Download, Clean, Repeat: Creating a Sustainable Data Dashboard Workflow

Kaypounyers Maye, Scholarly Engagement Librarian for Social Sciences and Data, Tulane University
Megan Sheffield, Interim Head of Open Scholarship, Data Services Librarian, Clemson University
Kelsey Sheaffer, Creative Technologies Librarian, Clemson University
Megan Palmer, Assessment Librarian, Clemson University
Christopher Vinson, Head of Digital Strategies, Clemson University

Purpose and Goals

Clemson Libraries recognized the need to centralize and present our data related to spaces, collections, personnel, and services after a series of internal activities related to benchmarking our organization against a set of similar peer institutions with Carnegie R1 status. The resulting report showed the significance of collecting and consolidating data from multiple sources to create captivating visualizations and to construct a narrative that demonstrated our current state and future aspirations as a new R1 library to stakeholders and users. While the R1 report was a useful model, it was nevertheless a static document that captured the condition of our organization at a particular moment in time. In order to build on the momentum gained by the R1 report and evolve our narrative, we realized the need for a project to establish best practices and workflows for ongoing data collection with minimal duplication, identify tools for data manipulation and visualization, and develop an accessible platform for easy retrieval and display of up-to-date data on our website—a data dashboard. This need for data transparency became more apparent after the announcement that Clemson would soon move to a responsibility-based budgeting (RBB) model, in which the Libraries would continually be expected to show evidence of its value and impact to campus administrators. Thus, our overall purpose with this project was to determine the staff time commitments involved and the tools and skills needed to develop a sustainable product, with the goal of producing a working prototype of a data dashboard for our internal and external audiences to test and provide feedback on for future development.

Findings

As is often the case, what seemed initially like a straightforward project quickly became more complicated. Although Tableau supports linking directly to data sources, our proprietary library systems do not currently have the corresponding functionality to accomplish this. As a workaround, we developed a schedule of specific reports to run and a shared Google Drive for these reports (mostly in the form of CSV files) to link to Tableau. We are currently partnering with our library IT to set up automated reports and university server space for the data so that it is not linked to an individual's Google account.

Practical Implications & Value

For our own institution, the dashboard will have the immediate value of providing quick visualizations for commonly requested statistics, such as the gate count for a certain date range, number of checkouts, or reference statistics. This will save time for employees that need these numbers, and present them in an official format consistent with university standards. More broadly, many academic libraries have either recently created or are in the process of creating data dashboards to communicate with stakeholders to show the impact of their work, and many libraries use the same software systems. Although no two institutions will have exactly the same dashboards, we hope to shed light on the work behind our dashboard so that others may use it as a starting resource.

Current Status

Although we were able to create the framework and proof of concept for our dashboard, we were less successful in making this work sustainable. Although all individuals involved see the value a dashboard could provide, no single person has the capacity to take on ownership of a project at this scale and the upkeep it would require. Currently, we are in a holding pattern as we try to identify ways to make this dashboard more sustainable; this may take the form of automating certain tasks (such as data cleaning) with the help of student labor or potentially hiring additional staff that can devote more time and expertise to the project. The work of this group was invaluable in identifying some of the potential challenges in combining different data sets into a single dashboard that will be used by stakeholders with varying priorities.

Methodology

1 *Administrative Preparation*

To determine the scope of the data dashboard, three project members met with three library administrators (Dean of Libraries, Head of Teaching and Learning, and the Head of Special Collections and Archives). Each administrators' answers were shared with the entire committee as they decided which (1) data to use, (2) how often to retrieve data, and (3) which dashboard visualizations to create.

2 *Extraction*

After completing the administration interviews, the project team met and identified the following data to extract (data source included) using the outlined extraction methods. Considering the variety of data, the extraction process for each dataset varied depending on the data's type and the extraction knowledge of the project members. Since each project member had varying degrees of expertise in this area, the team's long-term plan is to automate each data extraction using Python or some other tool.

Suggestions for Sustainability:

- Collaborate early with your system administrators to determine the best extraction approach.
- Automate as much extraction as possible to avoid data loss and error.
- Create an extraction schedule, preferably one that is automated
- Create a data model in case data needs to be collectively stored in a database

3 *Transformation*

Simple data transformation was conducted either using Excel or the pandas library in Python. The team hopes to automate a majority if not all of the transformation workflows in the future.

Suggestions for Sustainability:

- If project team members do not have the skills needed to perform data transformation, partner with a campus group that has this expertise.
- If you plan to use scripting to clean your data, use one programming language consistently. This code should be well documented with accompanying metadata.
- Determine which datasets are automatically produced and which are produced by staff. Reduce the number of staff-produced datasets to ensure data accuracy.

4 *Loading*

After being transformed in either Excel or using Python if necessary, all data was loaded to Google Sheets documents to be fed into Tableau for visualization. Each Google Sheet consists of 2 pages: the first page is a housekeeping page that identifies the person who updated the Google Sheet (Data Collector/Cleaner), the collection date (Date Collected), earliest record date, latest record date, the semester the data was collected, and the year of the data was collected. Data is updated at the end of each academic semester unless needed by library administration or librarians for special projects and programs. One project team member was also in charge of uploading the data to Tableau. This process was completed through a live connection between Google Sheets and Tableau.

Suggestions for Sustainability:

- Choose a visualization tool that is easy to use for all team members. If possible, provide training.
- If one of the data sources has the capability to process data for visualization, consider uploading all data to that one platform for easy access and use. Springshare's LibInsights could prove useful here.
- Loading should be automated whenever possible to free up time for other workflows.
- If possible, use a storage solution that is tied to the institution rather than an individual.

5 *Design*

The dashboard design is an ongoing process as librarians and administrators at Clemson University determine the appropriate way to represent the library's "story." However, the project members did identify the best visualization(s) for each data set included on the dashboard.

Suggestions for Sustainability:

- Determine visualization type before the design process
- Create prototypes to be reviewed by the library administrators
- Use your institutions color palette
- Visualization size should be proportionate to importance of the visualization.

Extraction Workflow

DATA SOURCE	DATA TO EXTRACT	EXTRACTION METHODS	Reason for Using Data
LibApps/Springshare	(i) Study and Meeting Space Statistics (ii) Library Guide Views (iii) Course Instruction Records (Not automatically generated) (iv) Virtual Chat Records	(i) Download and store (ii) Auto-Download link (iii) Auto-Download link (iv) Download and store	(i) Determine space modifications and library hours (ii) Determine popular library guides and modifications to library guides (iii) Determine information literacy standards being taught, departments with most library instruction, space considerations for instruction
Leganto	(i) E-Reserves	(i) Download and store	(i) Inform collection development
Alma/Alma Analytics	(i) Checkouts and Renewals (including material type field) (ii) Digital Requests (including vendor and database fields) (iii) Print v. Digital Circulation (count) (iv) Physical Holdings (including material type and location fields) (v)Digital Holdings (including material type field) (vi) Physical Reserves (including course subject field)	(i) Download and store (ii) Download and store (iii) Download and store (iv) Download and store (v) Download and store (vi) Download and store	(i) Inform collection development (ii)Inform collection development (iii) Inform collection development (iv)Inform collection development (v)Inform collection development (vi) Inform collection development
Illiad (InterLibrary Loan)	(i) Borrowing (including subject and material type) (ii) Loaning (including subject and material type)	(i) Download and store (ii) Download and store	(i) Inform collection development (ii) Inform collection development
WordPress	(i) Website Traffic	(i) Download and store	(i) Determine online interaction with libraries
Sensource	(i) Occupancy Data for all library locations	(i) Download and store	(i) Determine space usage for future library renovations

Transformation Workflow

DATA SOURCE	DATA	TRANSFORMATION PROCESS
LibApps	(i) Study and Meeting Space Statistics (ii) Library Guide Views (iii) Course Instruction Records (Not automatically generated) (iv) Virtual Chat Records	(i) No transformation needed (ii) No transformation needed (iii) Remove columns and standardize instructor name, course ID, and learning outcomes (iv) Remove columns that include user system information (IP address, web browser)
Leganto	(i) E-Reserves	(i) No transformation needed
Alma/Alma Analytics	(i) Checkouts and Renewals (including material type field) (ii) Digital Requests (including vendor and database fields) (iii) Print v. Digital Circulation (count) (iv) Physical Holdings (including material type and location fields) (v) Digital Holdings (including material type field) (vi)Physical Reserves (including course subject field)	(i) No transformation needed (ii) No transformation needed (iii) No transformation needed (iv) No transformation needed (v) No transformation needed (vi) No transformation needed Since Alma allows report customization, we were able to retrieve the exact information we needed.
Illiad (InterLibrary Loan)	(i) Borrowing (including subject and material type) (ii) Loaning (including subject and material type)	(i) No transformation needed (ii) No transformation needed
WordPress	(i) Website Traffic	(i) No transformation needed
Sensource	(i) Occupancy Data for all library locations	(i) No transformation needed

Design Workflow

DATA SOURCE	DATA	VISUALIZATION TYPE
LibApps	(i) Study and Meeting Space Statistics (ii) Library Guide Views (iii) Course Instruction Records (iv) Virtual Chat Records	(i) Bar chart (ii) Bar chart (iii) Bar chart, Donut plot, Line graph (over time) (iv) Bar chart
Leganto	(i) E-Reserves	(i) Line graph (over time), bar chart
Alma/Alma Analytics	(i) Checkouts and Renewals (including material type field) (ii) Digital Requests (including vendor and database fields) (iii) Print v. Digital Circulation (count) (iv) Physical Holdings (including material type and location fields) (v) Digital Holdings (including material type field) (vi)Physical Reserves (including course subject field)	(i) Line graph (over time), bar chart (ii) Line graph (over time), bar chart (iii) Line graph (over time), bar chart (iv) Line graph (over time), bar chart (v) Line graph (over time), bar chart (vi) Line graph (over time), bar chart
Illiad (InterLibrary Loan)	(i) Borrowing (including subject and material type) (ii) Loaning (including subject and material type)	(i) Line graph (over time), bar chart (ii) Line graph (over time), bar chart
WordPress	(i) Website Traffic	(i) Line graph (over time), bar chart (by day of the week)
Sensource	(i) Occupancy Data for all library locations	(i) Line graph (over time), bar chart (by day of the week)