The Brugd Undergraduate Student Data Solution: A Data Management Collaboration Between Global Environmental Sciences and Hamilton Library at the University of Hawai‘i at Mānoa

Oscar Ramfelt
University of Hawai‘i at Mānoa
Department of Oceanography
Hawaii Institute of Marine Biology

Michael W. Guidry*
University of Hawai‘i at Mānoa
Global Environmental Sciences
*corresponding author: guidry@hawaii.edu

Jonathan S. Young
University of Hawai‘i at Mānoa
Hamilton Library

I. Introduction

Our collective effort was to develop a sustainable means of combining past, present, and future institutional and programmatic-collected sources of undergraduate student data, specifically for the Global Environmental Science Program, into a common database for analysis and visualization. The resulting database also needed to be anonymized to both address student privacy concerns and so that resulting analyses could be easily shared and communicated amongst researchers, faculty, and other program stakeholders.

II. Needs of the GES Program

The Bachelor of Science degree in Global Environmental Science (GES) at the University of Hawai‘i at Mānoa (UHM) is administered by the Department of Oceanography in the School of Ocean and Earth Science and Technology (SOEST). In addition to the degree's general education and major-specific curricular requirements, all GES students must complete a faculty-mentored research experience, write a thesis, and publicly present their research findings in a symposium. Given the added value the thesis experience represents -- and the accompanying investment on part of the students, faculty mentors, and the program -- the tracking of students (demographics, academic progress, research progress, etc.) is one tool the GES Program can leverage to monitor and improve student retention and success. With access to the relevant data, tracking allows for both real-time (instantaneous) and longitudinal (historical or long-term) assessment of student outcomes and trends. The ability to track major trends as students enter, move through, and exit the program (either with or without their GES degree) is necessary to ascertain how well newly implemented program initiatives are working and also to identify potential areas for improvement, monitoring, and assessment.

Instantaneous vs. Longitudinal Data Needs

The GES Program needs (1) instantaneous information (defined as student data accessed during the current, active semester) to assess the immediate needs of currently enrolled students in the program, and (2) longitudinal information (meaning data from prior semesters) to assess longer-term trends and impacts of program requirements, new initiatives, etc. on the GES student population. Having a database

that can serve both these needs centralizes and economizes the data tracking effort and therefore more robust and frequent analysis.

III. Library Data Management Skills and Services

As universities move into the data science paradigm, there have been increasing needs along the lines of what the GES program wanted, for departmental decision making, but also for student tracking, learning analytics, and improving educational outcomes. The needs presented by the GES department have been shown to be able to be used to improve retention rates and student experiences.[1] Yet, often the expertise to design and implement these data systems are not readily available to departments. University IT departments are tasked with a different set of objectives. But there have been calls for academic libraries to provide this service, as libraries have long dealt with data management, and have built relationships with departments and faculty.[2]
Librarians are well suited especially to advise on the ethical and privacy issues inherent in data management. There is an inherent conflict in educational data science between the use of data to improve programs, and the rights and expectations of students to their private information.[3] Libraries have long been very careful at protecting the privacy of their users, and therefore can serve as a well respected counsel against initiatives that students, exposed to the excesses of the last decades of social media, may be wary against.[4]

In January of 2020, the GES program contacted librarians at UHM's Hamilton Library with questions regarding the possibility of designing a database system for tracking student progress through the program. Hamilton Library is organized into subject departments, each responsible for liaising with colleges in the university, as well as technical support departments. There is no specialized position or department in the UH library system with responsibility for data management service to the university. Instead, liaison librarians provide services based on their own skills and interests. The GES liaison librarian invited the Natural Sciences liaison and the Acquisitions librarian

---

[1] Sara de Freitas et al., "Foundations of Dynamic Learning Analytics: Using University Student Data to Increase Retention," *British Journal of Educational Technology* 46, no. 6 (2015): 1175–88, https://doi.org/10.1111/bjet.12212.
[2] Carol Tenopir, Ben Birch, and Suzie Allard, "Academic Libraries and Research Data Services: Current Practices and Plans for the Future," *An ACRL White Paper*, June 1, 2012, https://trace.tennessee.edu/utk_dataone/20.
[3] Tami Oliphant and Michael R. Brundin, "Conflicting Values: An Exploration of the Tensions between Learning Analytics and Academic Librarianship," *Library Trends* 68, no. 1 (2019): 5–23, https://doi.org/10.1353/lib.2019.0028.
[4] Kyle M. L. Jones et al., "A Comprehensive Primer to Library Learning Analytics Practices, Initiatives, and Privacy Issues," SSRN Scholarly Paper (Rochester, NY: Social Science Research Network, 2020), https://papers.ssrn.com/abstract=3567955.

to participate in several development meetings in 2020 to assist the GES program in this project. This group provided general data management advice, and demonstrated principles of designing a database system to the GES personnel tasked with implementing the solution.

IV. Solution - Brugd Development and Technical Report

Based on recommendations from meetings with the Hamilton librarians as well as various needs and requirements by the GES program, Brugd (Basking Shark in Swedish) was developed. The Brugd program sets out to solve the three primary issues related to data collection and analysis which were 1) merging data from different sources e.g. STAR or manual inputs into cohesive units that could be imported into a database, 2) protecting the privacy of the GES students, and 3) providing a platform to do this in a systematic manner. Brugd solves these issues by providing several pipelines each of which is related to a different “type” of student data. These pipelines are structured in a way where they usually provide one or more anonymized output files that can easily be imported into a database of the user's choosing. A general overview of the program's inputs and outputs can be seen in Figure 1. The following sections describe various aspects of the Brugd program.

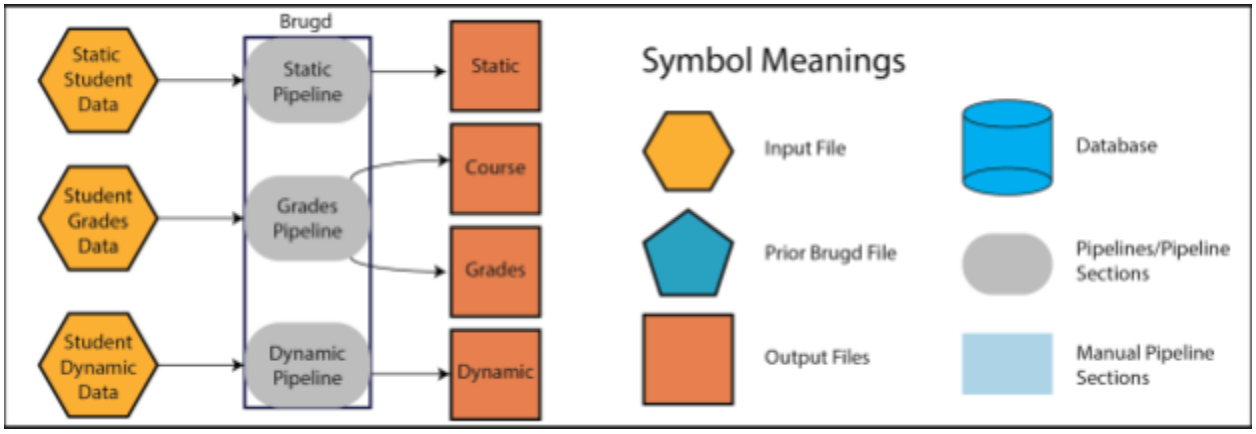

Figure 1. Summary of the Brugd program showing input and output file for each pipeline. Also shows symbols used in Figures 1, 2, 3, and 4.

The following sections describe the method used to protect student privacy and the three primary Brugd pipelines, the student static pipeline, the student grades pipeline, and the student dynamic pipeline.

A. Student Privacy

A central requirement for any pipeline that deals with student information is privacy. The GES Program needed the ability to handle student data and share general conclusions without putting student's privacy at risk. However, student information needed to be linked to other facets of a student's learning outcomes e.g. whether success in one course is correlated with success in another.

To solve this issue Brugd provides an Anon ID for each student that it encounters. This anon ID works similar to a student ID in that it is unique to each student i.e. there is a 1-1 relationship between student IDs and Anon IDs. However, the Anon ID value is selected randomly for each student meaning that there is no way to identify who a particular student is based solely on the Anon ID.

However, since students will likely be encountered multiple times by the program as they progress through an education program (in this case GES) there must be a way to repeatedly give students the same Anon ID based on their student ID. This is solved by a file (hereafter referred to as the index file) that contains the relationship between each student ID and each Anon ID. This file is used by the program to interpret both whether a student has already been encountered by the program and if so what Anon ID should be given to them.

This method allows the collection of longitudinal data without putting student's privacy at risk. However, notably the location where the index file is stored is of the utmost importance since it both contains a record of which student's have already been encountered by the program and a method to revert the Anon IDs in the database back into the original student IDs. The Brugd program does not implement a specific solution for the storage of this file and instead leaves it up to the user to provide a safe storage space for the file.

B. Brugd Static Pipeline

The student static pipeline is the core of the Brugd program and its output is key to both the repeated use of the program. When this pipeline is run it helps prepare new students to be identified throughout the rest of the program by adding them to both the index file. It also provides an output file called with information considered to be largely static for a student throughout their program. This output file can then be used as the basis for any new database or if this is a repeated run can be used to append additional rows to a database table focused on static student information. An overview of the pipeline can be seen in Figure 2.

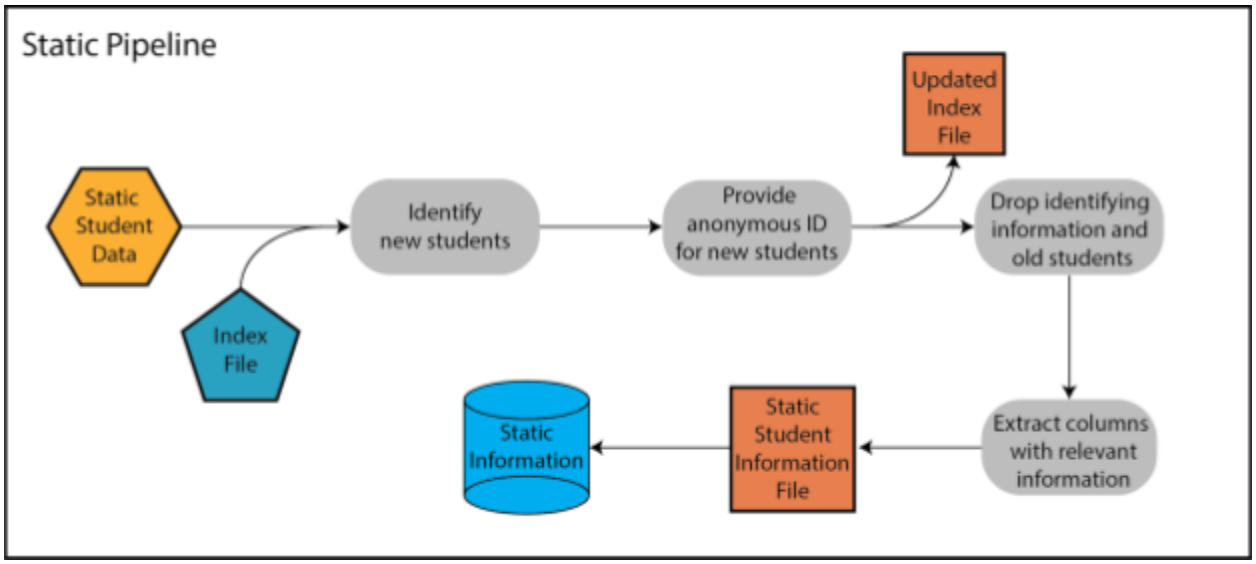

Figure 2. Summary of steps within the Brugd static pipeline.

C. Brugd Grades Pipeline

Once a cohort of students have been introduced to the program via the static pipeline and added to the index file various other pipelines can be run. A key pipeline within the Brugd program is the grades pipeline. It provides a method to systematically extract information from a STAR grades data request, anonymize it, and then format the output into two separate files. One file contains the information on the Anon ID, the semester CRN (i.e. the course CRN concatenated with the semester code), and the grade that the student received in the course. The other file contains information on each of the courses taken by at least one student. This file exists to provide additional context to the semester CRN, i.e. what course number it was or what section of the course it was.

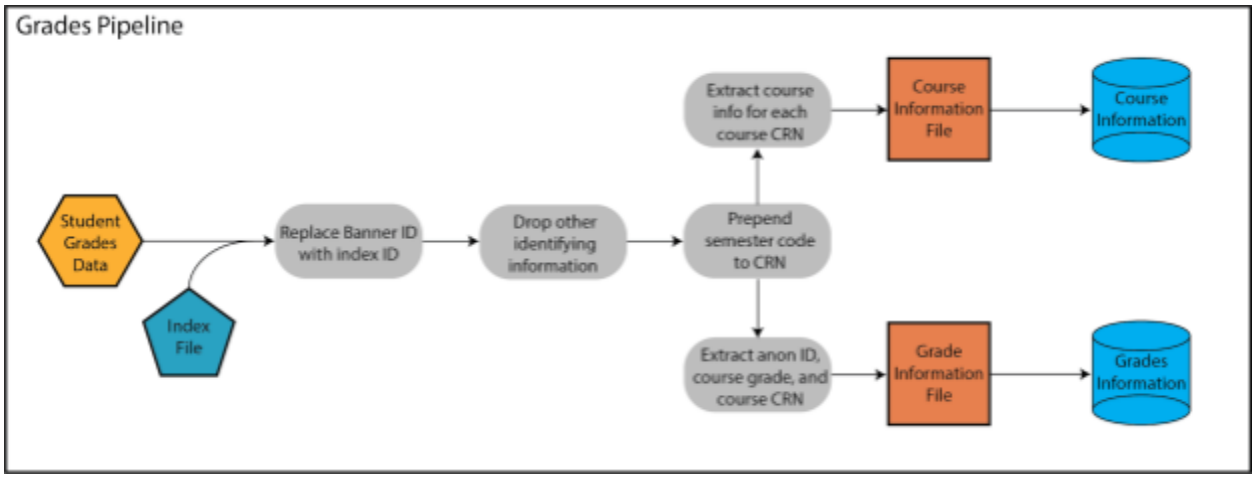

Figure 3. Summary of steps within the Brugd grades pipeline.

D. Brugd Dynamic Pipeline

The final pipeline is unique in that it does not anonymize student data immediately. This was due to there being portions of a student's program progress that cannot be queried from STAR, these include for example the amount of credits that a student entered the GES Program with or how a student entered the GES Program. In order to fulfill this requirement while also maintaining student privacy this pipeline allows administrators to anonymize the dynamic student information just prior to it being formatted for import into a database.

The cyclical section (i.e. the dynamic pipeline portion) of the flowchart exists to allow administrators to keep previously manually entered data columns while also updating other columns that extract information from STAR pulls on e.g. graduation. This allows the final dynamic information table to contain information that originates both from within STAR but also allows for the ability for administrators to focus on particular aspects of a student that STAR might not actively keep track of. Like when a student finished their thesis topic or what their thesis topic was in. Then once the administrators are prepared to format the data for import into a database they can run the anonymization pipeline that takes care of removing any identifying information and replacing the student ID with an Anon ID.

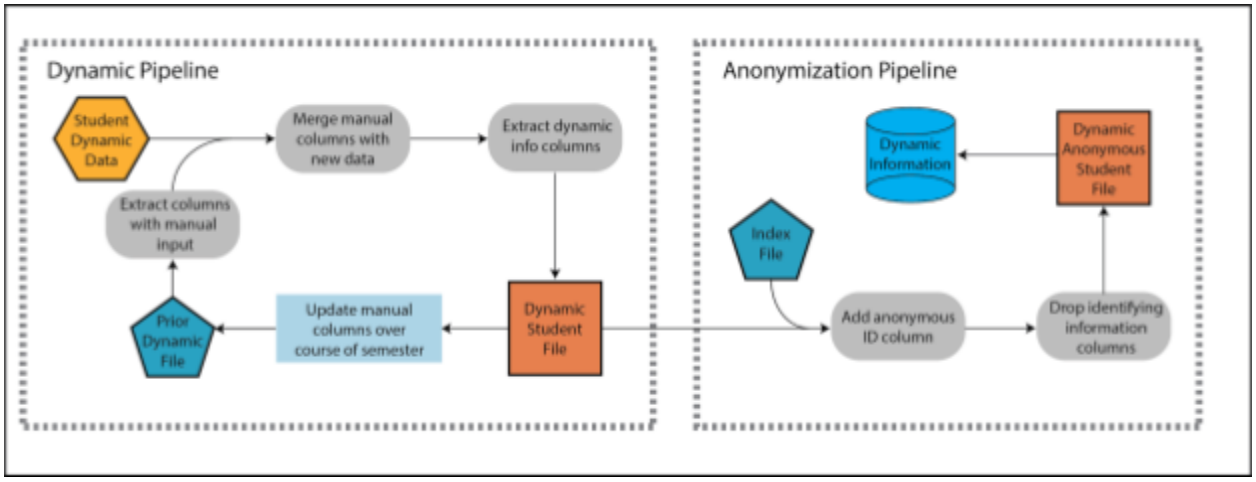

Figure 4. Summary of steps within the Brugd dynamic pipeline and subsequent anonymization.

Brugd Summary

The Brugd program contains various pipelines that focus on different aspects of a student and help provide a systematic and scalable solution for administrators to take data originating from STAR and enter it into a relational database. The resulting data is anonymized to protect student privacy and is set up in a manner to allow for greater flexibility when running analyses. Lastly, the Brugd program also allows a certain

amount of flexibility when it comes to custom columns with the dynamic pipeline. By providing columns and a method to incorporate them with the overall program administrators can supplement STAR data with manual columns that contain data not from STAR. This maintains a largely systematic method to incorporate data from STAR while also providing the flexibility for academic programs that might have additional requirements.

V. Moving Forward in GES - Expected Benefits

Our next steps moving forward are to begin conducting analyses in Microsoft® Access® of the relational database created by Brugd. As part of this analysis effort, and in parallel with using MS Access for analysis, we are beginning to use Microsoft® Power BI® for data visualizations. Figure 1 is an example of one way Power BI® can be used to visualize layered data to provide additional context. We are interested in investigating subjects such as student retention, persistence, and other outcomes. Power BI® also makes sharing visualized data (e.g., enrollment, student outcomes, etc.) relatively straightforward via connection to a program's website. We envision making visualizations of interests available to the public and also begin producing programmatic data reports using these visualizations.

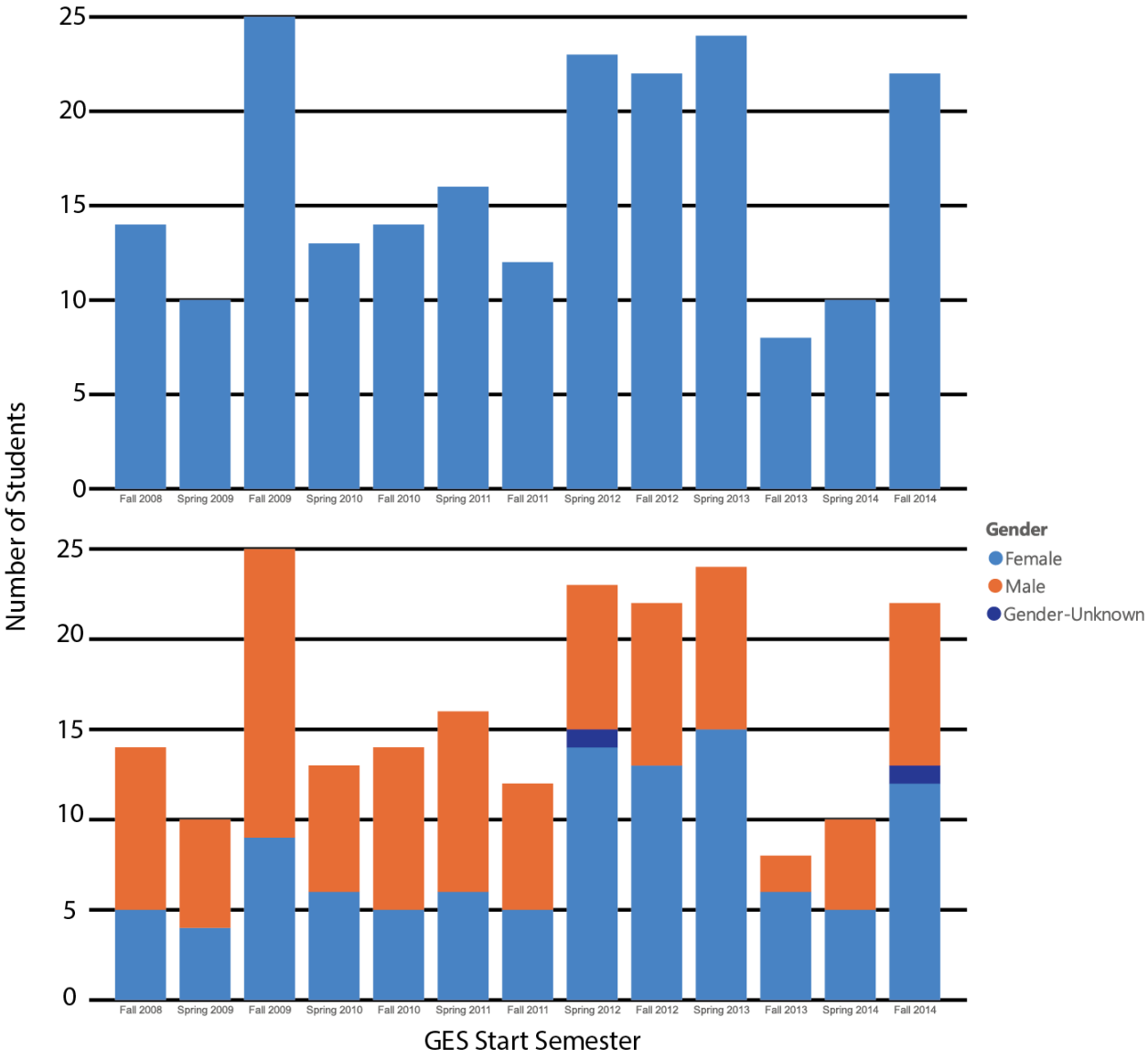

Figure 5. Example of using Power BI to provide additional context by layering data of student start semester in GES, number of students, gender of the students, and the student's start semester.

VI. Conclusion - Moving Forward with Library Data Management Services at UH

It might be asked why this case study includes much discussion at all of the role of the library. After all, the Brugd project was almost entirely completed by GES, and its success is a testament to their vision and hard work. However, just like many library endeavors, in this case, the assistance of the library at the start, while small and limited in duration, can be seen as a catalyzing force, a small directional guide that assists in the further independent development of the project.

This model can be used as a template for outreach efforts at other departments across the university and at other institutions. While a dedicated data management services librarian has been used at other institutions, UHM's distributed model can still produce successes. The greatest barrier in this model is awareness, in this case, the partnership happened fortuitously due to pre-established strong ties between the library liaison and the GES Program, but other departments may have similar needs but not realize the potential for the library to assist. This paper is one means to address that, but in the future the UHM librarians should also create a more centralized list of data service skills available.

VII. Appendix - Brugd Tables

Lists the various columns that are found in the output files by the Brugd pipelines. The symbol (*) indicates that the information in a column originates directly from STAR. Tables represent suggested database tables to be used alongside Brugd.

VIII. Appendix - Brugd Documentation

Contains the in depth documentation for the Brugd program that describes in detail the usage of each of the pipelines as well as other details.

IX. Bibliography

Freitas, Sara de, David Gibson, Coert Du Plessis, Pat Halloran, Ed Williams, Matt Ambrose, Ian Dunwell, and Sylvester Arnab. "Foundations of Dynamic Learning Analytics: Using University Student Data to Increase Retention." *British Journal of Educational Technology* 46, no. 6 (2015): 1175–88. https://doi.org/10.1111/bjet.12212.

Jones, Kyle M. L., Kristin A. Briney, Abigail Goben, Dorothea Salo, Andrew Asher, and Michael R. Perry. "A Comprehensive Primer to Library Learning Analytics Practices, Initiatives, and Privacy Issues." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, 2020. https://papers.ssrn.com/abstract=3567955.

Oliphant, Tami, and Michael R. Brundin. "Conflicting Values: An Exploration of the Tensions between Learning Analytics and Academic Librarianship." *Library Trends* 68, no. 1 (2019): 5–23. https://doi.org/10.1353/lib.2019.0028.

Tenopir, Carol, Ben Birch, and Suzie Allard. "Academic Libraries and Research Data Services: Current Practices and Plans for the Future." *An ACRL White Paper*, June 1, 2012. https://trace.tennessee.edu/utk_dataone/20.