New Developments for Journal Package Analysis and Data Visualization

Nat Gustafson-Sundell, Pat Lienemann, Luwis Andradi, Evan Rusch, and Jeff Rosamond

Presenters

ABSTRACT
The presenters demonstrated the most recent results and insights from a long-running, iterative Collection Management Technology (CMT) project. Demonstrations included a journal package level analysis report, an interactive subject analysis report, several data visualization designs for journal package analysis, and a program to automate report production, including data visualizations. Finally, the current, cyclical journal collection review process was compared to a past cycle to highlight how the CMT team’s development priorities have changed. Instead of focusing primarily on efficiency and information yield in the tools they build, the team has now prioritized the development of tools that enable interaction with the data and broaden understanding.

KEYWORDS
Library liaison outreach; collection analysis; collection development; collection outreach

Introduction
At Minnesota State University, Mankato, the Collection Management Technology (CMT) team of the Library Services Department has for several years continuously and iteratively developed journal collection analysis reports to meet several purposes, including journal-level and package-level collection development, and to support liaison outreach to academic departments. For the 2020-21 comprehensive journal collection review, the team assessed the processes and results of the last comprehensive review in 2018-19 to develop a revised approach, especially to increase library and campus stakeholder involvement in discussion and decision-making. This team presented new developments as a panel, each focusing on a facet of the team’s work, while the final section contextualized the overall effort.

Nat Gustafson-Sundell provided a quick background and review of the team’s work and demonstrated two reports, the Package Level Analysis Report (PLAR) and the Subject Analysis Scratchpad (the Scratchpad), as examples of Mankato’s most recent developments for journal package review and outreach. Pat Lienemann demonstrated how he has used Tableau as a development environment for journal data visualization and he highlighted useful ideas from other libraries which have most recently informed his development priorities. Luwis Andradi demonstrated how he automated final report production in Jupyter Notebook using the Python programming language and discussed the advantages of using Jupyter Notebook for report production. Evan Rusch compared how Mankato performed journal package collection development two years ago to what they are doing now, especially to highlight how the team’s newest developments can help improve library liaison outreach. Jeff Rosamond did not present, but he was available for questions about data matching and validation.

Background
The CMT team has previously presented other applications of their reports and how they develop the reports, including the data matching and validation methodology, at other conferences, and several of these presentations are available from the team’s library institutional repository, searchable under the
authors’ names. In summary, the team has developed four standardized collection analysis reports and numerous customized or ad hoc reports for the purposes of collection assessment (including collection development), collection evaluation (including accreditation and program review), and outreach. All of the reports rely on a key list, which can be any list of journals, and which serves to provide the basis for data integrity in the absence of database normalization. The key list is used as the foundation to match all other relevant, available journals-related data from a variety of sources, including the Integrated Library System (ILS) and Electronic Resource Management (ERM) system, vendors, and other third parties.

These data sources describe electronic and print journal holdings and coverage, journal quality (as a variety of citation-related variables), link resolver requests, interlibrary loan requests, print browses, prices (from price lists and purchase orders), and additional electronic journal usage data, as provided by vendors on the J1, J2, and J4 reports defined under the COUNTER Code of Practice. The resulting base reports each include 100–150 variables derived from these data sources. These base reports provide the basis for numerous standardized data visualizations, as well as final, executive summary reports, designed by selecting from among these visualizations.

**Highlighted reports**

For journal package collection development, Gustafson-Sundell demonstrated two of the reports, the PLAR and the Scratchpad. These two reports enable Mankato to address two fundamental areas of inquiry for journal package collection assessment: (1) How do journal packages perform relative to one another? (2) If it is necessary to cancel one or more journal packages because of budgetary constraints, what would be the subject level impacts of cancelling any given packages, and which journals would provide curricular value if continued as individual subscriptions? In addition, the Scratchpad was designed to increase liaison engagement with journal collection development and to enable liaisons to communicate more meaningfully with campus partners about journals as curricular resources.

The PLAR is provided as a Microsoft (MS) Excel workbook to the Journals Review Committee (JRC), which is a journal collection development committee, and to liaisons. The PLAR includes a summary worksheet including all of the variables, as well as numerous worksheets with a variety of plots and lists to improve the legibility of the data. The PLAR summary worksheet, also called the base report, is based on a different report, called the Collection Review (CR) report. The CR is a journal-level report including all journals subscribed by Mankato, whether they are in packages or not. For the PLAR, Mankato sums the journal level metrics up to the package level, and then the PLAR is augmented manually with additional package-level only data. The PLAR workbook also includes a rough glossary explaining the variables. These variables can also be understood categorically, as they describe (1) package coverage and overlap with journals provided by aggregators; (2) package quality; and (3) package usage and cost-effectiveness. An example PLAR, anonymized and truncated, was provided at the following link: [https://link.mnsu.edu/package](https://link.mnsu.edu/package).

Similarly, the Scratchpad is also provided to the JRC and to liaisons as a MS Excel workbook. A goal of the Scratchpad is to encourage immediate, intuitive interaction with the data through plots (as a pivot chart), so the interface is deliberately limited. The Scratchpad opens on a pivot chart where users can change the subject for analysis and the variables of analysis. The variables are inherited from the underlying standardized base report, so there are more than a hundred variables available. This underlying base report is hidden, so that users will not be intimidated by the data.

More advanced users could also change which packages are included in the analysis, to see, for example, how the cancellation of a specific package will impact a subject, but a hope is that users might graduate from the Scratchpad to design more complex approaches to the data on their own. The Scratchpad is designed to excite and intrigue potential users, including liaisons and campus-wide partners, who might otherwise not engage with the tabular data. By interacting with the pivot chart, the hope is that users will learn more about the data and its applications, so that they can understand
more clearly how students utilize journals, which will enable them to contribute more meaningfully to conversations ranging from collection development to program review. An example of the Scratchpad can be viewed at the following link: https://link.mnsu.edu/mankato-nasig-video (starting at 11:43).

The CMT Team offered a workshop to all librarians at Mankato at the end of the spring 2021 semester so that they could learn how to use both the PLAR and the Scratchpad. This workshop was also utilized by the CMT team as an informal focus group to gain ideas about how to develop these reports to make them more useful and accessible in future iterations. Another workshop will be offered in fall 2021, where the next versions of these reports will be shared.

The data visualization laboratory

When Lienemann started at Mankato in 2018–19, he joined the CMT Team specifically in the role of data visualization (viz) developer, in part because he had some background with data viz through an elective class on digital humanities tools he had taken in graduate school. While all members of the team contribute to data viz development using a variety of tools, depending on the given report or the staging of the report, Lienemann has primarily focused on implementations using the commercial product Tableau, due to his familiarity with the software, its usefulness for prototyping, and because Tableau can be used freely by academic researchers, upon request.

When Lienemann first joined the team, they were engaged in an earlier cycle of package analysis. Lienemann demonstrated two example plots developed in that cycle. The first example compared package usage trends over a five-year period.\(^3\) Usage via the subscription platform was distinguished from usage via other platforms, such as aggregators. This fundamental plot reveals the relative overall value derived from the package subscriptions specifically. The second example compared title and citable document counts within packages over a three-year time period.\(^4\) This plot provides a sense of whether the packages are growing or declining. This second type of analysis can be helpful to set the stage for annual package negotiations, because vendors often do not sufficiently acknowledge package decline in their pricing models.

In that earlier cycle, the team needed a decisive method to identify single title subscriptions to add back if a package was selected for cancellation. Lienemann developed several approaches to this problem, one of which he demonstrated because it was especially useful.\(^5\) He configured a treemap for each package so that individual journals would appear as boxes within the treemap.\(^6\) The sizes of the boxes would depend on all usage of the journals over five years, while the shading of the boxes would reveal how much of the usage was subscription-specific. Finally, a numerical value within each box provided a citation-based journal quality measure. This combination of variables made it much easier to identify which journals merited attention as candidates for individual subscriptions.

As Lienemann prepared for the 2020–21 cycle, he looked back at the team’s previous work. The team had been successful using some of the plots for decision-making and communication, but Lienemann began to question the usefulness of some of the team’s earlier efforts. He asked, “Do these viz tell us what we need to know or are they just cool looking?” and “What variables should we be focusing on (now) to inform decisions?” In order to make sure the team could benefit from the wisdom of other libraries, he conducted a literature review, especially to focus on data viz, but also including any innovative collection analysis variables which could be duplicated and plotted at Mankato.

As useful background, Kathryn Wissel and Lisa DeLuca identify two purposes for visualizations: (1) as a way to understand collections and (2) as a way to make decisions.\(^7\) Their paper reinforced Lienemann’s feelings that he, personally, was gaining a good understanding of Mankato’s journal collection through the process of developing viz, but that he had not yet learned how best to distill that knowledge in a way that could be shared so others could understand it intuitively. Lienemann decided to seek out variables and visualizations recommended by other librarians.

Multiple authors stated that they considered cost per usage (CPU) the most important variable for collection assessment. CPU is defined variously, but generally as cost divided by usage, where usage is most often derived from the J1 report defined by COUNTER. Lienemann focused on innovative
approaches to CPU. For example, Megan Kilb and Matt Jansen suggest plotting usage against CPU and grading the value of the journals as acceptable, good, problematic, low value, or unacceptable, according to a benchmark scale applied to the plot.\textsuperscript{8} Lienemann applied a similar concept to Mankato’s data by developing a rough overlay derived from the published paper. Using this overlay, it was simple to grade journals either within a package or the collection overall, for the purpose of cancellations or identifying add-backs.

Mathew Jabaily, James Rodgers, and Steven Knowlton provide another way to look at CPU with a metric called the adjusted CPU (ACPU). They argue that the actual value of a current subscription, if the deal provides post cancellation access rights to content from previous subscription years, is only provided by the content published during the current year of the subscription deal. An adjusted CPU can be calculated by dividing the cost of the current subscription by usage of only the content published in the most recent year (available via the J4 report, as defined by COUNTER).\textsuperscript{9} Utilizing the ACPU, the perceived value of a current-year subscription can differ drastically from what one sees using CPU alone. Lienemann plans to incorporate ACPU in Mankato’s newest data viz development efforts in order to distinguish packages and journals more meaningfully.

Lienemann also presented the California Digital Library’s (CDL) Weighted Value Algorithm (WVA), which combines utility, quality, and effectiveness into one metric.\textsuperscript{10} In a previous cycle, Mankato prototyped a weighted value by combining package content volume, quality, and usage, which can be calculated from a variety of different variables in varying proportions. Following the model of the WVA, Mankato may consider further permutations. While a weighted value can be used to rank packages or journals for retention or cancellation, as the CDL has done, Lienemann and Rusch are especially interested in utilizing weighted variables as a means to provoke conversations with campus stakeholders.

Data visualization standardization and automation

Andradi demonstrated programs he has developed using Python in the Jupyter Notebook environment to automate the production of plots and final reports. He explained why the team has preferred Python and Jupyter Notebook for automating the standardized plots, as distinguished from prototype or ad hoc plots, which might be developed in other environments. Jupyter Notebook is free and open source, efficient to learn and use, powerful, and extendable with additional packages. The environment enables reproducible results, in which manual effort and user errors are both minimized. The environment can also produce a variety of portable outputs, including MS Excel workbooks.

Andradi developed programs to create data viz for the PLAR as well as two other reports. These two other reports are the “SciMB,” which maps the entire universe of academic journals as reported on the ScImago Journal & Country Rank website\textsuperscript{11} to all of the other data sources readily available at Mankato, and the Liaison Journal Collection Analysis (LJCA) report, which is essentially any subject area or subject category slice of the SciMB.

As a graduate assistant at Mankato, one of Andradi’s earliest projects was to duplicate and automate production of the twenty-one plots and lists previously developed by the CMT team and standardized for both the SciMB and LJCA reports. He then added new plots and lists in consultation with Rusch, as additional elements which could be selected from a MS Excel workbook called a “pre-finished report.” The pre-finished report provides the basis for finished reports, or executive summaries, which can be customized to a variety of purposes based on the selection of elements. Along these lines, he also prototyped a program to output a final report as a MS Word document, which includes a small selection of high-impact plots and lists. This prototype final report was designed primarily to support liaison visits to academic department meetings, but other final reports could also be standardized and automated for other purposes. Finally, he developed new data viz for the PLAR, both on his own and in consultation with Lienemann. Andradi wrote a program to automate the production of these PLAR data viz, although this last project remains a prototype at this time, to the extent the team has not yet had time to review the results.
Andradi showed a sample of his code and explained how different packages, such as Plotly and Matplotlib, produce differing results, but he also made it clear he can duplicate and automate anything Tableau can do using packages like these. He also displayed how he wrote the code so that it could be easily updated from year to year, which is especially important for annualized data. Finally, he ran an example report as a live demonstration, which showed how his program can convert a massive table including over 100 variables and more than 30,000 journal records into a set of dozens of plots and lists in less than a minute, with multiple outputs including pre-finished reports as MS Excel workbooks and finished reports as MS Word documents.

**New journals collection review**

Rusch compared the last collection-wide review cycle to the current cycle to explain how the CMT team has developed new outreach priorities, tools, and methods. In 2018–19, the JRC conducted a comprehensive journal collection review utilizing the collection analysis reports generated by the CMT. That review’s first phase included an initial assessment of journal packages and relied heavily on the first iteration of the PLAR. This phase of the review winnowed the list of over twenty journal packages into a set of seven that were targeted for deeper analysis. Packages chosen for greater scrutiny tended to have higher CPU or significant content overlap with aggregators.

In the second phase of the 2018–19 review, individual journal titles within packages were analyzed. CPU was recalculated using a variety of approaches to determine which journals would be added back if packages were broken or cancelled. The cost of journals expected to be added back offset the savings from dropping a package, impacting the final cancellation decision. Three journal packages were cancelled because of the collection review. All three were subject-specific packages rather than broader multi-disciplinary packages.

In 2020–21, the JRC began a new comprehensive journal collection review. The CMT again provided reports to inform the work of the JRC and assessed the 2018–19 review to seek ways to improve the process. Although the JRC did not express concern, the CMT team wanted to address the fact that there was not much debate or discussion within the JRC about the collection development decisions in 2018–19. In the earlier review, extensive charts and data were demonstrated in JRC meetings by members of the CMT, and the JRC came to an easy consensus on cancellations without much discussion. While this could be considered a success, in hindsight it seemed possible that since most members of the JRC did not participate in the analysis itself, it was difficult for them to offer debate or alternative perspectives. As a result, important anecdotes about specific programs, curriculum, professors, and assignments that add depth to understanding of collections and their uses, and which might have risen to the surface in the course of a healthy debate, were likely missed.

Upon reflection, the CMT team also raised concerns regarding liaisons’ comfort in communicating about the collection decisions with the departments they served. If liaisons were not involved in the analysis that drove the collection decisions, it seemed they might not be comfortable sharing or discussing the plots or lists created for them or responding to questions from academic departments.

An additional concern was that the 2018–19 review might have relied too much on CPU. Because subject-specific packages tend to have lower usage than multi-disciplinary packages, and a much smaller pool of potential users, the CPU for these packages will tend to be higher, depending on price. In 2018–19, because of the reliance on CPU, subject-specific packages were targeted, with the consequence that a few individual departments were disproportionately impacted. Some departments expressed frustration later that they bore the brunt of cancellations. Other departments have since provided testimonials indicating that subject-specific journal packages are essential to support teaching and learning within their disciplines. The CMT team realized that they needed to develop a process which would inspire the collection of additional qualitative data, not just as librarian anecdote, which has always been offered, but as actual testimonials from teaching faculty about how journals and packages are used for specific classes or programs.
Finally, the CMT team also discussed concerns about how data visualization was implemented during the 2018–19 review. Tableau enabled the team to create complex new plots, but some of these visualizations may not have been understandable by all audiences, so again, participation in discussion and decision-making may have been inadvertently discouraged.

Following this assessment of the 2018–19 review, new goals were set to increase participation in the 2020–21 review, beyond just the CMT, and to improve the collection of qualitative information from academic departments about how they use journal packages especially. In addition, the CMT team chose to set a goal to identify one large multi-disciplinary package as a priority for cutting in the future, intentionally to avoid accidentally biasing cancellations again, and to ensure equitable impacts across academic departments. The CMT and JRC identified several reasons why it is important to start early to prioritize a large multi-disciplinary package for cancellation, rather than waiting until there is budgetary cause. Starting early would allow librarian liaisons and the JRC to begin campus discussions about the impacts arising from the cancellation of the package, and to identify alternatives if necessary. It would allow the JRC to respond quickly, without the need to rush any decision-making due to budgetary cause. It would be possible to weigh subject-specific curricular and research demand more thoughtfully, so that any loss of content would be shared across campus disciplines more equitably. Any multi-year license agreements could be modified as soon as possible so that the identified package would be cancelable at the time of need.

Given the goal of increasing participation in discussion and decision-making, the CMT team needed to develop an improved approach to subject analysis, so that the JRC, liaisons, and campus-wide stakeholders would understand how cutting a multi-disciplinary package would impact specific programs and departments. The CMT team wanted to ensure that any academic department could see the impact of cancelling a large package, especially to note three categories of impact: (1) loss of overall quality or high prestige titles, (2) loss of regularly used content, which might be understood as curricular supply, and (3) loss of collection diversity or the overall amount of content, which might be understood as research supply.

To meet the goal of bringing more people into the process of analysis and discussion, the CMT offered a workshop for all librarians to experiment with a variety of collection analysis reports, including the new Subject Analysis Scratchpad. The Scratchpad seemed like an immediate success to the extent librarians started interacting with the journal data and asking questions in ways the CMT team had never heard before. The workshop setting itself enabled a broader conversation than had ever previously occurred and will become a featured component of future efforts. As attendees explored and asked questions, they discussed their perceptions of the programs and departments they serve as liaisons and how library collections are used by those disciplines. The conversation yielded early examples of how qualitative data could be integrated into decision-making more meaningfully. Because they were actively participating in collection analysis during the workshop, it seemed the attendees would walk away with increased confidence to share information with their liaison departments. Also, it seemed some of the attendees had benefited by improving their competence using data, generally, and had learned some basic applications of data analysis tools.

If liaisons are empowered to interact with the data through the data viz, using tools like the Scratchpad, the CMT team’s impression so far is that they will find more meaning in the data which they will be able to communicate more readily. It may be too early to say, but the team sees the potential for a cascade of positive impacts, including improved communication by librarians, both within the library and across campus, which in turn will increase the reach of the library’s conversations with academic departments. Over the years of the CMT team’s overall project, they have already seen how the most data-competent liaisons have used LJCA reports to garner useful, and sometimes very positive, feedback from their departments.

Ultimately, of course, the overall purpose of the CMT team’s work is for the library’s journal collection to serve the university’s mission and goals as optimally as possible within budget constraints. If the library’s liaisons can use the new journal collection analysis tools confidently, like the
newest version of the PLAR and the Scratchpad, but also including older tools, like the SciMB and the LJCA reports, our hope is that they will have the information they need to understand and collect testimonials from academic departments. Such testimonials will provide insight during current and future collection review cycles, and they can be used to reaffirm library value to the university’s administration in the library’s annual reports.

Conclusion

In summary, the Mankato CMT team has revised its collection review strategy in several meaningful ways. They have shifted analysis away from overreliance on universal variables, such as CPU, and toward a better understanding of discipline-specific impacts. Through Lienemann’s work especially, they have sought increasingly nuanced and meaningful implementations of universal variables, such as CPU, and developed weighted variables as devices to improve communication specifically. They have successfully increased the involvement of more librarians in analysis and decision-making, with the hope this increase will spread further across campus as cyclical reviews continue. They have developed tools and data viz designed in the first place to be produced cost-effectively, but also, and increasingly, they are developing new tools to improve audience understanding and participation. By improving audience participation, they have increasingly collected more meaningful qualitative data, which in turn can be utilized in multiple ways.

The team has turned a corner in the latest iteration of this long-running project. Instead of exhausting the possibilities of collection analysis and report development over the years, the team has found new areas for development. Having focused primarily on technical advances earlier in the project to improve the efficiency of report production, the team is now increasingly working to enable library and university stakeholders to take advantage of the evidence-based approaches developed by the CMT. While the team focused on adding more variables to reports, or more data viz, in the past, the work now is to refine the selection of variables and viz to increase audience understanding, and to reframe reports to make them more interactive. In other words, the CMT team’s development priorities are evolving away from designing for efficiency and information yield primarily, and more toward facilitating involvement.

Notes

1. See https://cornerstone.lib.mnsu.edu/.
2. COUNTER was formerly an acronym, but is now treated as a brand-name; see https://www.projectcounter.org/about/.
3. See https://link.mnsu.edu/nasigfig1.
4. See https://link.mnsu.edu/nasigfig2.
5. See https://link.mnsu.edu/nasigfig3.
6. See https://www.geeksforgeeks.org/treemap-using-plotly-in-python/.
7. Kathryn M. Wissel and Lisa DeLuca, “Telling the Story of a Collection with Visualizations: A Case Study,” Collection Management 43, no. 4 (2018): 264–75, https://doi.org/10.1080/01462679.2018.1524319.
8. Megan Kilb and Matt Jansen, “Visualizing Collections Data: Why Pie Charts Aren’t Always the Answer,” Serials Review 42, no. 3 (2016): 192–200, doi:10.1080/00997913.2016.1207479.
9. Mathew J. Jabaily, James R. Rodgers, and Steven A. Knowlton, “Leveraging Use-by-Publication-Age Data in Serials Collection Decisions,” in Where Do We Go From Here? ed. Katina P. Strauch, Leah H. Hinds, Beth R. Bernhardt (Purdue University Press, 2016), 292–302.
10. Jacqueline Wilson and Chan Li, “Calculating Scholarly Journal Value through Objective Metrics,” California Digital Library, https://cdlib.org/cdlibinfo/2012/02/13/calculating-scholarly-journal-value-through-objective-metrics (accessed July 14, 2021).
11. See https://www.scimagojr.com/.

Disclosure statement

No potential conflict of interest was reported by the author(s).
Notes on contributors

Nat Gustafson-Sundell is Associate Professor and Collections Librarian, Minnesota State University, Mankato, Minnesota.

Pat Lienemann is Assistant Professor and eAccess & Discovery Librarian, Minnesota State University, Mankato, Minnesota.

Luwis Andradi was the Collection Management Technology Graduate Assistant, Minnesota State University, Mankato, Minnesota.

Evan Rusch is Associate Professor and Reference/Instruction Librarian, Minnesota State University, Mankato, Minnesota.

Jeff Rosamond is Technical Services Technician, Minnesota State University, Mankato, Minnesota.