

Recommendations for the Design of the Library Analytics Toolkit

Introduction

Academic Library statistics has been a topic of study and concern within the profession for many decades.¹ While there seems to be widespread agreement that statistics can be useful in assessing and evaluating library resources and activities, there is little agreement over the exact metrics that should be used and the best practices for collecting them. Even when libraries attempt to collect the same data, subtle differences in terminology and interpretation can make it difficult to make meaningful comparisons between libraries. Moreover, it is not clear that all libraries would benefit from the same statistics, or could collect the information necessary for these statistics in a cost-effective manner.

This lack of consensus is evident at the Harvard University libraries, where approaches to data collection vary greatly among libraries, and in some cases, even between departments of the same library. While most of the librarians I spoke with agree that statistics can provide valuable information, current practices in data collection vary widely among the departments. To be successful, the Library Analytics Toolkit dashboard must be designed to (1) address numerous potential uses of library statistics, (2) ensure data consistency through clear and concise definitions and guidelines, (3) pull data directly from the current data collection and storage tools, (4) be extensible, and (5) provide support for both qualitative and quantitative data.

By focusing on these priorities, the Library Analytics Toolkit will be able to streamline the collection and use of current statistics while still responding to future changes to the ways that data is collected and statistics are used. Moreover, this design will make the dashboard a powerful tool for the future assessment needs of the Harvard University libraries.

¹ See, for example, Radford, Neil A., “The Problems of Academic Library Statistics,” *The Library Quarterly*, Vol. 38, No. 3 (July 1968), pp. 231-248, Stable URL: <http://www.jstor.org/stable/4305902> and Stubbs, Kendon, “Apples and Oranges and ARL Statistics,” *The Journal of Academic Librarianship*, Vol. 14, No. 4 (September 1988), pp. 231-235.

Current Practices in Collecting Statistics

The various libraries, archives and special collections at Harvard employ a wide range of statistics and metrics in tracking activities and transactions. One fairly universal tool used across the Harvard libraries is ALEPH, which tracks a range of statistics related to technical services and circulation activities. While ALEPH does provide access to reliable data on these activities, it can be difficult to access this data since doing so requires running a COGNOS report, an activity which can only be done by certain authorized individuals and which requires training. To supplement the information available through ALEPH and COGNOS, many departments count other transactions and patron interactions by hand. This includes both tracking space usage by counting the number of patrons in specific locations over the course of the day using tick marks on a sheet of paper and tracking transaction types and duration using check marks on a form designed to reflect the categories selected for tracking. This data is usually later input into another larger database or spreadsheet, but this second stage is often done at a later date and by a different staff member, which can introduce errors and lead to delays. Typically, this information is ultimately collected in an excel spreadsheet, which facilitates later analysis and manipulation, but some departments also use databases or even Word documents to track this information.

Recently, many libraries have begun to evaluate new software that can streamline data collection processes. Some of these tools provide other functionality, such as the Aeon system that has recently been adopted by several Harvard collections to track requests² and Request Tracker, an open source issue tracking system used by several archives and special collections at Harvard.³ Both of these systems are used primarily by archives and special collections at Harvard. Aeon primarily tracks requests for items whereas Request Tracker is more focused on reference requests and transactions, but both can be used to track at least some statistics.

² The Aeon system is currently being used to track requests for special collection items at Houghton Library, the Harvard Map Collection, Tozzer Library, the Loeb Music Library, the Fine Arts Library and Harvard-Yenching Library. More information on the system, including screen shots is available at http://hcl.harvard.edu/info/special_collections/.

³ More information about Request Tracker is available at <http://bestpractical.com/>.

Another newly adopted system that has potential applications for a wide range of library statistics is Springshare's LibAnswers system.⁴ Both Baker Library and the Harvard Law School Library have already started to use the analytics functionality of the LibAnswers system to track reference transactions and certain other library events, such as circulation transactions, but the other libraries that are using the system have not yet started using the analytics functionality. While not all Harvard libraries are currently using the LibAnswers system, more libraries may implement the system once Springshare has added some additional functionality. Also, other libraries that have implemented LibAnswers are considering adopting the analytics functionality in the future. Springshare has recently launched a new feature for LibAnswers, which allows libraries to track statistics for different departments in separate "instances" of the LibAnswers analytics system to increase the types of statistics that are tracked with LibAnswers, which may increase its use for data tracking purposes. While this system is not expected to incorporate all of the library statistics that are kept at Harvard, this system or a similar service is likely to become more prevalent in the future. The data collected in the system also can be exported into an excel file for easier access, manipulation and sharing.

Another major component of the statistics collected across the Harvard libraries is statistics that come from electronic sources and outside vendors. Examples of these sorts of statistics include both data about electronic resource usage and data collected through entrance gates and turnstiles. These statistics can be a source of frustration since they are often the statistics that librarians are most interested in but at the same time librarians have very little control over the format and granularity of these statistics. Many librarians across a range of libraries and collections stated that they would like to have more granular information about who was using the library resources (for example, what department patrons were affiliated with and each patron's status at the University) and more specific information about usage patterns within electronic databases. Currently, however, this information is not easily available. This is a

⁴ More information about LibAnswers is available on Springshare's website at <http://www.springshare.com/libanswers/> and information about the Harvard implementation of LibAnswers, including information about how Baker Library and the Harvard Law School Library use the analytics component, is available on the Virtual Reference User Group iSite at <http://isites.harvard.edu/icb/icb.do?keyword=k71132&pageid=icb.page343393>.

problem that has been studied fairly extensively both at Harvard⁵ and elsewhere,⁶ but no fully satisfactory solution is available given the current technology and the inconsistencies in data collection and presentation between vendors. Libraries will likely continue to negotiate with vendors for better information in the future, but it is unlikely that the libraries will ever have complete control over the format or types of data collected by third party vendors.

A significant source of variability in data collection among the Harvard libraries is the frequency with which data is both collected and compiled. While some libraries already collect data on an ongoing basis and create frequent reports, as often as biweekly or monthly, other libraries only generate statistical reports on an annual basis for inclusion in the annual report or the ARL statistics. Additionally, some libraries choose to collect data only during isolated periods rather than collecting it on an ongoing basis throughout the year. This can mean either collecting data for one week out of every month or collecting data for two or three weeks out of the year. This data is then used to extrapolate averages or trends for the entire year.

The current state of data collection at the Harvard University libraries, archives and special collections reflects the wide range of approaches that can be taken to evaluating and assessing library services and programs. While this can be seen as a complication to the development of a dashboard that will be able to accommodate all of these types of data, the vast array of approaches seen at Harvard does reflect the diversity of data collection methods that libraries use. If the dashboard is designed to incorporate the different types of data collected around Harvard, it will be well situated for adoption by other libraries with minimal modifications.

⁵ For more information about the current state of electronic resource usage statistics at Harvard, see Abigail Bordeaux, Sharon Boland Bayer, Wendy Gogel, Laura Morse and Noelle Ryan's Usage Statistics Analysis, which was submitted to the University Library Council and the Collections and Content Coordinating Committee on February 23, 2010.

⁶ See, for example, Gayle Baker and Eleanor J. Read, "Vendor-Supplied Usage Data for Electronic Resources: A Survey of Academic Libraries," *Learned Publishing*, Vol. 21, No. 1, pp. 48-57 (January 2008) and Abigail Bordeaux, Alfred B. Kraemer, and Paula Sullenger, "Making the Most of Your Usage Statistics," *Serials Librarian*, Vol. 48, No. 3, pp. 295-299 (2005).

Goals for the Toolkit

After meeting with individuals in a wide range of job functions throughout the Harvard libraries and archives, it is clear that there is enthusiasm for the library analytics toolkit. The idea of a dashboard of statistics that would give users a clear picture of library statistics is one that many found very helpful. In fact, some libraries, such as Baker Library at the Harvard Business School, have already developed internal “dashboards” that display select library statistics and OIS has developed a website to collect all the “statistical reports or other use metrics for systems supported by OIS.”⁷ The toolkit will be able to build on these existing tools by incorporating the statistics that are currently in use and providing a place for additional types of data to be displayed. To be able to fulfill the current and future needs of a wide range of users, the dashboard should focus on three main goals.

- 1) It should allow users to identify and visualize trends

Statistics alone are of limited value. Their true value comes from an ability to compare statistics over time to find trends in resources and service usage. One potential advantage of developing a statistics dashboard is that it can help librarians to identify these trends by tracking data over time and visualizing changes. Some existing library dashboards show how this sort of trend information can be incorporated into a dashboard’s display, such as the dashboard used by the Watson Library at the Metropolitan Museum of Art, which incorporates past months’ data into its display,⁸ or the New York Public Library Executive Dashboard, which was developed by the Office of Management and Budget and shows graphs that show changes over a ten year period.⁹ Another example of this ongoing data display method has also been developed by the Indianapolis Museum of Art.¹⁰ This sort of display gives users the flexibility to focus on current numbers or to track change over time to predict future activity, cost or transaction levels.

⁷ Available online at <http://hul.harvard.edu/ois/reporting/>.

⁸ The Watson Library dashboard is available online at <http://libmma.org/dashboard/>.

⁹ The Executive Dashboard is available online at <http://www.cfcrew.com/omb/expense.cfm?type=New%20York%20Public%20Library>.

¹⁰ The Indianapolis Museum of Art dashboard is available online at <http://dashboard.imamuseum.org/>.

2) It should give users an up-to-date and dynamic view of data

One issue with current data collection methods is that information is often collected by a single individual and stored off-line. This means that most staff members only have access to statistics when regularly scheduled reports are released, which can be as infrequently as annually. While the dashboard will be limited by the frequency with which data is collected and input, it could provide a more dynamic display for data that is collected and input frequently, including data that is automatically collected through ALEPH. This data could be set up to update automatically to provide wider access to data between reports. In addition, once such a feature is implemented, it would likely increase the frequency with which data was input as data would be immediately useful in a way that is not currently possible when data is kept in off-line excel files or databases. This functionality would not only make library statistics more useful but would also be a major factor that could encourage adoption of the dashboard across a range of libraries.

3) It should facilitate the sharing of information

A major advantage of displaying data in a dashboard is the ease with which this information can be shared. This can include sharing information between departments within a single library, between libraries or even with patrons and the public. As seen with the existing dashboards highlighted above, data can be made publicly available to engage with members of the public, whether this means patrons or professionals at other similarly situated libraries. The dashboard should include features that make it easy for libraries to make statistics public or to otherwise share their data and statistics even if the libraries choose not to make such data publicly available initially.

Recommendations

While these three main goals provide a starting point for the dashboard, there are many approaches that can be taken, all of which likely satisfy these three basic needs. However, based on the research conducted over the past several months, I would recommend designing the initial toolkit with the following features in mind.

1) High-Level Statistics with Add-On Module Options

To facilitate use by a wide variety of different libraries, the toolkit should be designed to accommodate both high level statistics that will be relevant to a wide variety of libraries, such as number of items held, total expenditures across various categories and number of patron interactions such as reference or circulation transactions, and more granular statistics that may only be applicable to certain types of libraries and collections. Users should be able to configure their individual dashboard to display the types of statistics that are most useful for their work, whether high-level statistics that allow comparisons between different libraries or more granular statistics for use in monitoring workflow in specific departments. In this manner, the dashboard can be configured for users who are focused on tracking changes in the types and number of activities in an individual department and for users who want an overview of basic statistics for multiple departments. Moreover, by including high-level statistics that represent commonalities between libraries, the dashboard can facilitate comparisons between similarly situated libraries.

This level of configuration will also increase the use of the dashboard since it will make it easier to integrate the dashboard into existing workflows. Currently, there is a wide degree of variance between libraries in terms of the level of statistics that are collected, and even variance in the types of statistics to which the libraries have access. Some of these differences are a matter of policy, such as libraries that have decided not to collect detailed information about patron activities for privacy reasons, and some are based on technological differences, such as the difference between turnstiles that require users to swipe their IDs and counters that only track the number of patrons with no further information. Accommodating both granular and high-level statistics will respond to these different collection methods. In addition, to further address different types of data collection methods while still allowing comparisons between departments using different methods, the modules can be designed so that more granular levels of data can be combined together within the higher level modules. For example, many reference departments track the amount of time spent on a reference request. Most do not record the exact amount of time spent on a request, but instead use time ranges (such as 1 to 10 minutes, 11 to 20 minutes, etc.). However, different departments may define the exact time range differently. By taking advantage of the open source nature of the toolkit to allow libraries to create modules that reflect their exact ranges but also including higher level module that includes all reference transactions

below 30 minutes and above 30 minutes, the dashboard will permit users to track local transactions in the same way that they are currently being tracked while simultaneously allowing users to aggregate or compare data across departments. This will facilitate use of the dashboard and will also encourage libraries to contribute to its ongoing development.

The question of what the high-level statistics should look like, however, remains difficult. While several people I spoke to agree that ARL statistics were of limited use because they only collect very high-level information about certain library functions,¹¹ I would nevertheless recommend using them as the basis for the highest level modules for a few reasons. First, all of the Harvard libraries have experience collecting ARL statistics. In fact, the ARL statistics were some of the only statistics that people across the libraries, archives and special collections at Harvard referred to consistently. Even departments that otherwise took very different approaches to statistics would refer to ARL statistics and had experience using their collection methods to fulfill the requirements set forth by ARL. This seems to suggest that the dashboard could easily be used on day one if it included modules that matched the ARL requirements. Second, other members of ARL across the United States and Canada will, by necessity,¹² also be collecting these statistics, which will improve the odds of other libraries adopting the dashboard. Widespread adoption of the toolkit would be beneficial to its goals since it will not only benefit the greatest number of libraries but will also improve discussions about library metrics by simplifying comparison between libraries and putting everyone on the same page with respect to metrics. While ARL statistics should by no means be the sole functionality of the dashboard, basing the high-level modules on ARL statistics would spur adoption across a wide range of libraries and encourage comparisons across libraries by offering a uniform system of statistics with which users are already familiar.

¹¹ The problems with ARL statistics also been highlighted in the literature. As Hendrix notes, for example, “[t]raditional ARL measures, such as volume counts and total expenditures, have severe limitations qualitatively, as they only measure size and temporal growth. Assessing the effectiveness and quality of library services through traditional ARL statistics is not possible.” Hendrix, Dean, “Relationships between Association of Research Library (ARL) Statistics and Bibliometric Indicators: A Principal Components Analysis,” *College & Research Libraries*, Vol. 71, No. 1, p. 39 (January 2010).

¹² Members of ARL are required to “contribute the data necessary to establish the membership indices and to compile the annual ARL Statistics.” Procedures for Membership in the Association of Research Libraries (October 18, 2001), available at: <http://www.arl.org/arl/membership/qualproc.shtml#pIII>

2) Include Clear Definitions and Guidelines

Once modules and statistics categories have been set, it is very important that the system include clear definitions of each category and guidelines for the way that the data should be input and the types of information that should be included. Failure to provide this information in a detailed and easy-to-understand manner can undermine the usefulness of the data and will make meaningful comparisons impossible. As Hendrix noted, “Libraries may define the same measures differently, thus reporting inconsistent numbers” which can impact currently collected data such as the ARL statistics.¹³ By focusing on providing clear definitions and guidelines, it will be possible to eliminate this problem in advance for these statistics. As an added benefit, this level of attention to the specifics of the statistics being tracked will also help to facilitate the design process. To ensure that this level of clarity extends to all of the statistics included in the dashboard, users who develop new modules to reflect additional library statistics should also be encouraged, or even required, to including definitions and guidelines when sharing these modules.

3) Draw from Current Collection Methods to Minimize Data Entry Requirements

A recurring concern expressed by many librarians is that using the dashboard would add an additional step at a time when librarians are already strapped for time. While librarians are enthusiastic about the idea of a statistics dashboard, adoption is likely to be limited if the dashboard will require a significant amount of additional work to input data or a high degree of data re-entry. Wherever possible, the toolkit should be designed to pull data from existing data sources either automatically or with one or two clicks.

Given the wide variety of data collection and storage tools that are used, it will not be possible to design a system that can automatically draw data from every possible source. However, certain file types are more common than others which means that it will be possible to captures a high percentage of the data by focusing on a limited number of file types. In particular, many different software tools and vendors offer the option to export data into an excel file, which can then be imported into the dashboard. Many libraries that collect data by hand

¹³ Hendrix, Dean, “Relationships between Association of Research Library (ARL) Statistics and Bibliometric Indicators: A Principal Components Analysis,” *College & Research Libraries*, Vol. 71, No. 1, p. 39 (January 2010).

also store this data in excel files. Another source of a large amount of automatically updating library data is ALEPH and the related COGNOS reports. By designing the toolkit to automatically pull data from these two sources, or at a minimum, to allow data to be easily uploaded from these sources, it will be easier for users to integrate the toolkit into the existing workflow and data collection processes. A significant amount of data is also available in XML, making this a good candidate for another data input mechanism. By focusing on these three types of data files, users will be able to upload a majority of the data currently collected without having to re-enter data into the dashboard.

4) Designed to be Extensible

While the open source nature of the library analytics toolkit will facilitate future changes to the dashboard by users, one of the most important qualities that the toolkit can have is an ability to be expanded and extended in the future. This is true for several reasons. First, given the wide range of statistics outlined above, it would be impossible to design the initial toolkit in a way that could accommodate all of the existing types of statistics being collected. Rather than waiting to release the dashboard until it has modules for every possible type of data, it will ultimately be more practical to provide the dashboard to users with modules for the most common types of statistics while making it straightforward to add additional modules in the future. Second, many librarians noted that they believe that methods of data collection will change in the coming months or years, particularly as a result of the library transition. Even if the dashboard can be designed to incorporate most of the types of data that are currently being used, this could change in the future, making it important that the dashboard has the flexibility to incorporate new statistics in the future. Third, as noted above, libraries have limited control over the types of data they can extract from certain sources such as electronic databases. To be successful in the future, the dashboard needs to be easily modifiable if the types of statistics provided by third party vendors change over time. Finally, our research has primarily been focused on the statistics collected by libraries at Harvard University. However, in the future we hope that other libraries will also use the dashboard, which means that it is important to design the dashboard so that it can be modified by other libraries to meet their specific needs. It is anticipated that the modular design of the dashboard will make future additions and modifications fairly straightforward. The model for this approach should be the Brown

University Dashboard which consist of discrete “widgets” and includes “[a]n administrative form [that] will allow any library staff-member to create a widget by filling out basic information, pasting in data-values in a specified format.”¹⁴

Assuming that the dashboard will be extensible in the future and that it will not attempt to incorporate all statistics in its initial release, the question next becomes what statistics should be included in the initial dashboard. I would recommend focusing on those statistics that are most uniform and most easily captured across a range of Harvard libraries. This means an initial focus on statistics that can be drawn from ALEPH and from the statistics on electronic resources that are provided by OIS. This will include circulation and usage data, which was the initial focus of the Library Analytics Toolkit proposal. Additional types of data that may be most easily incorporated into the toolkit include information on reference transactions, which are generally tracked at most libraries, albeit in differing formats, and gate count statistics.

5) Designed to Incorporate Both Qualitative and Quantitative Data

While the traditional focus of library statistics is on quantitative data, several libraries at Harvard collect qualitative data as well, including focus group data and survey responses. This data is likely to become even more prevalent in the future as libraries are beginning to collect more qualitative data to “[add] extra dimension and depth to quantitative data.”¹⁵ While it may be difficult to fully integrate these types of data into the dashboard layout of the toolkit, several individuals expressed an interest in being able to store and manipulate this data in the same application that stored their library’s quantitative data. To this end, it would be helpful if the toolkit could be designed to either provide a place for qualitative data, or at a minimum with an eye towards allowing this sort of expansion in the future.

6) Build on Already Existing Dashboards

A final recommendation for the toolkit would be that it take advantage of the work that has already been done in this area at other libraries. As discussed above, several libraries have designed software applications that can be used to track and display library data through an

¹⁴ <http://library.brown.edu/dashboard/info/>.

¹⁵ Cathie Jilovsky, "Library Statistics: reflecting yesterday, today and tomorrow," presented at the 6th Northumbria Conference on Performance Measures, August 2005.

online dashboard. As part of the research that has already been conducted, we have been in contact with the team at NCSU that has been working on that University's Data Views project.¹⁶ This team hopes to offer their dashboard and its open source code in the coming months. Additionally, the code for the Brown University Library dashboard is already available online.¹⁷ While the Library Analytics Toolkit should be designed to meet the specific needs of the Harvard libraries, we should also be open to learning from the already existing tools in this space and should seek to collaborate with these other teams when possible.

¹⁶ More information on this project is available at <http://www.lib.ncsu.edu/dli/projects/dataviews/>.

¹⁷ The code is available at <https://bitbucket.org/bul/dashboard-app/wiki/Home>.