Do We Collect That Information and If So, How Can I Access It? Designing a Statistics Depository

Michael Perry and Gina Petersen
Northwestern University, USA

Introduction

In the fall of 2015, Northwestern University Libraries restructured. The goals of the structure change included empowering frontline staff to make operational decisions and allowing administrators to spend more time focusing on strategic issues. Of course, all organizations require data in order to make strategic and operational choices; Northwestern University Libraries now had a renewed vigor for needing staff throughout the organization to access the information required to make the best decisions.

The library assessment specialist and the head of assessment and planning worked together to first understand the statistical and data landscape of the organization and then to facilitate access to and comprehension of this information. This paper outlines the process taken by the authors to conduct a data audit and create an infrastructure for storing and facilitating access to information.

Rationale

Aside from being new to their roles, the authors saw two principal reasons that an audit of data collection was needed: the decentralization of information collection and storage and a lack of knowledge regarding which people were responsible for which pieces of information.

Data collection and analysis does, and the authors believe should, occur throughout the organization. However, the authors wondered if there was duplication of effort. Are units A and B both pulling the same metric and if so, are they getting the same result? Northwestern University Libraries has a long-standing culture of democratizing information by giving many staff members access to the reporting features of products, such as LibAnalytics and Alma. Within such systems, it is important that the criteria for the reports are correct. For example, contrary to what may be intuitive, in-house uses must be manually excluded from reports of circulation numbers, as they are considered a type of circulation within the ILS. If unit A correctly excludes in-house uses and unit B does not, competing circulation numbers about the same collection could be disseminated throughout the organization.

Further, once information is generated, the authors wondered how it was stored. Do final fiscal year numbers live primarily in annual reports? Is it easy to compare changes year to year? Is the raw data stored in file formats and directories that allow others to access said information?

The decentralization of information collection also facilitated another problem; there was no good way to learn who the best person was to ask to generate a specific piece of information. The authors saw a need for a list of point people for various types of data. Frequently there would be an e-mail message sent to all supervisors asking who has data about headcounts in a specific space or computer use during interim periods. There was a need to describe what information is being collected where and ensure that the data is stored such that more than a single person has access to it.

After mining annual reports for specific pieces of data, the authors knew a lot of the information that was being collected and by whom, but realized that other pieces of data, which described operations and could inform decision making, did not rise to the level of being included in annual reports.

After considering these issues, the authors decided that they needed the help of others in order to conduct a more complete data audit and outlined four project goals for the data audit:

- To understand who was collecting data where and ensure that effort was not being duplicated.
- To clearly delineate who was responsible for collecting data within the library, which, in turn, will make it clear who the point person was for each piece of data.

- To develop and implement a central depository location. This, in turn, would make access to some pieces of data easier. Further, the authors hoped that by being able to analyze data side-by-side, additional insights could be developed.
- To establish a community of practice regarding data and its stewardship and analysis.

Process

The authors drafted, piloted, and distributed a data stewardship form. The form was distributed to all supervisors, who were asked to record information about the data their department compiled, generated, and kept. The criteria for submission included that the data: be generated by library or user workflows, be used for planning purposes, or be included in annual reports and statistics. For each piece of data the form asks for a name and description of the data, the system from which the data is generated, the schedule for compiling/pulling data, the file type(s) of the reports, the department responsible for the data, whether the data contained personally identifiable information, and where the data is stored. The full text of the form is available in Appendix A.

The authors received 55 submissions to the form. A small subset of the Library Assessment Committee reviewed all submissions while considering the following questions:

- Is the information clear?
- Is the response correctly coded for personally identifiable information?
- Does the record contain multiple data sources that need to be split up?
- Does there appear to be any missing information?
- Does this data appear useful for further analysis (such as meta-analysis, visualizations, etc.)?

Challenges

Upon first review, the data submitted provided a number of challenges. First, it was unclear who this information should go to. Some responses were clearly handled by supervisors while other departments had spread data collection and reporting among a number of different staff members. Additionally, the relationship between library administration and the collection of data seemed to vary based on the data source in question. Bringing together the analysis and creating more formalized processes would require communication at all levels across the organization.

As the analysis of the responses continued, it also became clear that Northwestern University Libraries were collecting a variety of data that often did not have a clear upstream purpose. We were often collecting data that did not appear immediately useful, sometimes based only on the idea that it might someday be useful. This seemed to also impact buy-in for more established data collection methods as it makes it unclear what is ultimately useful or what is not. Further, some information needed at the front-line level is not needed at the strategic level. There was a need to explore the use of specific data in-depth.

What we learned

The data stewardship form submissions revealed that Northwestern University Libraries is collecting data from 35 different systems and manually collecting at least nine data points. The manual collection is in some ways underreported as some of the systems from which data is pulled require that transactions (such as reference interactions) be manually added one at a time. A review of the systems revealed that there are occasions where multiple modes of data collection are used in order to gather and triangulate parts of what could as first be thought of as a single statistic.

Entrance and exit counts are one example of this. Users scan their university ID cards when entering the library. Affiliated users who arrive at the library without their ID fill out a paper form and, once verified as active, are allowed to pass through the gate without generating a record in the entry system. Meanwhile, visitors are issued a paper day pass which must be scanned by the barcode readers that scan IDs. Visitor entries do create a record in the entry system. The numbers of visitors is pulled from the visitor system. Further complicating matters, there are periods of time, such as orientation week, alumni weekend, and graduation, when the gates open. Anyone is free to visit the libraries during these periods. Therefore the entrance gate system data is useful for determining patterns of traffic (e.g., Are many students entering the library after 11 p.m. or have most already arrived? Is the library being used before noon on Sundays?). However, it does not generate a reliable number regarding the total number of users entering the space. Instead, the library uses the exit counts in order to have a grasp of the number of people using the library. The exit counts are generated by an infrared visitor counter. This system is also potentially unreliable as it has difficultly tracking multiple people exiting together and it provides no additional data about users.

At the onset of the project, the authors hoped to find areas of duplication that could be eliminated. Instead, they found areas of overlapping data.

What we did/are doing

Reviewing the work, the authors developed a plan to centralize and provide more accessibility to data. A page was created on the library's intranet outlining the data sources, an explanation of what data is contained, a primary contact for that system, and the location where data or reports can be found. Alongside this guide, a central repository for data was created using the shared network drive. This allows for varied control of permissions for data that includes personally identifiable information as well as more permissive access to data that might be useful across the library.

The authors also hope to utilize Tableau Server to display visualizations of key pieces of data, displaying trends and aggregates. This will be particularly helpful when the underlying data is formatted, labeled, or coded in such a way that there is a learning curve to understanding the outputs. These visualizations will allow staff and administrators to answer key questions without the requisite work initially needed to understand the underlying data.

Providing an inventory and access to data is only the first step however. Developing a culture that sees data analysis as a foundational element requires champions willing to demonstrate skills and help to train fellow staff members. The authors are exploring the creation of a data analysis group as an addition to the Assessment Committee. This group will aid in the stewardship of data, provide assistance in analysis and visualization, and aid in the preparation of reports and surveys. This group may also be helpful in the efforts to create more data-focused work group annual reports. Reflecting on the first year of the reorganized library, work group leaders were asked to identify five to seven key performance indicators for their work group. These may be metrics that are already being tracked and are part of our data inventory or new metrics that better reflect the new work group's focus. This may require work in altering data reporting, developing new data sources, or adopting new modes of data analysis. The data analysis group can aid in this work while also understanding its context in the larger data environment.

There is hope that collocating data will provide the opportunity to develop new insights by comparing data points which were previously separate. For example, is this blending data or simply having the ability to consider variable A in the context of variable B by virtue of easily being able to view trends of both data points? This could be achieved by displaying related information in one Tableau dashboard. Or more simply, it could be that, since staff members can easily access both data points from the intranet landing page, the relationships more naturally emerge.

Even if we are able to succeed in collocating many data points, there are, as there always are, technological challenges to displaying data in a central place.

Conclusion

The goal of this project was to better understand the data landscape of Northwestern University Libraries as it existed during a time of transition. The hope was that this would better inform the use of data in decision making and allow for new insight from comparing data that was previously siloed. Collecting information about what data sources existed and how data was reported started to show the scope of issue. What the authors found was a data landscape that included often redundant or unnecessary data that made reporting difficult and data that was stored in a variety of places often with limited accessibility. A plan was formulated to gather data in a single, accessible place and index the data sources and reports that are available. In conjunction with the formation of a data analysis group, the hope is that the library will be able to move in the direction of more informed data-driven decision making and glean new insights from a more robust analysis plan.

Appendix A: Data Stewardship Form

1. Data: *What is the data?*

2. Data Description: *Description of data*

3. System(s): *Through what system(s) is the information generated?*

4. System Administrator(s): *Work group in charge of those systems*

5. Calendar Type: *Is data reported on the academic or fiscal calendar? Most user data should be reported on the academic calendar.*

6. File type(s) of reports: *What type(s) of file(s) are generated when this information is pulled/compiled?*

7. Work group(s) responsible for depositing data: *What work group is responsible for pulling this data?*

8. Personally Identifiable Information: *Does data contain personally identifiable information about users?*

9. Deposit Location: *Where is the data (once pulled/compiled) currently stored/saved?*

10. Written procedure for extracting and reporting data: *Optional*