Smart Data, Smart Library: Assessing Implied Value through Big Data

Jin Xiu Guo
Stony Brook University, USA

Gordon Xu
Northern Michigan University, USA

Abstract

The growing expenditure on electronic resources has become a new norm for academic libraries. It is crucial for library administration to measure the impact of such investment consistently and persistently, and then develop collection strategies. Big data technology provides such an arena for management to gain insights through meaningful data and allow libraries to optimize collection operations in real time. The purpose of this study is to assess the implied value of a research library by analyzing Cost per Use with BigQuery—a cloud-based data warehouse. The authors developed a systematic approach to process structured data including e-resource usage and interlibrary loan transactions, and then analyzed the data in BigQuery. Google Data Studio was employed to visualize the results. The findings of this study not only manifest the implied and exchange values of the research library but also offer an innovative approach to predict the future collection needs. The methodology employed in the study also provides a new opportunity for libraries to adopt big data technology and artificial intelligence to tackle intricate problems and make smart and informed decisions in this big data era.

Introduction

Academic libraries have been working tirelessly to sustain library collections to meet the needs of teaching and research. However, the exponential cost increase of electronic resources has surpassed what the library budget can afford. Meanwhile, the open-access movement has made more and more scholarly publications freely available to the public. It is impossible to develop an effective collection strategy without assessing the values that library collections have brought to users.

The impact of electronic resources on teaching and research has changed the landscape of collection development. For example, 89% of the overall use of library collections at the University of Massachusetts Amherst happens outside of the library buildings and 53% of the use is to support teaching or class work. In STEM disciplines, 92% of use by graduate engineering students also occurs outside of the libraries and 45% of the use is for their theses or dissertations.[1] Academic libraries have met these requirements by subscribing to electronic resources and creating digital collections over the last decade. Researchers and students have been enjoying convenient access to these rich and diverse resources. However, with the growth of digital resources, subscriptions to online resources have become the primary consumption of the collection budget. It has never been so crucial for libraries to examine their collection strategies critically and seek solutions to this challenge which libraries face today and tomorrow.

Measurement

It has been 16 years since the COUNTER initiative was launched. The COUNTER Code of Practice provides a mechanism for libraries and publishers to gather usage statistics consistently across publishers and libraries. One of the metrics listed in the code is *Journal Report 1* (JR1) which is the number of successful full-text articles requested by month and journal title.[2] Therefore, it is possible to calculate the *Cost per Use* (CPU) with JR1 and journal cost. CPU is a widely accepted criterion to assess electronic journal subscriptions, but is by no means the only metric for libraries to adopt. Every library is different and unique. Libraries must conduct assessment and interpret findings in their respective contexts.

It is complicated to measure the value that an academic library brings to the university or college. Student success might be the results of contributions from many campus constituents. Studies have been conducted to assess library values with appropriate measure indicators, as Tenopir pointed out that libraries could measure implied value with usage statistics.[3] However, libraries must consider the cost when assessing the value. CPU potentially integrates the use value into the exchance value in the way of a ratio, which allows both elements considered in a calculated value. CPU is a relative value determined by both use value and exchange value.[4] Therefore, CPU can be a consistent means to measure the implied value of electronic resources as well as exchange value.

This study explores a systematical method to consolidate and analyze the big data including usage statistics, interlibrary loan transactions, and operational and institutional data in BigQuery to assess the library-implied value and provide evidence for management to make evidence-based decisions and develop collection strategies in its respective context.

Methodology

Libraries have adopted various tools to collect usage statistics systematically and run operations with integrated library systems (ILS). However, the communication between different information systems has not been directly established. To resolve this problem, the authors developed a process to gather, process, and analyze structured data in BigQuery.

BigQuery is a product of Google Cloud Platform,[5] which is a suite of cloud computing services. Except for a set of management tools, the Google Cloud Platform provides a series of modular cloud services including computing, data storage, data analytics and machine learning. BigQuery is a scalable and fully managed enterprise data warehouse for analytics. Currently the cloud platform is free through a registration service.

Data Source

The authors generated the *Journal Report 1* (JR1) for 2015, 2016, and 2017 individually with the EBSCO Usage Consolidation. The cost of each e-journal is critical to this project. To facilitate a large amount of data processing, the authors chose the *Collection Assessment Reports* in EBSCONet and the Journal-holding Report generated in EBSCO Holding Management. The acquisitions data in the integrated iibrary iystem (Aleph) supplements the cost data that is not available in the EBSCONet reports. The most requested journal reports for each calendar year from 2015 to 2017 were created with ILLiad reporting. The ILL expenditure was a part of the ILL operational data.

Data Cleanup and Preparation

Data cleanup is a crucial step before analyzing data. To select the data that is meaningful to the final results, the authors believe that at least 70% of data analysis involves cleaning and selecting the most useful data. It is worth pointing out that saving a copy of the original data may prevent researchers from losing data permanently. The first step is to identify whether an Excel workbook contains irrelevant or excessive data, such as plots, graphs, irrelevant headings, or explanation information in reports. Such data should also be removed. It is necessary to add the data that is missing. For example, if a journal has not been assigned one of the four subject categories—namely STEM, arts and humanities, health science, and social science—give it a category. If no cost information is included for a journal or journal package, add the data correspondingly. These data manipulations can be easily handled in Excel.

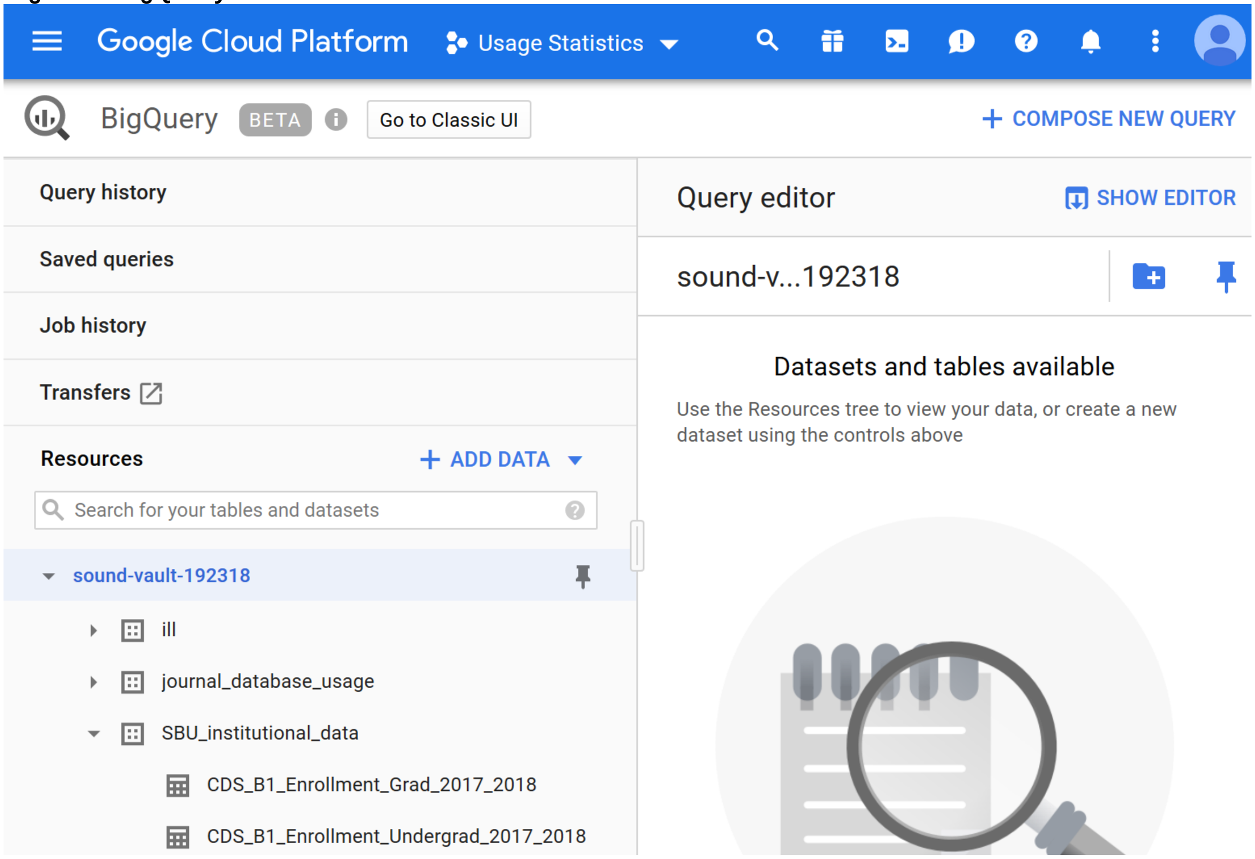
The second step is to convert Excel worksheets to CSV files. BigQuery only accepts some specific file formats, including CSV, JSON, Avro, Parquet, and ORC. At present, there is no way to upload an entire Excel workbook at a time. Data must be transmitted from a single worksheet. In this study, all data sources obtained are Excel workbooks with multi-worksheets. The converting process is simple. However, if Excel spreadsheets contain non-ASCII (American Standard Code for Information Interchange) symbols, such as foreign characters (tildes, accents, etc.) or hieroglyphs, a particular treatment should be taken.[6]

Import Data into BigQuery

A local data source can be loaded either via a BigQuery web UI or CLI (command-line interface). Below is the process of loading data into BigQuery with a web UI:

- Create a new project in BigQuery.
- Create a dataset in the new project. The authors created three datasets, namely *ILL*, *Journal_database_Usage*, and *SBU_Institutional_data* (see Figure 1) for this project.
- Create tables within a dataset.
- Import an Excel worksheet into a BigQuery table. In this study, the authors uploaded interlibrary loan transactions and journal usage statistics from 2015 to 2017 into the dataset *ILL* and *Journal_database_Usage* created in the previous step respectively, and then import Stony Brook University institutional data into the dataset *SBU_Institutional_data* as well.

Figure 1. BigQuery Datasets andTables



Datasets and tables available

Use the Resources tree to view your data, or create a new dataset using the controls above

Data Analysis

The authors analyzed data and sought the relationship among variables by operating SQL (Structured Query Language) queries in BigQuery. The BigQuery standard SQL complies with the 2011 SQL standard and has extensions that support querying nested and repeated data. By default, BigQuery runs interactive query jobs on demand, which means that the query is executed as soon as possible. Query results are always saved to either a temporary or a permanent table.

Data Visualization

Google Data Studio is a business intelligence tool used to visualize data through dashboards and reports. Besides Google analytics products, it also collaborates with Facebook, Amazon, YouTube, and more than 120 business partners to meet various needs. In consideration of adopting BigQuery for future studies, the authors decided to choose the Google Data Studio as the data visualization tool for this study. When a query is finished, it can be imported into Google Data Studio for data visualization.

Results

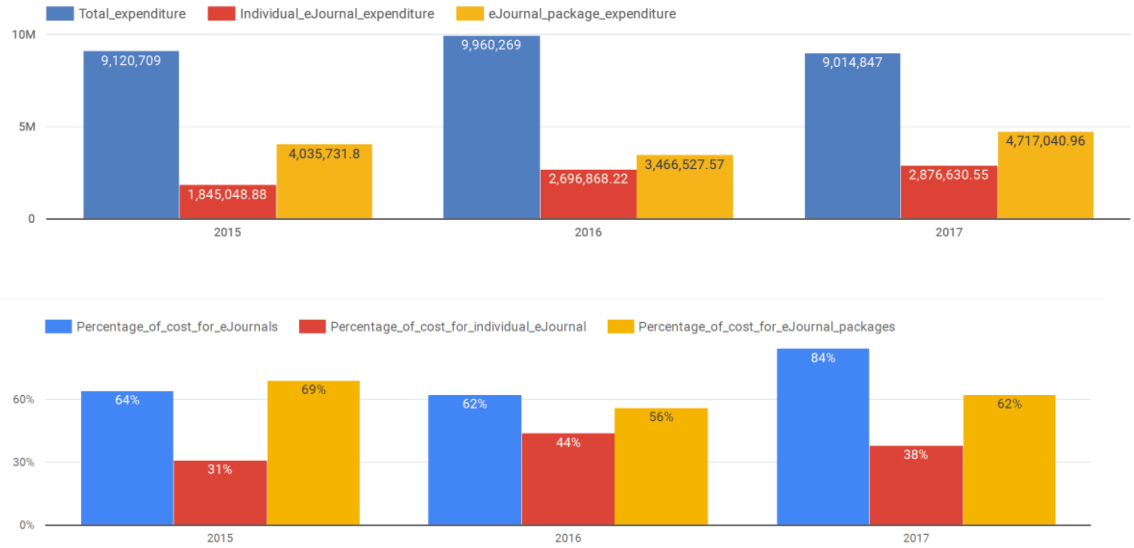
Library Collection Budget

The library collection budget is consistent with the resource expenditure. Figure 2 shows the change of the Stony Brook University (SBU) Library collection budget from 2015 to 2017. Compared to the budget for 2015, it stays flat for 2016, but slightly increases for 2017. The finding also signifies that most of the budget is spent on electronic journal subscriptions. For instance, 64% of the total budget was spent on e-journals in 2015 and 62% in 2016, but this expenditure jumped to 84% of the entire collection budget in 2017.

When examining the expenditure on e-packages, it shows that 69% of e-journal budget is for e-packages in 2015 and the number shifted to 56% in 2016 and 62% in 2017 respectively, which indicates a growing expenditure on e-packages over the last three years. Higher spending on e-packages in 2015 was due to one-time purchased backfiles.

Notably, more than half of the e-journals were subscribed to through e-packages or big deals. It is common for libraries to subscribe to e-journals via a Consortium Member License Agreement to bring down the cost of per journal title. The SBU expenditure for the fiscal period 2015–2017 is in accord with this practice.

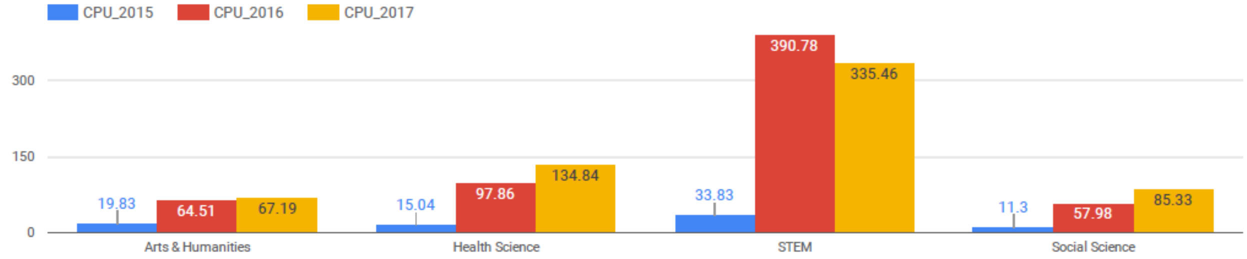
Figure 2. SBU Library Collection Budget



Average CPU for Individual Journal Subscriptions

The SBU Libraries tracks the usage statistics for primary electronic resources through EBSCO Usage Consolidation. The JR1 shows that 88% of e-journals were used at least once in 2015. The same ratio is 60% for 2016 and 95% for 2017 similarly. These e-journals are available to SBU users through licensed e-journals or full-text databases. In this study, the authors firstly calculated CPU for each journal by dividing the journal cost by the value of *Reporting Period Total Use* in JR1, where the journal cost is grouped by a subscription model, such as individual subscription, e-package, or database, then computed the average CPU for each subject category. Figure 3 is the CPU for individual journal subscriptions.

Figure 3. Average CPU for Individual Journal Subscriptions



The results show that the average CPU increases across STEM, health science, social science, and arts and humanities over the last three consecutive years. The average CPU for this period is $253 for STEM, $83 for health science, $52 for social science, and $51 for arts and humanities. Overall, the average CPU for individual subscriptions is about $110. Surprisingly, the average CPU for STEM in 2016 is about three times higher than it is for health science but decreased to 1.5 times in 2017. The possible factors include the increase of cost per title, less use, or a combination of both elements. It is worth mentioning that a study on periodical price also signifies that the cost growth in 2016 is more than the increase in 2017.[7]

However, the results also raise the concern on the effectiveness of the traditional journal subscription model. Would it be more cost-effective to purchase an article via pay-per-view (PPV) rather than a journal subscription? Especially in the STEM field, SBU might not consider a journal subscription until the number of PPV reaches a predefined limit. Libraries may implement the decision-making process with a prediction model. The model should consider the factors including discipline, labor cost, access convenience, and business transactions. On the other hand, publishers, vendors, and libraries could develop a new subscription model collectively to sustain the journal affordability.

Average CPU for e-Packages (CPUP)

To compare CPUs for two subscription models, the authors utilized the same method to calculate CPU for e-packages (CPUP) (see Figure 4).

Figure 4. Average CPU for e-Packages (CPUP)

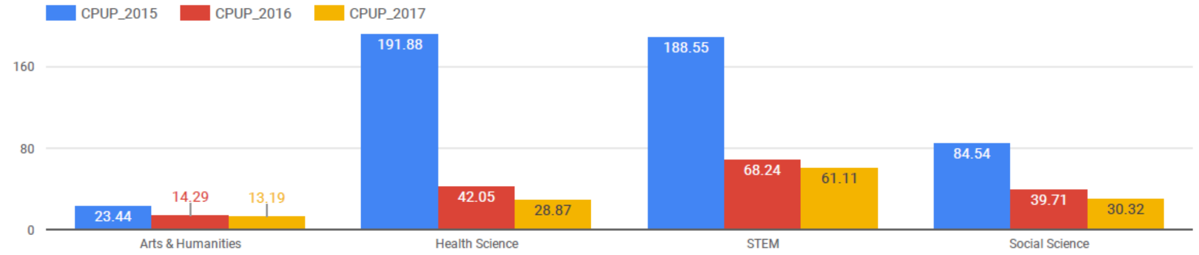


Figure 4 shows that CPUP dramatically decreased across four domains and is much lower than for individual journal subscriptions. The CPUP is $106 for STEM, $87.6 for health science, $51.5 for social science, and $17 for arts and humanities. Remarkably, the CPUP for STEM is only 42% of the CPU for individual journals.

Likewise, the high CPUP for 2015 across four subjects is expected as a result of one-time purchased back files. The findings demonstrate the e-package model is more cost-effective than the journal subscription model for SBU Libraries, but the annual increase rate for e-packages could affect its efficiency. The authors also suggest that libraries should closely monitor the usage to timely adjust journal titles selected in a respective e-package.

Cost Per View (CPV) for ILL

The SBU Libraries adopt the pay-per-view model to acquire articles that cannot be fulfilled via ILL or when the charge of an ILL article is higher than the PPV price. Dividing total cost by the number of pay-per-view articles, the authors calculated the average CPV for the period of 2015 to 2017 (see Figure 5).

Figure 5: Average CPV for ILL

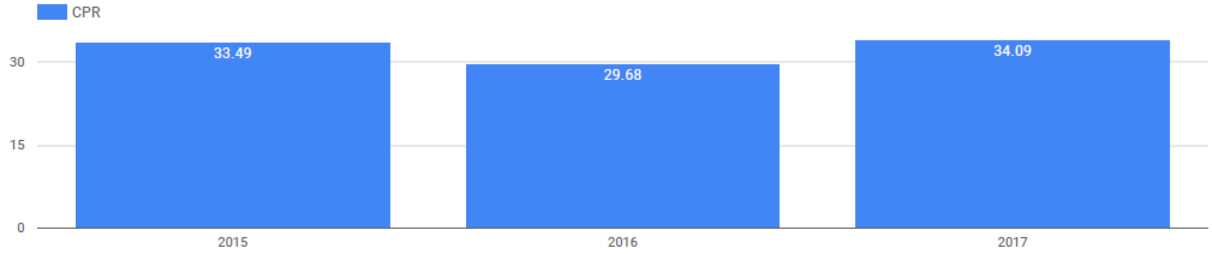
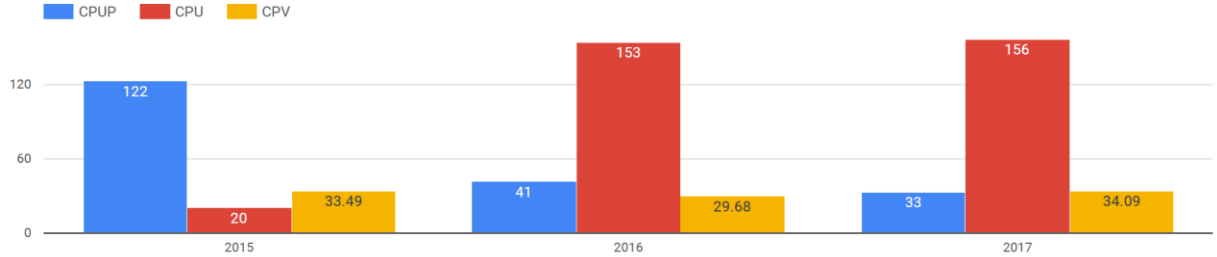


Figure 5 indicates that the average CPV is $32.42 for the last three years. If looking into the CPV for each year, the difference is less than $5 among 2015, 2016 and 2017.

To understand the relationship among CPU, CPUP and CPV, the authors compared CPU with the CPV in Figure 6.

Figure 6. Comparison of CPU, CPUP, and CPV



The average CPV is $33.49 for 2015, $29.68 for 2016 and $34.09 for 2017. While CPU for individual journals is $20 for 2015, $153 for 2016 and $156 for 2017, CPU for e-packages is $122 for 2015, $41 for 2016, and $33 for 2017. In consideration of one-time purchased back files in 2015, the PPV model is the least expensive, e-package is next, and the individual journal subscription model is most expensive for SBU Libraries. Therefore, the pay-per-view model is more effective than journal subscriptions and e-packages when the number of requested articles via ILL is manageable without increasing personnel.

Most Requested ILL Journals

To recognize the pattern of CPV articles, the authors tracked the journals that were requested more than once via ILL from 2015 to 2017 and grouped them by four disciplines as well.

Figure 7: Most Requested Journals via ILL

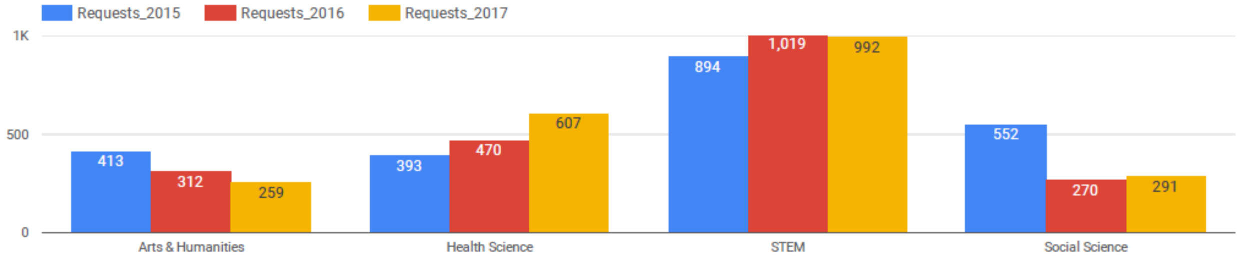


Figure 7 shows that STEM articles are highly requested and account for nearly half of ILL borrowing transactions. The total number of most requested journals is 2,095 for STEM and 1,470 for health science for the last three years. The numbers of most requested journals in both disciplines present a steady growth from 2015 to 2017. On the contrary, the numbers of most requested journals for arts and humanities and social science decrease, which means that the effort put into collection development in these disciplines is rewarding. It also suggests that the electronic resources for STEM and health science need to be improved. For example, the library should actively review journal titles selected in e-packages and add high quality open-access journals to the collection.

Conclusions

Academic libraries have supported universities and colleges to achieve their educational missions for decades. Today, it is even more important for libraries to articulate their values to gain more support to meet the ever-changing needs of users in this digital age. CPU incorporates exchange value into implied value and can be an essential metric of measuring library values and effectiveness in the respective context. The purpose of this study is to assess the implied value of library collections with CPU by utilizing big data technology.

The study shows that the cost of e-journal subscriptions has increased dramatically from 2015 to 2017. Particularly, the expenditure on e-journals reached 84% of the library collection budget in 2017. Also, the CPU for e-packages in the STEM field is about 42% of the CPU for journal subscriptions. The findings indicate that the e-package model is more cost-effective than a journal subscription, especially for STEM and art and humanities at SBU Libraries.

The PPV model can be a valuable addition to the e-package and journal subscription models, which allows libraries to provide resources beyond existing collections in a cost-effective manner. SBU Libraries has employed this model to acquire articles on STEM and health science for a few years.

Libraries may improve the journal subscription model by developing a prediction model to alert the Acquisition Department when to switch from a pay-per-view model over to a journal subscription by adopting big data technology and artificial intelligence. In the meantime, publishers, vendors, and libraries can develop a more sustainable PPV model collectively to meet the emerging and growing needs of scholars. The ability to add and maintain high-quality open-access content to library collections is also critical to library success.

Endnotes

[1] Rachel Lewellen and Terry Plum, “Assessment of E-Resource Usage at University of Massachusetts Amherst: A MINES for Libraries® Study using Tableau for Visualization and Analysis,” *Research Library Issues* no. 288 (Jan. 2016): 5–20.

[2] Oliver Pesch, “A Brief History of Counter and SUSHI: The Evolution of Interdependent Standards,” *Information Standards Quarterly* 27, no. 2 (summer, 2015): 5.

[3] Carol Tenopir, “Building Evidence of the Value and Impact of Library and Information Services: Methods, Metrics, and ROI,” *Evidence Based Library and Information Practice* 8, no. 2 (2013): 270–274, https://doi.org/10.18438/B8VP59.

[4] Matthew Harrington and Connie Stovall, “Contextualizing and Interpreting Cost per Use for Electronic Journals,” *Proceedings of the Charleston Library Conference* (2011): 1–8, http://dx.doi.org/10.5703/1288284314928.

[5] “Why Google Cloud Platform?” Google Cloud, Accessed January 4, 2019, https://cloud.google.com/.

[6] Svetlana Cheusheva, “How to Convert Excel to CSV and Export Excel Files to CSV UTF-8 Format,” last updated September 11, 2018, https://www.ablebits.com/office-addins-blog/2014/04/24/convert-excel-csv/.

[7] Stephen Bosch and Kittie Henderson, “New World, Same Model: The Shifts to Online and OA Continue Apace, but neither is Causing a Sea Change in Pricing,” *Library Journal* (April 19, 2017), https://www.libraryjournal.com/?detailStory=new-world-same-model-periodicals-price-survey-2017.