of cumulative history on students' borrowing of monographs. It removed individual student account information immediately upon separation from the university. This policy extended to other services as well. Interlibrary loan reports were organized by journal title with no personal information attached to the records. Similarly, library system reports pro-

vided aggregated proxy log-ins and database usage with identifying fields encrypted or stripped. While some campuses advocate for the use of swipe technology to record information at service points, CSULB never asked for that level of individual data analysis.[18]

The most complete library data set derived from entries in its room-scheduling software, Meeting Room Manager. Across two different reservation platforms, the library had recorded the number of students, instructor names, duration of instruction, and course numbers for multiple years. Similar to other libraries, attendance is based upon faculty-supplied class size. It was difficult, however, to consistently and accurately calculate how many students attended library sessions or how many times a student came to the library for instruction while at CSULB. The library could amass data for library instruction by college, department, and course level and for the number of books a student checked out. The library could also identify students who created library passwords, required for remote access to databases. Based upon a willingness to collaborate, efforts to merge library and campus data sets focused on IR&A.

Takeaway: Develop Friends in High Places

For libraries, collaborating with campus units by sharing data is a win-win proposition. The majority of studies that discussed details of data collection worked with their version of campus institutional research. As at CSULB, libraries relied on these offices to combine their data sets with those from other units on campus and to anonymize the data.[19] At CSULB, the Office of Institutional Research and Assessment began to supply more user-friendly ways to access and analyze data, primarily with dashboards using Tableau software. The ability to include library data is attractive because it provides an additional measure for student success. Any library interested in collaborating with units on campus should leverage the large amount of data available in library systems by contributing to campus data sets. Library data have value that can be used for university planning and to gain insight into how the library supports the college curriculum. Any library embarking on this type of initiative should strive to collaborate with the units that handle the learning management system, advising, and enrollment data, in addition to the institutional research office.

Library data have value that can be used for university planning and to gain insight into how the library supports the college curriculum.

At CSULB, the changing culture on campus contributed to willingness to work with the library. With the incorporation of library data, departments and colleges could see which majors use the library most, what percentage of their students get formal library instruction, and where information literacy is concentrated in the General Education program and the academic majors. The university will also benefit from additional data to incorporate into accreditation reports. While collaboration was not possible across all units on campus, it is worthwhile to develop partnerships for future collaborative projects with these offices.

Support for Big Data Analysis

For many librarians interested in conducting library data analysis, the largest roadblock can be a lack of expertise. The advanced level of programming and statistical knowledge needed may lie beyond the skill set of many librarians. To prepare for this project, the primary investigator wrote a grant request to fund the combining of data from multiple data sets. The grant included monies to consult with a statistician and to hire students to clean the data. Finding expertise on campus can help contain the costs. Utilizing a professor from the Department of Mathematics and Statistics at CSULB who had experience using library analytics was far less expensive than hiring an outside consultant.[20] His expertise was using statistical analysis tools and applying predictive models to library data. Alan Safer assisted in constructing the request to IR&A, in developing the methodology, and in the final analysis of results. An additional benefit is his interest in participating as a coauthor in any publications ensuing from the project.

For many librarians interested in conducting library data analysis, the largest roadblock can be a lack of expertise. The advanced level of programming and statistical knowledge needed may lie beyond the skill set of many librarians.

An analytics project also requires knowledge of statistical software. Remaining grant funds were used to hire graduate students with expertise in statistical methods to troubleshoot problems with data modeling. Initially, three students worked on the project, but after the delays with getting the data sets, only one student continued. This student has become a coinvestigator on the project and, in addition to cowriting this case study, will take part in submitting results to conferences and academic journals.

Takeaway: Big Data Is Hard to Analyze on a Library Budget

The level of expertise needed to analyze data can be a burden if a library does not have house staff to do it. Another challenge in analyzing data is obtaining the technical support to do the analysis, which requires tools more sophisticated than commercial statistical software. Depending on the amount of data compiled, a small team of data analysts may be required. Without a grant to pay for students and a consultant to help apply advanced algorithms, the CSULB project would have failed. Likewise, many libraries cannot afford to spend time troubleshooting for an entire year. External funding was vital to achieve the ambitious goals of this project. Both staffing and money are necessary to set up a framework for collating and analyzing data in a continual assessment cycle. If the data are collected and cleaned regularly, merging data with other campus units becomes easier. Additionally, knowledgeable team members must remain on hand to supply technical support and check the programming to ensure accurate results.

Gaining Access to Data

Navigating the IRB Process

Another major hoop to jump through was gaining Institutional Review Board (IRB) clearance for the research project. Since at least the 1970s, the guidelines for institutional research have centered on preserving the rights and privacy of disadvantaged groups. In 1978, the Belmont Report by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research outlined three principles to guide the work of IRBs: respect for persons, beneficence, and justice.[21] IRBs at academic institutions are referred to as "locals" and serve to protect the human subjects for all research conducted at the university.

At CSULB, students, staff, and faculty members engaged in research must gain clearance from the Institutional Review Board for the Protection of Human Subjects.[22] Since the submission guidelines for IRBs vary by university, it is important to review guidelines from your campus. If more than one campus is involved, the primary investigator must obtain IRB clearance before research begins at the home campus, and participating universities must gain permission on their own campuses.[23]

Since the submission guidelines for IRBs vary by university, it is important to review guidelines from your campus.

Research involving secondary data is one of the least invasive data collection methods because the data do not come directly from human subjects and usually contain no identifiable information. This situation changes when gathering potentially sensitive personal data without formal student consent in big data research. Massive data sets of student information are unique in that they may contain identifiable details when cross-referenced with other sets of data.

Takeaway: Be Realistic about the Timeline and the IRB Review

When cross-referencing large sets of data, dealing with an IRB, and working with complex formulas, be realistic about the timeline. The initial request for data may take, as at CSULB, more than a year to complete. When establishing a method for regular data mining, apply for IRB clearance at least six months before beginning the request and collation of data.

Research how your system or university handles requests for student data before you start to design your data collection. After the initial research study, the IRB at CSULB modified its requirements. Researchers must provide the source of data, apply for administrative review, and include a formal letter from the unit providing data. While there are still no formal guidelines for big data ethics in the CSU System, the recognition that not all data are created equal ensures a higher level of scrutiny.

FERPA and Big Data

A second, just as significant consideration when using learning analytics is the protection placed on the use and access to student information by the Family Educational Rights and

Privacy Act of 1974 (FERPA). FERPA is a federal law that protects the privacy of student records. It applies to all schools that receive funds under an applicable program of the U.S. Department of Education.[24] Cross-referenced or longitudinal data for a student's academic career have the potential to conflict with FERPA due to the intricate level of information included. The secondary data may be anonymized on one level, but problems arise when large data sets with many unique data fields are amassed. Whereas most data used in research may give some demographic indicators, such research rarely violates the privacy of an individual. However, analytics assembles raw data from a variety of units across campus, which increases the danger of developing data sets that can be identifiable. Big data has the potential for providing so much information that even "anonymized data" can result in violations of privacy. Metcalf and Crawford note, "There have been several recent cases where de-identified data that was released publicly was able to be re-identified."[25] In these cases, FERPA safeguards may be inadequate to protect the privacy of students.

Big data has the potential for providing so much information that even "anonymized data" can result in violations of privacy.

The CSU System reviews all requests for potentially sensitive student data. The Office of the Executive Vice Chancellor and General Counsel for the CSU System must grant permission to prevent a violation of FERPA. The main concern from the legal office pertained to the amount of information collected at the student level. Because the data were longitudinal and included every academic movement of every student for many years, simply removing student IDs was not sufficient to protect privacy of individual students.

The Office of the Executive Vice Chancellor and General Counsel granted access to the requested student data, as long as two criteria were met. The first was that the sample sizes not be so small that they would unintentionally identify individual students. In the data analysis, student clusters smaller than 15 were discarded. Consequently, some identifiers were excluded from the data set, including participation in learning communities and program-specific sections, such as the honors program and the Educational Opportunities Program. To safeguard the confidentiality of students, all data supplied by the IR&A were stripped of student ID numbers.

The second stipulation was the primary purpose of the data. As the coordinator of assessment and information literacy, the primary investigator needed a letter of support from the dean of the library stating that the data would be used for practice and not solely for research. After many delays, the data request was approved halfway into the semester that had been allotted for the research project.

Takeaway: Student Data Privacy Is More Complex Than Imagined

A desire to help and improve services is often the primary reason for any research dealing with students. Predicting the best ways to insert library instruction and increasing students' usage of resources, materials, and even physical space are desirable but not the only considerations for libraries. Take into account the potential long-lasting effects

of learning analytics on student confidentiality. Stigmatizing students or making them feel unsafe should outweigh the benefits of increasing use of journals or streaming video services. To address this gap in the current guidelines, some universities have set up their own rules for the ethical use of large learning data sets.[26]

Before implementing big data, investigate the privacy rules at your own campus. If it lacks a policy specifically addressing the use of patron data, consider developing frameworks like those available on the Stanford website "Responsible Use of Student Data in Higher Education,"[27] or utilize checklists such as DELICATE (determination, explain, legitimate, involve, consent, anonymize, technical aspects, and external partners).[28] In addition to developing guidelines for staff, the library should also communicate to patrons how it uses information. San José Public Library has an excellent example on its website.[29] The most important role of a librarian is to ensure that students become lifelong learners who use libraries and explore information without fear of their actions being utilized in a hurtful way. Security and privacy are top priorities, including how the library follows campus protocols to combine data and protect students' privacy.

The most important role of a librarian is to ensure that students become lifelong learners who use libraries and explore information without fear of their actions being utilized in a hurtful way.

Best-Laid Plans: Analyzing the Final Data Set

The authors gathered more than 1 million records consisting of all freshmen cohorts from 2012 to 2016 and all transfer cohorts for the same years. Student data for over 50 unique courses with multiple sections were included. IR&A merged library instruction data with cohorts by instructor name and course start time and then removed all unique student markers (names and student IDs) before providing files to the library.

Takeaway: Expect the Unexpected

After overcoming the hurdles of developing a methodology and data collection plan, navigating the IRB, and getting the data sets needed for analysis, hidden issues may only be revealed after running preliminary data queries. At CSULB, the team ran test queries against the data set to determine the level of cleaning needed and reveal any incomplete fields in the data. When a test was done to see how many total students attended library instruction sessions for all years, the results were off by hundreds of students because data points selected to merge the data were missing. This underscores the importance of testing analysis from multiple access points, which may reveal huge errors in the data.

The greatest challenge to overcome was the inconsistent quality of data supplied by the library's systems. Much time was required to clean library-supplied data, which contained incomplete fields and variances that hindered the ability to merge library data with other campus data. Librarians entered data manually, which increased the likelihood of errors or omissions.

What Librarians Can Learn from This Experience

This research project was a learning opportunity not only for the authors but also for any librarian who wants to attempt to use big data to analyze library usage. Application of big data has limits but can be effective as a tool for curriculum planning and integration of library services. Accreditation bodies for most U.S. colleges and universities require evidence of student learning of information literacy. Big data can show patterns and numbers but not provide context and evidence of learning. It is better to regard data as a factor in decision-making rather than the sole justification for action. Using big data to show patterns in conjunction with analysis of student work offers a better picture of how students use the knowledge they gain from library usage or attendance at IL sessions led by librarians.

The library at CSULB will use the data from this project to examine patterns of library instruction in specific majors and in the General Education program. Almost every major at the university has a departmental learning outcome related to information literacy. Identifying the number of majors who have attended library sessions and which courses have a research component can support program assessment and accreditation of some departments. Results were not compiled for the first set of research questions until the end of 2019, two years after the project was scheduled for completion. From this journey, the authors hope librarians can identify units on campus for data support, consider the role campus culture may play in data collection, understand the issues with the IRB and using big data for analysis, and troubleshoot issues in analytics.

Tiffini A. Travis is an adviser for information literacy and library instructional assessment at the University Library, California State University, Long Beach; she may be reached by e-mail at: tiffini.travis@csulb.edu.

Christian Ramirez is a graduate student in the Department of Math and Statistics at California State University, Long Beach; he may be reached by e-mail at: christianramirez138@gmail.com.

Notes

1. Larry Johnson, Rachel Smith, Holly Willis, Alan Levine, and Keene Haywood, "The 2011 Horizon Report," New Media Consortium, 2011.
2. "7 Things You Should Know about IPAS [integrated planning and advising services]," EDUCAUSE Learning Initiative, November 5, 2014, https://library.educause.edu/resources/2014/11/7-things-you-should-know-about-ipas.
3. Megan Oakleaf, *The Value of Academic Libraries: A Comprehensive Research Review and Report* (Chicago: American Library Association, 2010), http://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/value/val_report.pdf.
4. Graham Stone, David Pattern, and Bryony Ramsden, "Library Impact Data Project," *SCONUL* [Society of College, National and University Libraries] *Focus* 54 (2012): 25–28.
5. Melissa Bowles-Terry, "Library Instruction and Academic Success: A Mixed-Methods Assessment of a Library Instruction Program," *Evidence Based Library & Information Practice* 7, 1 (2012): 82–95.
6. Ben Showers, *Library Analytics and Metrics: Using Data to Drive Decisions and Services* (London: Facet, 2015).

7. Mary O'Kelly, "Correlation between Library Instruction and Student Retention," *Presentations* 55 (2015), https://scholarworks.gvsu.edu/library_presentations/.
8. Andrew Asher, "Evaluating the Effect of Course-Specific Library Instruction on Student Success," 2017, https://scholarworks.iu.edu/dspace/handle/2022/21277.
9. Krista M. Soria, Shane Nackerud, and Kate Peterson, "Socioeconomic Indicators Associated with First-Year College Students' Use of Academic Libraries," *Journal of Academic Librarianship* 41, 5 (2015): 636–43, doi:10.1016/j.acalib.2015.06.011; Tiffany LeMaistre, Qingmin Shi, and Sandip Thanki, "Connecting Library Use to Student Success," *portal: Libraries and the Academy* 18, 1 (2018): 117–40, doi:10.1353/pla.2018.0006; John K. Stemmer and David M. Mahan, "Investigating the Relationship of Library Usage to Student Outcomes," *College & Research Libraries* 77, 3 (2016): 359–75, doi:10.5860/crl.77.3.359.
10. Joni Blake, Melissa Bowles-Terry, N. Shirlene Pearson, and Zoltan Szentkiralyi, "The Impact of Information Literacy Instruction on Student Success: A Multi-Institutional Investigation and Analysis," Greater Washington Library Alliance, 2017, https://scholar.smu.edu/cgi/viewcontent.cgi?article=1015&context=libraries_cul_research.
11. *EDUCAUSE Review* 47, 4 (2012).
12. Kyle M. L. Jones and Dorothea Salo, "Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads," *College & Research Libraries* 79, 3 (2018): 304–23.
13. D. Grant Campbell and Scott R. Cowan, "The Paradox of Privacy: Revisiting a Core Library Value in an Age of Big Data and Linked Data," *Library Trends* 64, 3 (2016): 492–511.
14. Jacob Metcalf and Kate Crawford, "Where Are Human Subjects in Big Data Research? The Emerging Ethics Divide," *Big Data & Society* 3, 1 (2016), doi:2053951716650211.
15. Celeste Lawson, Colin Beer, Dolene Rossi, Teresa Moore, and Julie Fleming, "Identification of 'At Risk' Students Using Learning Analytics: The Ethical Dilemmas of Intervention Strategies in a Higher Education Institution," *Educational Technology Research and Development* 64, 5 (2016): 957–68.
16. M. Brooke Robertshaw and Andrew Asher, "Unethical Numbers? A Meta-Analysis of Library Learning Analytics Studies," *Library Trends* 68 (2019).
17. United States Department of Education, National Center for Education Statistics (NCES), "Library Statistics Program: Academic Libraries," http://nces.ed.gov/surveys/libraries/academic.asp; NCES, "Integrated Postsecondary Education Data System," http://nces.ed.gov/ipeds.
18. Michelle Burke, Amelia Parnell, Alexis Wesaw, and Kevin Kruger, "Predictive Analysis of Student Data," National Association of Student Personnel Administrators, 2017, https://www.naspa.org/rpd/reports/predictive-analysis-of-student-data.
19. LeMaistre, Shi, and Thanki, "Connecting Library Use to Student Success."
20. Lesley S. J. Farmer and Alan M. Safer, *Library Improvement through Data Analytics* (London: Facet, 2017).
21. United States Department of Health, Education, and Welfare, National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, "The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research," 1978.
22. Jason Wang, Mary Walker, and Tiffany Rose, "CSULB IRB [California State University, Long Beach, Institutional Review Board] Submission & Review Process," 2018, https://www.csulb.edu/sites/default/files/groups/office-of-research-and-sponsored-programs/content_orsp_irbworkshop-spring2018.pdf.
23. Ibid.
24. "Family Educational Rights and Privacy Act (FERPA)," 20 U.S.C. § 1232g; 34 CFR Part 99 (1974), https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.
25. Metcalf and Crawford, "Where Are Human Subjects in Big Data Research?"
26. Stanford Data Science Initiative, Stanford University, "Ethics and Data Science," https://sdsi.stanford.edu/about/ethics-and-data-science.

27. Stanford Graduate School of Education, Stanford University, "Responsible Use of Student Data in Higher Education—A Project of Stanford CAROL (Center for Advanced Research through Online Learning) & Ithaka S+R," http://gsd.su.domains/.
28. Hendrik Drachsler and Wolfgang Greller, "Privacy and Learning Analytics: It's a DELICATE [determination, explain, legitimate, involve, consent, anonymize, technical aspects, and external partners] Issue: A Checklist for Trusted Learning Analytics," in *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (New York: Association for Computing Machinery, 2016), 89–98.
29. San José Public Library, "Our Privacy Policy," 2018, https://www.sjpl.org/privacy/our-privacy-policy.