Visualizing the Intersections of Impact and BTAA Libraries' Investments in the Research Enterprise Using Open Government Data: An Exploratory Model Using Tableau

Sarah Murphy
The Ohio State University, USA

Abstract

Purpose and goals

Grant funding serves as an important proxy for quantifying the value of the academic library. Past studies rely on researchers to self-report whether they used and cited library resources when crafting successfully funded research proposals (Kaufman 2008, Tenopir et al. 2010). More recent studies seek to quantify library support for publication and the grant seeking process using data gleaned from Scopus, Web of Science, Journal Citations Reports, and other tools (DeGroote et al. 2020; Monroe-Gulick, Currie, and Weller 2014). With the advent of the Federal Funding Accountability and Transparency Act of 2006 and the NIH Public Access Policy, libraries can now leverage open government data to explore the relationship between grant funding and investments in library collections and services. This study explores modeling open data on government spending and federally funded research outputs to (1) visually demonstrate how libraries contribute to the research enterprise by providing information scholars need to both develop and sustain their research agendas, and (2) allow libraries to visualize and utilize this same data to inform the development of library services and collections.

Design, methodology, or approach

Open government data now allows libraries to identify publications authored by their institution's faculty that result from federally funded research. Libraries may model this data to determine what journals faculty chose to publish their research in, as well as what journals faculty cite, and the value of these individual research outputs. The model for this project was created by first identifying all NIH project grants awarded to BTAA faculty between FY2010 and 2018 using the NIH RePORTER data portal. BTAA schools collectively expended more than $11 billion on research in FY2019 and have a robust program for optimizing researchers' access to member libraries' collections. A list of publications associated with these grant projects was then downloaded and enhanced by pulling lists of citing papers and reference papers to identify what journals were used to inform the author's research and what journals cite the author's research. NLM IDs (nlmids) were added wherever possible using the journal title for each publication to later identify and use MeSH terms as visualization filters. All data was then modeled in Tableau using a series of relationships and visualized in a series of interactive dashboard charts.

Practical implications or value

Modeling and visualizing the outputs of successful grant-seeking using Tableau allows libraries to explore this data at both a high aggregate and lower level of detail. This project demonstrates how to assemble and utilize such data to both illustrate libraries ongoing contributions to the research enterprise and inform library collections and services.

Libraries help scholars develop and sustain their research, and scholars in turn fund academic libraries. Open government data now offers libraries the opportunity to showcase their contributions to the research enterprise, and in turn, quantify their value. This study examines data available through the NIH RePORTER and iCite tools to explore the relationship between grant funding and investments in library collections and services at the fourteen current members of the Big Ten Academic Alliance (BTAA). Specifically, the study explores modeling open data on government spending and federally funded research outputs to:

1. visually demonstrate how libraries contribute to the research enterprise by providing information scholars need to both develop and sustain their research agendas and

2. allow libraries to visualize and utilize this same data to inform the development of library services and collections.

Literature review

With the advent of the Federal Funding Accountability and Transparency Act of 2006 and the NIH Public Access Policy, libraries can more easily link publications authored by their institution's faculty to funds competitively awarded to their institution by the NIH, NSF, and other federal agencies. Such data is not only freely available through the NIH RePORTER, but also via USAspending.gov and a number of proprietary databases, including Clarivate's Web of Science. Further, libraries can use these tools to determine what journals their faculty choose to publish research in and what journals faculty reference and cite. This information, when combined with other library data may be used to calculate a return on investment (ROI) for each journal title. When combined, or displayed with other measures, such as the Relative Citation Ratio (RCR) value for each article BTAA faculty publish in a journal, download trends, and more, this calculation may both quantify library value and guide library decision-making.

Previous studies of library value relied on researchers to self-report whether they used or cited library resources to develop successfully funded research proposals. Kaufman (2008) surveyed faculty at the University of Illinois at Urbana-Champaign to determine "the role that the library plays in their research and grant process." The survey asked researchers to share the importance of references in their grant seeking process and whether the references they used in proposals were accessed through the library. Tenopir et al. (2010) surveyed faculty across seven institutions, specifically asking participants to identify the number of references they cite in grant proposals, reports, and published articles and approximate the number of articles and or books they read but did not cite when preparing proposals. Both Kaufman and Tenopir et al. combined survey data with other publicly available financial data to calculate a high-level aggregate return on investment value quantifying the library's contribution to the research enterprise.

More recent studies seek to quantify library support for publication and the grant seeking process using data gleaned from Scopus, Web of Science, Journal Citations Reports, and other tools. Monroe-Gulick, Currie, and Weller (2014) examined citations included in successful NIH and NSF proposals to confirm that the University of Kansas Libraries' collections supported KU researchers. DeGroote et al. (2020) explored the relationship between research productivity, collection use, collection size, library funding, and research and development expenditures at 81 research-intensive universities in the United States, finding "full-text article requests, followed by library material expenditures and research expenditures, were … the best predictor of research productivity."

Other studies indirectly show how libraries support university research while using bibliometric data to support the evaluation of large, federally funded programs. Qua et al. (2021), for example, used RCR values with publication counts, citation counts, and other measures to evaluate the success of KL2 funded scholars affiliated with 3 NIH funded Clinical and Translational Science Award (CTSA) hubs. Available through the NIH Office of Portfolio Analysis' iCite application, the RCR value measures influence of an article by examining its co-citation network and benchmarking the article in relation to other articles in its field to the median value 1.0 (Hutchins et al. 2016). The measure, as calculated is field- and time- normalized. Yu and Hayes (2018) used RCR values with other bibliometrics to evaluate the impact and influence of the University of North Carolina at Chapel Hill's Cancer Cell Biology (CCB) program.

Design, methodology, or approach

The Data Model

Collectively, the BTAA expended $11.5B on research in FY2020 and roughly a third of these expenditures reflected NIH funded projects.[1] To model and then visualize the relationship between grant funding and investments in library collections and services, a list of NIH funded projects between FY2010 and 2018 for each BTAA school was first downloaded from the NIH RePORTER and compiled into one file. Since the NIH Public Access Policy requires "anyone submitting an application, proposal or report to the NIH must include the PMC reference number (PMCID) when citing applicable papers that they author or that arise from their NIH-funded research," a list of publications associated with these grants with the publication years 2010 to 2022 was then assembled using the publications tab in the NIH Reporter (https://publicaccess.nih.gov/include-pmcid-citations.htm). The export file for publications includes the core project number, publication title, publication authors, journal title, the PMCID, the PMID, the RCR value for the publication and more. The file also includes links to each publication's PubMed abstract, as well as the full-text of

[1] Calculated using FY2020 total cost for NIH funded BTAA projects listed in the NIH RePORTER and total research expenditures listed on Big Ten Academic Alliance University Data At-A-Glance 2020 information sheet available at https://btaa.org/docs/default-source/reports/2020-at-a-glance.pdf?sfvrsn=86845b67_6.

the article in PubMed Central, lists of related publications in both PubMed and Google Scholar, and lists of articles citing the publication in PubMed and Google Scholar.

In all, NIH RePORTER data for a total of 18,845 unique core project numbers awarded to BTAA schools between FY2010 and FY2018 was gathered. These projects to date have produced more than 280,000 publications that have referenced close to 5.40 million resources and have been cited by nearly 4.72 million authors (Table 1).

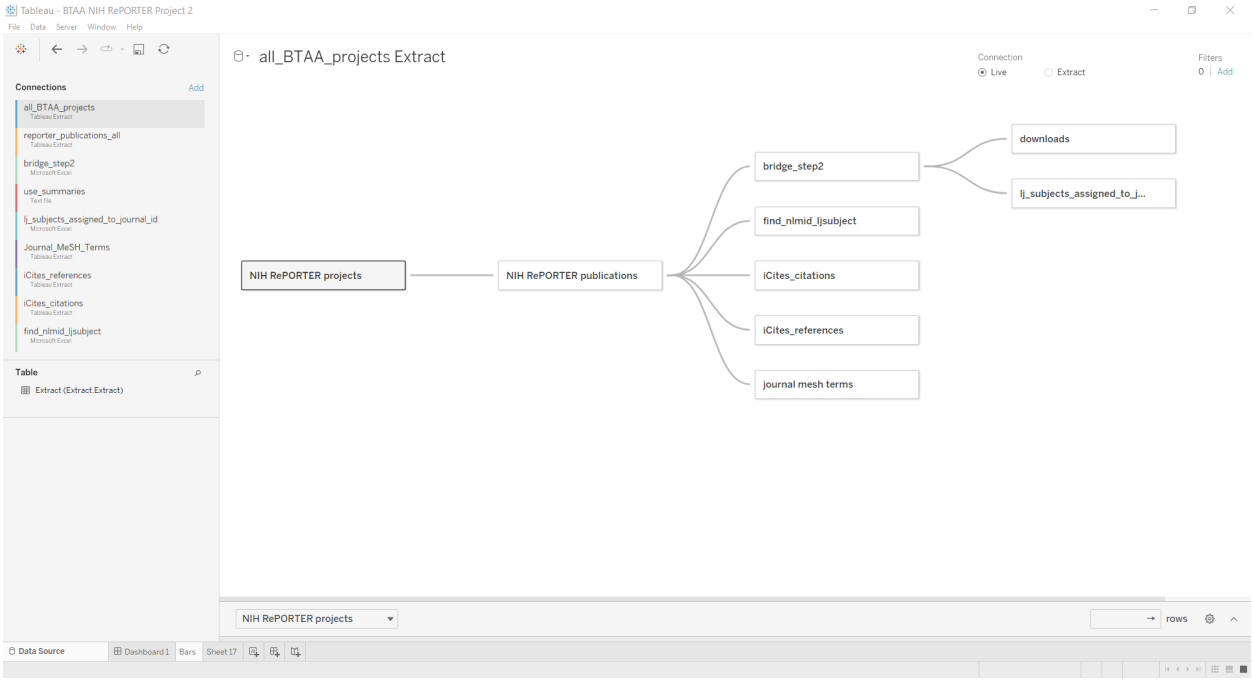
Table 1. Summary of data gathered

Category	Value
Number of NIH grants 2010–2018	18,845
Average annual value of NIH funding for each school	$193M
Range NIH funding across BTAA schools	$15M–$616M
Number of publications	284,822
Number of references 2015–2017	5,403,273
Number of citations 2015–2017	4,720,449

The data was then further enhanced by looking up the PMID for each publication and requesting a list of PMIDs referenced by the publication and PMIDs citing the publication using the iCite API. Journal title and pub year for each reference and citation PMID was next identified in PubMed using the NIH's Entrez Programming Utilities (E-utilities). MeSH terms for individual journal titles was also gathered to facilitate visualization filtering. This required that NLM IDs be added to journal titles wherever possible by matching titles on the PubMed journal list to each publication's journal title (National Library of Medicine 2021).

The export files for projects, publications, journal MeSH terms, references, and citations were last modeled in Tableau using a series of relationships (Figure 1). To model investments in library collections and calculate a ROI, a bridge file was added to map journal titles to a file containing the average 2018 cost of journal titles in each subject (Bosch, Albee, and Henderson 2018). This file was used in a related project that examined Web of Science data combined with other library operational data to visualize the value of library collections (Murphy et al. 2022). Journal usage data for 2015 to 2017 was finally added for the author's home institution to show how this data might be integrated with the other project data to inform decision-making.

Figure 1. Screenshot of Data Model in Tableau



The ROI Calculation

ROI is typically calculated by dividing net income by expenses, or net program benefits by program costs. Universities negotiate facilities and administrative (F&A), or indirect costs with the government for federally funded research. This rate is intended to support research infrastructure, including the construction and maintenance of buildings, utilities, and libraries. The 2018 F&A rate, broken down by individual components, earmarked 1.7% of total indirect costs for libraries at the author's home institution. To approximate or model the contribution of grant funding to supporting the purchase of an individual journal, a series of calculations is required to compute ROI (Table 2).

Table 2. Calculations required for ROI and other visualizations

Field	LOD Calculation
Number of grants	countd([Core Project Number])
Total grant value	sum([Total Cost]
Total indirect cost value	Sum([InDirect Cost IC])
Number of articles published by grant 2010-	{fixed [Core Project Number]: countd([Pmid])}
Number of grant articles published in this journal 2010-	countd([Pmid])
Estimated journal price	Sum([Average Cost Per Title 2018])
Adjusted library F&A rate	(countd([Pmid])/sum([annual number articles published by grant]))*.017
Library institutional cost	sum([InDirect Cost IC])*[adjusted library fa rate]

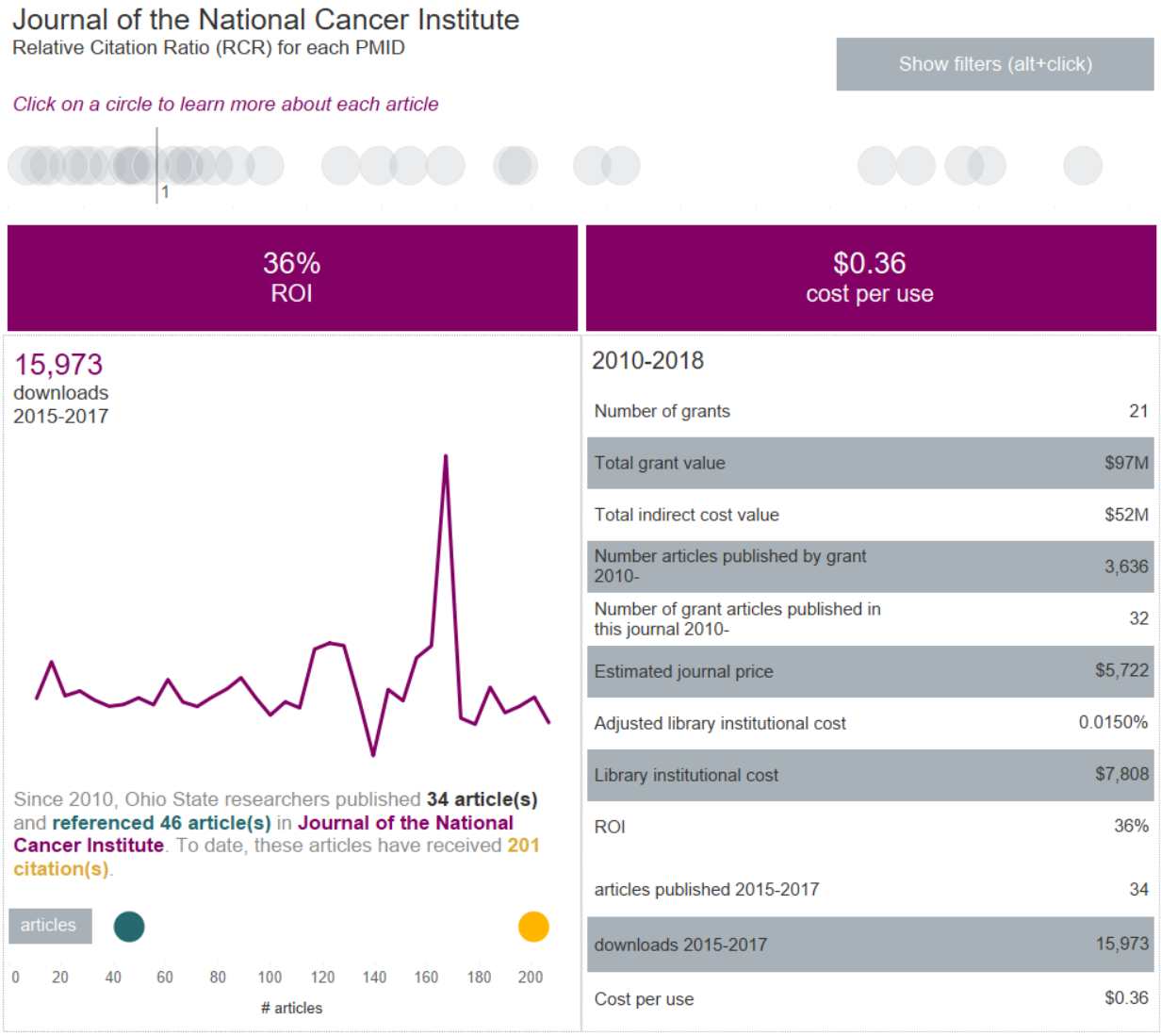
Field	LOD Calculation
ROI	([Library IC]-sum([Average Cost Per Title 2018]))/sum([Average Cost Per Title 2018])
Cost per use	min([Average Cost Per Title 2018])/sum([Downloads])

First a level of detail calculation fixes the total number of articles published by each grant to the grant's core project number. This is necessary to calculate an adjusted library F&A rate. The adjusted library F&A rate counts all articles published in one journal that were supported by NIH funding, and divides this number by the total number of articles published by the grants associated with each article. One article, for example, might have three to four NIH core grant numbers assigned, indicating that the research was supported by multiple NIH awards. This number is then multiplied by .017, or the 2018 F&A rate for the author's home institution. Thus, if the total number of articles associated with a grant, regardless of journal title being examined is 4, and only 1 of the 4 articles associated with the grant is published in the journal title being examined, the adjusted library F&A rate would be 0.4250% (1/4 * .017) This percentage is then multiplied by the total indirect cost for all grants associated with research BTAA authors published in the journal to determine the library institutional cost. To calculate the net income required for the ROI numerator, the library institutional cost serves as a proxy for revenue and the estimated journal price functions as a proxy for expenses. The net income is then divided by the estimated journal price to determine the ROI.

The Dashboard Visualizations

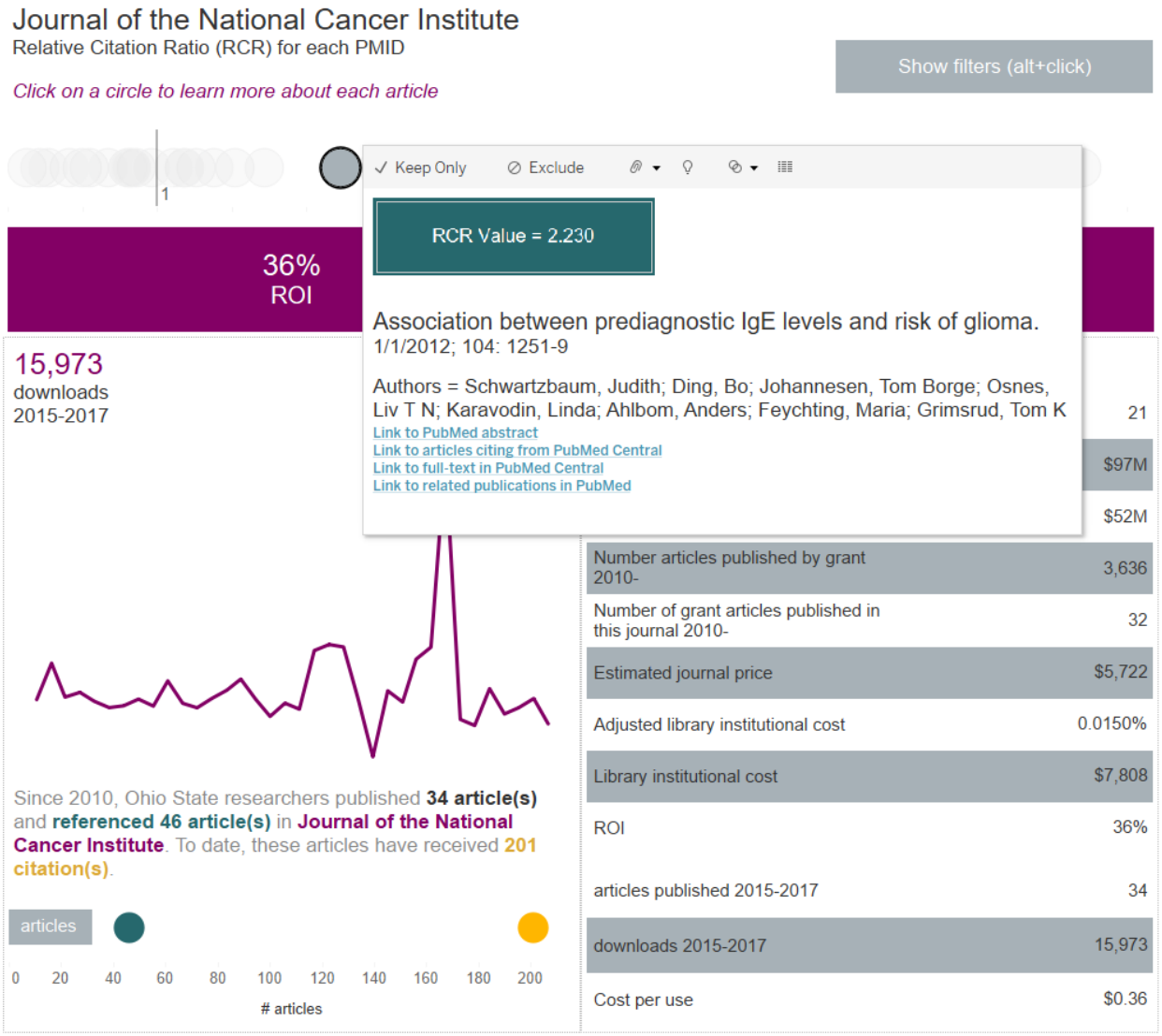
The dashboard displays of series of interactive visualizations modeling the use of, value and influence of research published in each individual journal by faculty at one or more BTAA institutions (Figure 2).

Figure 2. Dashboard for Journal of the National Cancer Institute



Hidden filters allow the user to select one or all BTAA schools, choose one of 2,020 MeSH terms, and select one journal title associated with the MeSH term. RCR values are displayed via a strip plot at the top to show the influence of individual articles published in the journal in relation to other articles within a researcher's specialty. The user may click on an individual circle to view the actual RCR value, the article title, the authors, and the citation information for the article. Clicking on an individual circle also allows the user to link out to the PubMed abstract, a list of articles citing the article from PubMed Central, the full-text of the article in PubMed Central, and a list of related publications in PubMed.

Figure 3. Dashboard for Journal of the National Cancer Institute with Tooltip Open



Call-out numbers in the two boxes immediately below the RCR values show the ROI and cost per use calculations. An interactive trend chart immediately below on the call-out numbers on the left summarizes monthly downloads of the title. Users may roll over the line to see the actual number of downloads each month, and if desired, highlight and download the summary data points for further analysis. The chart below the trend chart shows the number of articles the selected institution's researchers published in the journal each year from 2010 to 2018, the number of times these articles were cited, and the number of articles BTAA authors referenced in the journal. A text table summarizing the data assembled for the ROI and cost per use calculations is provided on the right.

Practical implications or value

Data available through freely available government tools offers libraries the opportunity to directly quantify their value. The dashboard visualizations in figure 2 are a first attempt to model and visualize the outputs of federally funded grant research with other useful library data, such as downloads, journal cost, and more to better understand academic libraries' contributions to institutional research. Libraries may better tailor collections and services to support both the institution and individual researchers, by assembling and viewing operational data with RCR values and other representations of research influence and impact.

Libraries may also further explore the data modeled for this project at both high aggregate and lower levels of detail using filters, sets, parameters, actions and other features in Tableau. Filters, when strategically applied, structure the outputs of Tableau analyses, and in this instance, may help libraries better understand how they support or could better support university research programs. Other potential projects of value using NIH RePORTER and iCites data include an exploration of researcher networks using chord diagrams or visualizations of interdisciplinary research connections using the MeSH terms assigned to each individual article published by a selected institution's researchers.

The NIH Public Access policy with the Federal Funding Accountability and Transparency Act of 2006 enabled this project. Libraries no longer must rely on self-report to quantify faculty use of library resources, but can piece together a holistic view of how faculty utilize library resources using freely available data. By visualizing aggregate level data showing what journals faculty choose to publish in, reference, and cite, with cost, influence, and impact data, libraries can demonstrate how they contribute to the research enterprise. Libraries help scholars access information needed to both develop and sustain research agendas.

—Copyright 2023 Sarah Murphy

Acknowledgments

The author would like to thank Jim Stemper for inviting her to collaborate on a research project using Tableau and Web of Science data and introducing her to Lee Konrad, Mary Schoenborn, and Stephen Meyer. This project grew from our weekly conversations, and I appreciate your patience when I went off to explore a way to visualize grant data with citation data. I also want to thank Shaun Fontanella helping me learn python and pull the reference and citation data for this project.

References

Bosch, Stephen, Barbara Albee, and Kittie Henderson. 2018. "Death by 1000 Cuts | Periodical Price Survey 2018." *Library Journal*, April 23. Accessed October 18, 2022. https://www.libraryjournal.com/story/death-1000-cuts-periodicals-price-survey-2018.

DeGroote, Sandra L., Beyza Aksu Dunya, Jung Mi Scoulas, et al. 2020. "Research Productivity and Its Relationship to Library Collections." Evidence Based Library and Information Practice 15 (4). doi:10.18438/eblip29736.

Hutchins, B. Ian, Xin Yuan, James M. Anderson, and George M. Santangelo. 2016. "Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level." *PL* 14 (9): e1002541. doi:10.1371/journal.pbio.1002541.

Kaufman, Paula T. 2008. "The Library as Strategic Investment: Results of the Illinois Return on Investment Study." Liber Quarterly 18 (3/4): 424–436. https://liberquarterly.eu/article/view/10533/12385.

Monroe-Gulick, Amanda, Lea Currie, and Travis Weller. 2014. "Local Citation Analysis of National Science Foundation and National Institutes of Health Grant Applications: Meeting the Needs of Researchers." *The Serials Librarian* 67 (4): 399–414. doi:10.1080/0361526X.2014.983291.

Murphy, Sarah Anne, Mary Schoenborn, Jim Stemper, Lee Konrad, and Stephen Meyer. 2022. "Visualizing Value of Library Collections Relative to the University Teaching and Research Enterprise: An Application of the CDL Journal Weighted Value Algorithm by Three BTAA Libraries." *Library Assessment Conference*. Association of Research Libraries. *In press*.

National Institutes of Health. 2016. *What is the NIH Public Access Policy*. September 9. Accessed October 18, 2022. https://www.nih.gov/health-information/nih-clinical-research-trials-you/what-is-nih-public-access-policy#:~:text=The%20Public%20Access%20Policy%20ensures,.gov%2Fpmc%2F).

National Library of Medicine, National Institutes of Health. 2021. *List of All Journal Cited in PubMed*. June 10. Accessed October 18, 2022. https://www.nlm.nih.gov/bsd/serfile_addedinfo.html.

Qua, Kelli, Fei Yu, Tanha Patel, et al. 2021. "Scholarly Productivity Evaluation of KL2 Scholars Using Bibliometrics and Federal Follow-on Funding: Cross Institution Study." *Journal of Medical Internet Research* 23 (9): e29239. https://www.jmir.org/2021/9/e29239.

Tenopir, Carol, Amy Love, Joseph Park, et al. 2010. *University Investment in the Library, Phase II: An International Study of the Library's Value to the Grants Process*.

White Paper, San Diego, California: Elsevier. https://roi-couperin.sciencesconf.org/data/pages/Tenopir_whitepaper_roi2.pdf.

Yu, Fei and Barrie E. Hayes. 2018. "Applying Data Analytics and Visualization to Assessing the Research Impact of the Cancer Cell Biology (CCB) Program at the University of North Carolina at Chapel Hill." *Journal of eScience Librarianship* 7 (1): e1123. doi:10.7191/jeslib.2018.1123.