What Do We Collect and Why? Conducting a Self-Study to Improve Data Collection Practices

Summer Durrant and Suzanne Chase
University of Mary Washington, USA

Abstract

This paper describes a self-study the University of Mary Washington Libraries conducted to examine our data collection practices. A small working group was formed to interview all full-time library staff about the data they routinely gather and challenges they experience when collecting and using library data. The results were used to compile a data inventory spreadsheet that contains detailed information about 81 library metrics. The findings were also used to improve data collection practices in several key areas, including automating data collection that was previously gathered manually, simplifying data collection tools, and centralizing where data files are stored. Consequently, the libraries are now better positioned to make effective use of the data collected about our library resources and services.

Introduction

Academic libraries collect a tremendous amount of data about their resources and services; however, making effective use of these data can be challenging. One obstacle is that data collection responsibilities are often decentralized in libraries, making it difficult to know exactly what data are collected, who is responsible for collecting them, and how they are used. Further, data are stored in multiple places, including library systems, vendor administrative portals, shared storage spaces, and on individual staff computers. This can make locating data a complicated and time-consuming process. Lastly, making use of data requires specialized knowledge of data management, data analysis, and data visualization techniques and best practices, which are not skills librarians typically receive training in. These challenges are particularly acute for small academic libraries that often do not have a full-time position devoted to assessment activities.

This paper describes a self-study the University of Mary Washington (UMW) Libraries completed during the 2015–16 academic year to improve our data collection practices. The impetus for this project came after a daylong strategic planning retreat in which library staff expressed an interest in developing a better understanding of what data are collected and how they are used. Staff agreed that a holistic view of our data collection practices would help the libraries make better use of our data for decision making and reporting purposes. For this self-study, the libraries formed a small working group, the Data Collection & Analysis Team (DCAT), which included staff from different library departments (e.g., Access Services, Collection Services, Technical Services, Special Collections & University Archives) to ensure all perspectives were represented. The goals of the self-study were to: (1) compile an inventory of all library metrics collected by staff and systems, (2) devise and implement strategies to standardize and optimize our data collection methods, and (3) assist staff with analyzing data to make informed decisions and demonstrate impact.

Methodology/Approach

To examine our data collection practices, DCAT interviewed all 26 full-time library staff members. We developed a semi-structured questionnaire (see appendix) to ensure all staff members answered the same set of questions. The first part of the questionnaire solicited information about the types of data staff collect. Staff were asked to identify the data they are responsible for collecting, and to provide a brief description of the data, along with information on the collection method, storage and access location, frequency, and how the data are used (e.g., for reporting to external organizations). When possible, staff were asked to provide examples of the data, such as a spreadsheet or system-generated report. The second part of the questionnaire contained four questions that focused on specific challenges staff experience when collecting or using data. Lastly, staff were asked to reflect on how the libraries could improve our data collection practices. A separate interview, with

slightly modified questions, was conducted with the university librarian.

Information gathered from the first part of the questionnaire was compiled into a spreadsheet. We used the card sorting software, OptimalSort, to categorize the library metrics into functional areas. The final product, the data inventory spreadsheet, is discussed in more depth in the next section. Responses to the second part of the questionnaire, which focused on challenges staff face when collecting or analyzing data, were coded and analyzed in NVIVO. Key themes that emerged during staff interviews are discussed in the Staff Challenges section.

Data Inventory Spreadsheet

The data inventory spreadsheet contains detailed information about 81 library metrics. These metrics are organized into four broad functional areas: *staffing*, *expenditures*, *collections*, and *services*. Three of the functional areas are further subdivided: *expenditures* into *general*, *wages and salaries*, *collections*, and *other operating*; *collections* into *holdings*, *usage statistics*, and *collection maintenance*; and *services* into *information services*, *library spaces and equipment*, *collection services*, *discovery and access*, and *marketing and outreach*. Figure 1 shows the data inventory spreadsheet organized into functional areas.

Figure 1: Data inventory spreadsheet organized into functional areas

Library Metric	Brief Description	Library Unit	Collection Method	Data Source	Compiled	Purpose	Reporting
Staffing							
Expenditures							
General							
Wages & Salaries							
Collections							
Other Operating							
Collections							
Holdings							
Usage Statistics							
Collection Maintenance							
Services							
Information Services							
Library Spaces & Equipment							
Collection Services							
Discovery & Access							
Marketing & Outreach							

The spreadsheet has eight columns. The first column provides the name of the library metric followed by a column with a brief description adapted from the ANSI/NISO Z39.7-2013 Information Services and Use: Metrics and Statistics for Libraries and Information Providers—Data Dictionary[1] and the Project COUNTER Code of Practice.[2] The third column identifies the library department responsible for collecting the data. *Collection method* indicates whether the data are automatically captured in a system or manually collected by library staff. *Data source* provides the name of the system where the data are stored or the file format (e.g., Excel). The fifth column notes how often the data are aggregated, such as at the end of the fiscal year or academic semester. The final two columns describe how the data are used and whether they are reported to an external organization like ACRL or IPEDS. The *gate count* metric, for example, has a brief description of "total number of persons who physically enter the library." Access Services is the library unit responsible for collecting it. The data are collected manually in an Excel spreadsheet and compiled monthly as well as at the end of the fiscal year. The data are used for a variety of purposes, including making staffing decisions and demonstrating use of the physical library, and the metric is reported to ACRL. Figure 2 shows how the gate count metric appears in the data inventory spreadsheet.

Figure 2: Example of gate count metric from the data inventory spreadsheet

Library Metric	Brief Description	Library Unit	Collection Method	Data Source	Compiled	Purpose	Reporting
Services							
Library Spaces & Equipment							
Gate Count	Total number of persons who physically enter the library	Access Services	Manual	Excel	Monthly	External Reporting; Demonstrate Impact; Decision-Making	ACRL

A second spreadsheet was created to capture additional details about data collected manually in the library. This spreadsheet provides more in-depth information on a subset of 18 metrics, including the

collection method, file format, file location, history, and additional notes. Using the same example as above, the collection method for *gate count* is, "every morning a staff member records the gate count number on a paper calendar, which is later tabulated by month and fiscal year." The printed calendar is stored in a binder and eventually converted to an Excel spreadsheet that is stored on the libraries' shared network drive. The data are available from 1989 to present. Figure 3 shows the gate count metric as it appears in the second spreadsheet for manually collected data.

Figure 3: Example of gate count metric on second spreadsheet for manually collected data

Library Metric	Collection Method	Format	File Location	History	Notes
Access Services					
Gate Count	Every morning a staff member records the gate count number on a paper calendar, which is later tabulated by month and fiscal year	Paper Binder; Excel	N:\Public\Circulation\Circulation\Patron Counts	1989-present	The printed calendar is stored in a binder and eventually converted to an Excel spreadsheet

Staff Challenges

Examining the responses to the second part of the questionnaire, which focused on staff challenges to collecting and using data, revealed three main sources of frustration. First, many staff indicated that while they routinely collect data, they were unsure of how (or if) the data are used. Thus, there seemed to be a disconnect between staff collecting data and understanding how the data are used for making decisions or demonstrating the value of the library. Several staff also noted that they were unsure whether they should continue gathering data that had not been requested in a long time. Additionally, staff expressed an interest in knowing more about what data are reported to external organizations like ACRL or IPEDS. This feedback helped DCAT recognize that staff needed a clear sense of purpose for collecting data. They also wanted guidance on when it was appropriate to stop gathering data that was no longer relevant.

A second source of frustration was that staff felt uncertain about how to accurately record certain data. The most often-cited example was reference transactions. Since the libraries did not have an established procedure for recording reference transactions, each staff member entered the information slightly differently, making the entire dataset less reliable. Additionally, over time, the online form for capturing reference transactions became unwieldly with new questions, response categories, and tags added on an ad hoc basis. As a result, the form was tedious to complete and distinguishing between options was not always straightforward. After consulting with staff, DCAT decided that it would be beneficial to bring together all staff members who use the entry form to agree on a common definition for reference transactions and find ways to streamline the form.

Lastly, all staff reported that finding library data was an onerous process. The first barrier was knowing whether the information was even available. Since (at the time) the libraries did not have a data inventory to consult, it was not always clear whether the information was being captured by a staff member, system, or vendor. A second barrier was locating the data. Library data are stored in many different places, including library systems, vendor administrative portals, various folders on the libraries' shared drive, and even on individual staff computers. This made accessing the data difficult since it was not necessarily obvious where to look; it also raised concerns about the long-term availability of data, especially those stored on staff computers. From these conversations, DCAT learned that staff wanted a single storage solution—a data warehouse—where all library data could be housed and easily queried. Further, staff wanted frequently requested statistics, like gate count and circulation, to be compiled on a regular basis so that it would be easy to retrieve the information without having to run a report.

Outcomes

DCAT has used the findings of this self-study to improve the libraries' data collection practices in numerous ways. One immediate application was to use the data inventory spreadsheet to convey information about the purpose of each library metric. The data inventory spreadsheet contains a column for *purpose* that indicates how the data are used. For each metric, one or more of the following categories are selected: (1) external reporting, (2) decision making, (3) demonstrating impact, (4) internal record keeping, or (5) quality assurance. An additional column, *reporting*, was created to list the names of external organizations the aggregated data are reported to (e.g., ACRL, Petersons, Wintergreen Orchard House). Including this information on the data inventory spreadsheet has provided staff with

a clearer sense of how data are used in the libraries. In some cases, it has also helped staff decide to stop collecting data that was no longer relevant.

Another outcome was improving how reference transactions are captured. As discussed earlier, many staff felt uncertain about how to correctly record this information. To address this concern, DCAT hosted a half-day meeting in which staff agreed on a standard process for collecting reference transactions. The first part of the meeting focused on adopting a common definition for reference transactions using the ANSI/NISO Z39.7-2013 Data Dictionary. Next, we discussed ways to streamline the online form to include only information that was needed for external reporting or internal decision making and quality assurance purposes. In the end, a number of response categories were eliminated, which greatly simplified the form. These changes were put into effect on July 1, 2016 to coincide with the start of the FY17 data collection cycle.

Near the completion of the self-study, the libraries began implementing a new integrated library system (ILS). Alma, the new ILS, has sophisticated built-in data gathering and analysis tools that have enabled more automated collection of data that has been difficult to capture in the past. For example, Alma is able to harvest e-resource usage statistics using the SUSHI protocol. Staff have also automated data collection for book repair statistics and in-house use of print serials in Alma. The move to a next-generation ILS has enabled the libraries to move closer to our goal of a central data warehouse; however, a large amount of data still resides outside of Alma. While it is unlikely that we will implement a solution that pulls together data from all of our disparate data sources like Google Analytics, Gimlet, Springshare, and others, DCAT is exploring how to best organize data files that are not housed within a system, but are stored as individual Excel files.

Throughout the process of conducting staff interviews, DCAT members were able to observe how staff collect data. This led to opportunities to propose new methods for gathering data that improved quality and reduced staff time. One example is public computer usage statistics. Prior to the project, reference librarians counted the

number of individuals using public computers each hour and recorded the information in an Excel spreadsheet. Instead, DCAT recommended the libraries use LabStats, a commercial software product licensed by the IT department that tracks computer lab usage. The transition to LabStats has not only freed up staff time, but has also provided more detailed information about how our public computers are used. Finally, during staff interviews, DCAT members became aware that many staff were unfamiliar with how to run reports in library systems to retrieve information they needed. As a result, DCAT has started to offer a consultation service in which staff can meet with the group for advice on optimizing their data collection methods. DCAT has also begun provided training sessions on how to create reports in various library systems.

Conclusion
Completing a self-study of the libraries data collection practices was a beneficial process. In particular, the data inventory spreadsheet provided the libraries with a better understanding of the data staff routinely collect, where the data are stored, and how they are used. Responses to the second part of the questionnaire highlighted common challenges staff experience when collecting and using library data. The findings of this project have been used to improve data collection practices in numerous ways, including automating more of our data collection efforts, organizing where data are stored, and providing staff with training on how to retrieve data from library systems. Consequently, the libraries are now better positioned to use data to make more informed decisions about library resources and services, and demonstrate our value to the wider UMW community.

Endnotes
1. NISO, "ANSI/NISO Z39.7-2013, Information Services and Use: Metrics & Statistics for Libraries and Information Providers—Data Dictionary," *National Information Standards Organization,* 2013, http://z39-7.niso.org.

2. Project COUNTER, "The COUNTER Code of Practice," *Project COUNTER*, 2014, https://www.projectcounter.org/code-of-practice-sections/general-information.

Appendix

Staff Semi-Structured Interview

1. **Can you provide us with an overview of the types of data or usage statistics your service area collects?** Take notes on each item and follow up to find out:

   a. **Description**—what information is being collected? (e.g., library website traffic, ILL requests)

   b. **Collection Method**—how is the information collected? We want to know if the information is manually recorded (entered into a spreadsheet, document, online form, etc.) or automatically captured by a system like Virtua, ILLiad, Springshare, or Google Analytics.

   c. **Storage/Access**—details about where the information is stored or accessible. Does the information reside on a server that can be queried as needed, or is it stored as a spreadsheet, document, or other format? Where are the paper or electronic files located?

   d. **Frequency**—how often is this information captured or compiled? Examples include daily, monthly, quarterly, annually, or occasionally.

   e. **Outcome**—what is the importance of this information and how is it used? For example, usage data may be used to support resource renewal or cancellation decisions. Also, some data may need to be reported to external organizations like ACRL.

2. **What are some of the challenges you've experienced with collecting data or being able to use data to make decisions or demonstrate impact?**

3. **Are there other kinds of data or statistical information you wish we collected or had access to?**

4. **What data do you think we could stop collecting? Why?**

5. **Do you have any suggestions or comments you'd like to share with the Data Gathering Group?**