Cost per Use in Power BI using Alma Analytics and a Dash of Python

Authors

Lydia Harlan
Budget Analyst for Collections, Discovery, and Digital Strategy, University of Oregon Libraries
lharlan@uoregon.edu
0000-0001-6957-1942

Kristin Buxton
Head of Science Liaisons, University of Oregon Libraries
kbuxton@uoregon.edu
0000-0002-1693-7039

Gabriele Hayden
Head of Data Services, University of Oregon Libraries
ghayden@uoregon.edu
0000-0003-4740-4187

Abstract

A trio of personnel at University of Oregon Libraries explored options for automating a pathway to ingest, store, and visualize cost per use data for continuing resources. This paper presents a pipeline for using Alma, SUSHI, COUNTER5, Python, and Power BI to create a tool for data-driven decision making. By establishing this pipeline, we shift the time investment from manually harvesting usage statistics to interpreting the data and sharing it with stakeholders. The resulting visualizations and collected data will assist in making informed, collaborative decisions.

Introduction/ Background

At University of Oregon (UO) Libraries high level collections decisions are made by a collection managers group consisting of representatives from acquisitions, electronic resources, subject librarians, and access services. Facing a possible large cut to the collections budget in Fiscal Year 2022, this group started talking about how to decide which serials to cut. Cost Per Use (CPU) was raised as one starting point for making data driven decisions.

Historically at UO, usage data has been stored in manually created spreadsheets with data compiled from all vendors, while cost data has been stored in Alma. When small package cancellations were on the table in previous years, CPU was hand generated for those titles. The employees who manipulated the spreadsheets were understandably reluctant to hand generate CPU for the over 78,000 rows of usage numbers for combined vendor statistics at that time.

Two of the authors, members of the collection managers group, began exploring alternative options. We knew that there were commercial products that would provide this data but didn't have buy-in from the necessary stakeholders to purchase a tool. Instead, we pulled in a third colleague with data experience to help create our own tool. We wanted to provide CPU for as much of the continuing resources

collection as we could in ways that would be easily understandable for the subject librarians who would need to use the data.

Literature Review

As Harker (2022, 355) has observed, the cost per use metric "has become ubiquitous and nearly universal for evaluating resources, especially renewable resources such as journals and databases," because it is straightforward and easy to grasp. Most articles discussing cost per use address controversies regarding whether cost per use should be used, the specific metrics and processes that should be included in a cost per use analysis, or limitations to the specific metrics (such as COUNTER 5) currently in use.

Less common are articles describing the technical details of reproducible pipelines for gathering and deploying cost per use analyses. This may reflect the challenges of using proprietary software (such as Power BI) to work with data protected by license agreements. The few articles that did address technical access questions were very helpful to our project. Pesch (2015) offers a basic primer for using SUSHI and COUNTER, while Boardman and Thompson (2022) describe using Alma Analytics to programmatically query SUSHI APIs and visualize Counter statistics, documenting that it could be done even if they did not share the technical details needed to execute the project. This paper aims to fill a gap in the scholarship by offering a clear, step-by-step description of what we did that may be of use to others.

Approaches to the Pipeline

In the first iteration of this project, we created a CPU table in Excel by merging spreadsheets of usage and cost data. (Figure 1)

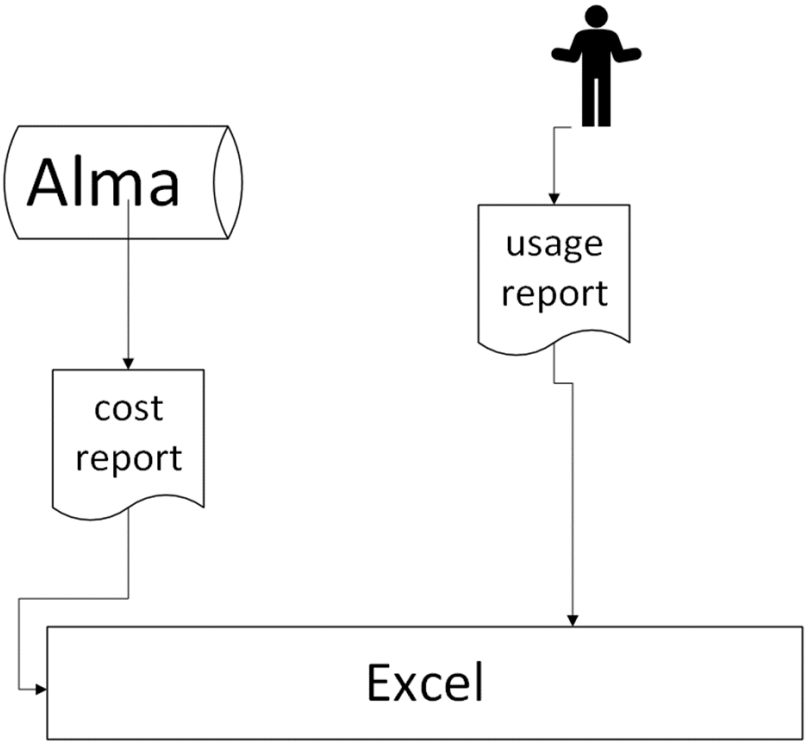

Figure 1

We performed lookups on title and ISSN for journals where we had both usage numbers and payment history. We were able to calculate the CPU for many of our journals and created graphs to make the data more visual. We ended up not needing to make major cuts to our journals budget at that time, so this tool wasn't needed immediately. We decided there was still value in improving our report. This Excel-only version had two major limitations: 1) it was challenging to update with new usage numbers and 2) sharing the underlying usage along with the visual meant sharing a very large Excel file vulnerable to user error, easily unintentionally manipulated to show false results.

We next decided to import the Excel files into Microsoft Power BI. (Figure 2)

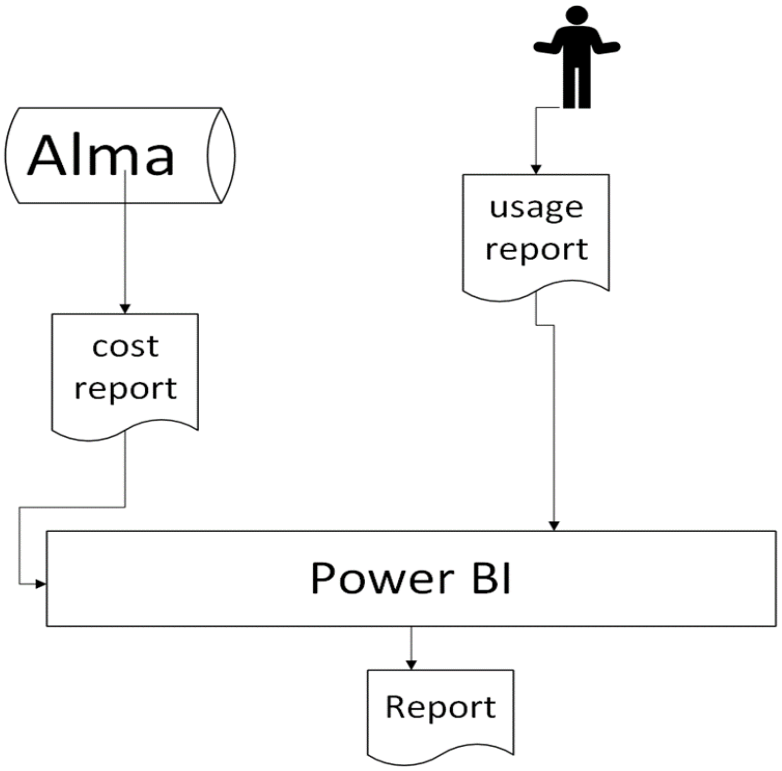

Figure 2

Power BI allows easier update of datasets, provides additional visualization options, and offers a way to share visuals and tables of data without needing to share the underlying data files. We successfully translated the data transformations we had conducted in Excel into Power BI. However, this version still relied on the laboriously hand-created Excel usage data file.

The next step was to gather usage statistics dynamically by querying the SUSHI APIs of our vendors. We experimented with making SUSHI calls directly from within Power BI and were successful in getting the process to work with a static table, but we couldn't find a way to continually add the next month's usage in a way that would store the data properly. (Figure 3)

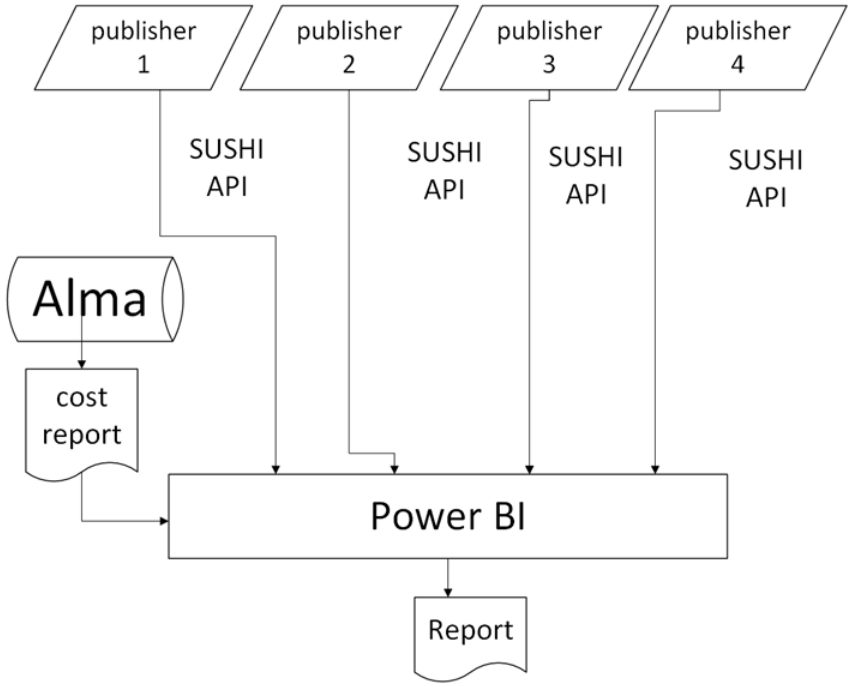

Figure 3

We learned that this is a deliberate limitation of the Power BI software, which is not designed to function as a data source but instead to draw on more stable data sources such as SQL databases. We would have needed to store the data outside of Power BI, which negated the elegance of establishing SUSHI calls from within Power BI.

We finally settled on using the SUSHI harvester in Alma to gather and store our data. (Figure 4)

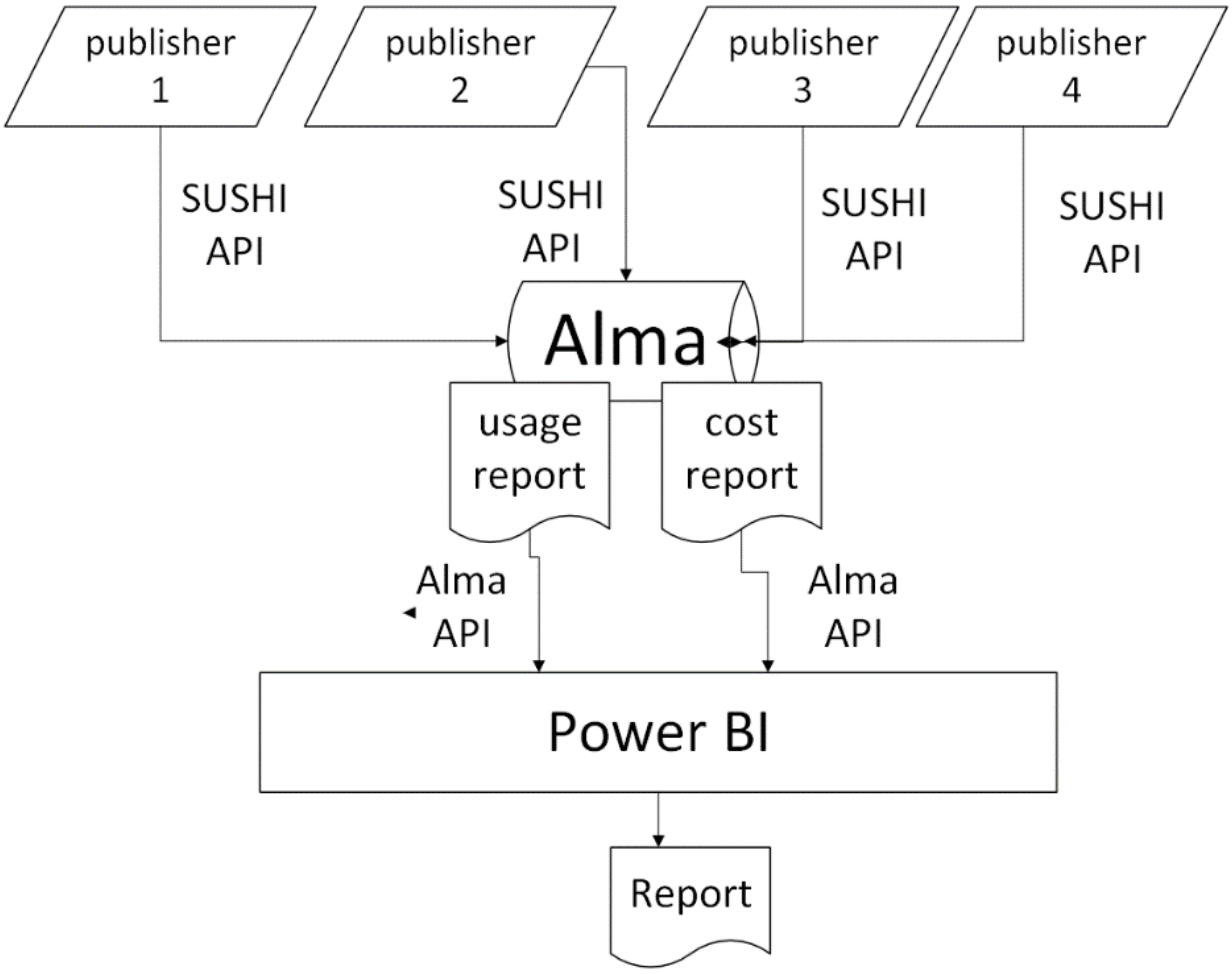

Figure 4

Alma is designed to both harvest and store usage data, and we knew we could get the information out of Alma using Alma Analytics' SQL database. However, if we wanted a fully dynamic report, we would then need to query Alma to retrieve cost and usage reports on a monthly basis. We explored querying the Alma Analytics API using Power Query (M) within Power BI but found it very difficult to troubleshoot errors. (Figure 5)

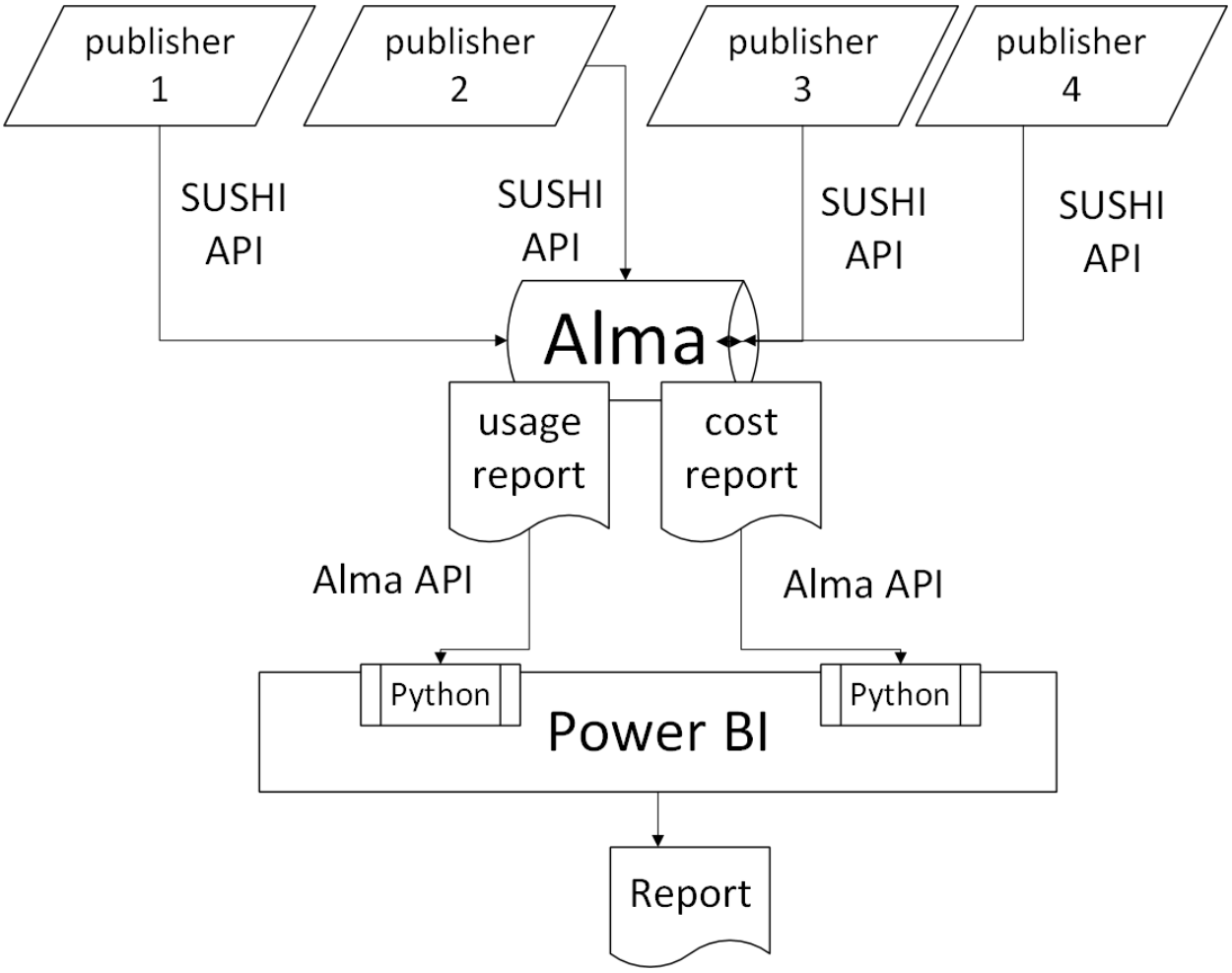

Figure 5

We decided instead to use a Python script embedded in Power BI to query the Analytics API to retrieve both cost and use data. We then used Power BI to transform the data, calculate and visualize cost per use, and distribute our dashboard to internal stakeholders.

In this final model SUSHI API accounts are set up for as many vendors as possible in Alma, which harvests the tr_j1 reports monthly. Alma Analytics reports are established to capture cost and usage statistics. Python scripts embedded in a Power BI M query retrieve each report via an API and loop through the data using a resumption token (required for data of more than 1,000 rows). The Python code pulls in the complete dataset each time we want to update it, which we can do by simply pressing the Refresh button in Power BI. More M code simplifies the data. DAX code then merges the data from the two datasets to calculate cost per use. Power BI's visualization tools are used to create a dashboard that can be shared with stakeholders.

If you want to implement a model like this one in your system, full step-by-step details can be found in Appendix A.

The Dashboard Visualized

Our current dashboard includes four views. The first uses a bar chart to display journal titles and their cost per use with the ability to filter by fund and fiscal year. (Figure 6) The second provides a table of usage data across multiple fiscal years, filterable by fund and searchable by platform, publisher or title. (Figure 7) Our third display is a table of titles that might be cheaper to obtain by ILL than subscription, based on the usage patterns and average ILL cost. (Figure 8) Finally, we have an About page that provides information about the dashboard.

Figure 6

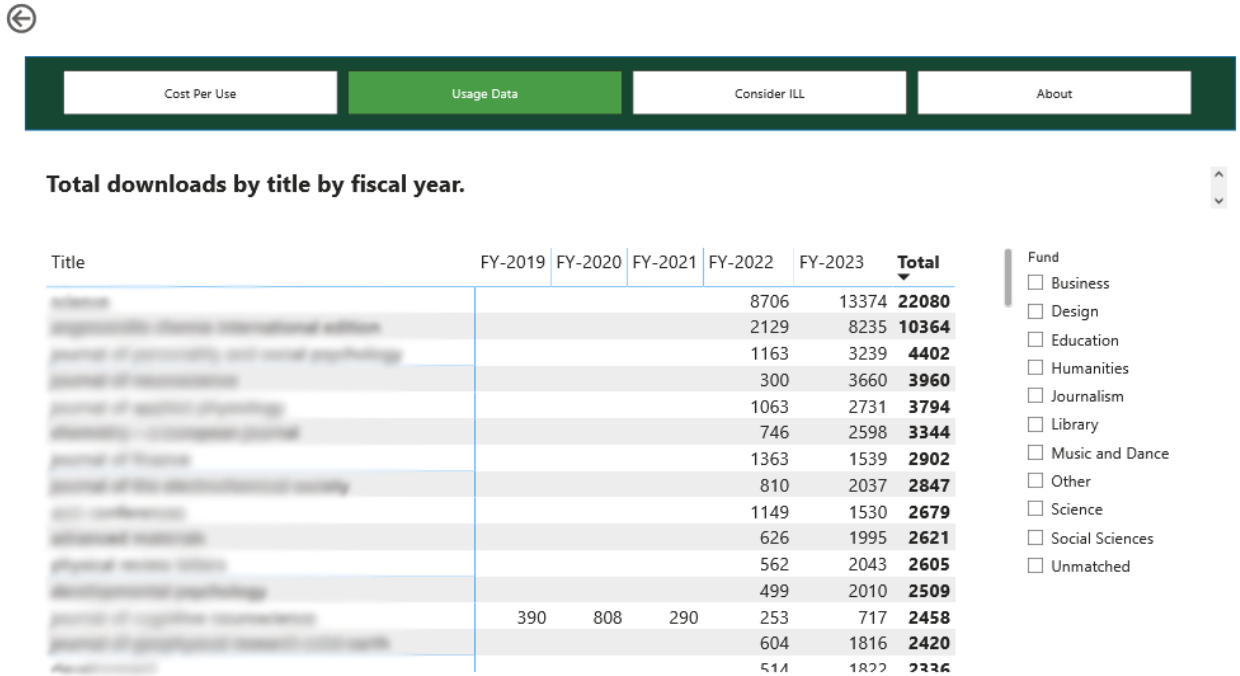

Figure 7

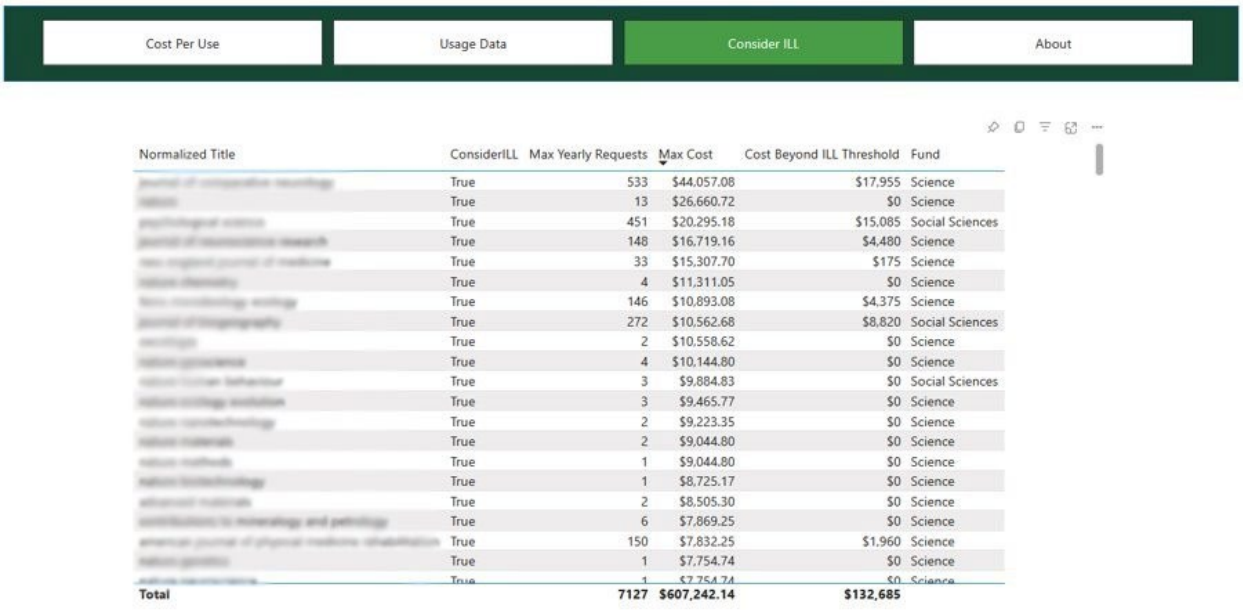

Figure 8

Cost Per Use Defined

Because this is a proof of concept for our technical pipeline, it uses a very simple definition of cost per use, as the number of uses as defined by COUNTER tr_j1 reports divided by the annual amount we paid for the product — counting only journal use and excluding databases and eBooks. We do not, for example, address any of the issues of open access articles discussed by Jabaily (2020) or addressed by the Unsub [1] product. However, the nature of Power BI as a tool that can aggregate and visualize data from a variety of sources means that we hope that this pipeline will be the basis for creating more robust analyses. In the future, for example, we might adjust for open-source access, include eBooks and databases, or incorporate elements such as tracking the journals our faculty and graduate students cite or publish in.

Assumptions and Limitations

In the process of writing up this project we had to distill it down to steps that are communicable and helpful to others. In reality, this was an iterative experience with several dead ends. We frequently tested elements as proof of concept before investing time in developing those same elements on the larger scale. It's possible there are more efficient approaches, especially related to the normalization of titles and ISSNs. We decided to stick with what was working and not get bogged down in perfection, in the spirit of getting it done. We know there are steps we'd like to improve, some quirks that need troubleshooting, and plenty of opportunity for additional report types and other upgrades, as well as regular ongoing maintenance. We welcome feedback and inquiries via email or via issues on our GitHub repository.

There are gaps in the data where publishers don't provide SUSHI access, or COUNTER 5 formatted reports, or don't do so in a timely fashion. Additionally, we have not yet systematically filled in gaps between the beginning of the existence of COUNTER 5 and the initiation of SUSHI harvesting within our instance of Alma. At times the titles or ISSNs don't line up perfectly, or we have usage reported from multiple platforms. We removed the Gale and JSTOR aggregator platforms from our data, and don't currently have a way of evaluating this usage. We do not include print usage of our collection at this time. In our initial Excel model print usage was very low and usage was not of recent issues, so we determined that including print usage would not add value to our model.

We made several assumptions along the way that led to the data being useful but not always completely precise. For example, if we don't have a payment for the current year, we base our calculation on the previous year's payment, even though it is likely not the same. We also have instances where the calculated CPU is zero. This includes cases where the title and/or ISSN aren't a perfect match between our usage and cost data, journals that we pay for as part of a package but receive usage on a per-title basis, and journals only available on platforms we removed from consideration because they're primarily backfiles, like JSTOR.

In short, the data is still messy, and while we have plans for how to fix some of these issues, it will never be perfect. Power BI's visualization can be deceptive, even seductive. We think our model looks pretty, but we still must put our thinking caps on to be sure that those attractive charts mean what we think they mean. Each of our displays is meant to be a starting point for cancellation decisions rather than a rigid list of things that must be cancelled.

Discussion and Future Work

We set up this model using only the tr_j1 report, and once we gain some experience consulting the model for assistance with deselection decisions, we plan to add more report types. We're interested in adding historical usage data by uploading COUNTER 4 stats to Alma in order to add more years into the model. We may also consider additional calculations beyond CPU that might be useful for our collection assessment needs. The way these reports are written, we shouldn't have to make major changes until there are significant changes with either COUNTER or Alma Analytics. This is intentional, so that we can keep the pipeline running with as little maintenance as possible.

Power BI lets us share out a dashboard that any of our colleagues can use to see the visualizations and tables of data. We have yet to use this model for significant cuts, and surely that experience will teach us a lot about what else we need to do to improve this tool. In the meantime, we have sent the model out to our collection managers and subject librarians for review, feedback, and to begin using it to help inform decisions. We encourage our colleagues to use the data as a starting point for collections decisions, rather than as the single tool to make the decisions. Along with the data we've shared information about the limitations of that data. So far, we have used this model to extract statistics for administrative reports, and as a starting point for database assessment.

Visualization can help us analyze the data to uncover patterns, trends, and correlations that might not be apparent from a spreadsheet alone. These deeper insights can help us make better decisions. By automating the process, it makes year-over-year comparison more achievable. It also sets us up for expanding the types of reports to intake and types of content to analyze in a reproducible way. With regular updates, the tool will grow more useful over time as gaps are filled in. We expect this tool, and permutations of it, will help our collections management team make data-driven decisions to meet the collections and access needs of our students and faculty.

Notes

[1] Unsub: https://unsub.org/

References

Boardman, Erika, and Jennifer France Thompson. 2022. "Visualizing COUNTER Metrics with SUSHI: Exploring Alma Analytics for E-Resource Evaluation." *Serials Review*, September, 1–3. https://doi.org/10.1080/00987913.2022.2120328.

Gribaudo, Ben. 2018. "Power Query M Primer (Part 5): Paradigm." February 28, 2018. https://bengribaudo.com/blog/2018/02/28/4391/power-query-m-primer-part5-paradigm.

Harker, Karen R. 2022. "The Depths of Cost-per-Use: Historical Context and Applications." *Library Trends* 70 (3): 355–86. https://doi.org/10.1353/lib.2022.0000.

Jabaily, Matthew J. 2020. "Recalibrating Cost-per-Use: Implications of COUNTER Release 5 and Unsub." *Serials Review* 46 (4): 292–99. https://doi.org/10.1080/00987913.2020.1850040.

Pesch, Oliver. 2015. "Implementing SUSHI and COUNTER: A Primer for Librarians: Edited by Oliver Pesch." *The Serials Librarian* 69 (2): 107–25. https://doi.org/10.1080/0361526X.2015.1063029.