Local Taxonomies Supporting Campus Analysts: Integrating Research Impact and Data Visualization Services

Sarah Anne Murphy, The Ohio State University

Sheila Craft-Morgan, The Ohio State University

Purpose & Goals: Various campus entities use bibliographic data to 1) analyze and benchmark the institution's position in relation to other universities, 2) understand collaborative networks, and 3) explore interdisciplinary connections. This data, however, is frequently not used in coordinated ways. Units independently harvest, repackage, and report it using their own procedures, definitions, and practices, leading to inconsistent results. Academic librarians are uniquely positioned to address this issue. As trainers and metadata experts, they already educate campus constituents to locate, gather, utilize, understand, and apply bibliographic tools and measures to assess research impact, both ethically and effectively. But stretched campus institutional research and research analysis units still invest a tremendous amount of time and human capital matching data collected from various public and proprietary sources to local discipline groupings, research programs, and other analysis units. Academic libraries can help these groups save time by using their data skills in concert with existing campus infrastructure to help facilitate consistent, timely reporting.

This study asks how academic libraries can engage campus analysts and use existing taxonomies with unique identifiers to develop tools that save campus planners time and improve reporting consistency across units. The study offers new ways for envisioning the application of library assessment by showing how integrated library research impact and data visualization services can successfully position the library as a partner for campus organizational performance improvement while simultaneously helping researchers make scholarly work discoverable and administrators understand the breadth, scope, and success of the university's various research programs.

Design & Methodology: Aware of an ongoing campus need to link funding data to publications, patents, and clinical trials, and filter this data using local taxonomies, the authors developed a list of use cases to share with campus planners to surface mutual goals and guide the development of tools. Because the possibilities of this type of work are infinite, the authors believed a prototype was needed to inspire discussion and focus the linking of bibliographic data to local systems. They initially developed a series of Tableau dashboards that integrate NIH RePORTER, iCites, and PubMed data. These dashboards not only raise awareness of the existence of and richness of the NIH data but show how to meaningfully model and visualize both grant and bibliographic data, as well as filter this data using local constructs.

The prototype includes federal funding data for grants awarded to the university between 2018-2022, lists of research publications associated with this data from the NIH RePORTER, corresponding reference and citation lists along with RCR values from iCites, and bibliographic information for the reference and citation lists from PubMed. To allow meaningful filtering, the data is linked to local systems using a simple join table that matches local identification numbers for each author with the research publication PMID, and the Scopus author id. Since most campus researchers do not have the requisite skills to harvest and clean author and location data from the messy affiliation field in PubMed, the join table helps campus planners to normalize faculty names, as well as university college, department, and division assignments. They can filter data using local directory information - including faculty ranks, discipline groupings, research programs, buildings, and more.

Conclusions: The prototype is now complete, and the authors are engaging various constituents across campus using the use cases and prototype. The dashboards include embedded instructions for using the views and document the sources the data was assembled from. A data dictionary is also available to facilitate a shared understanding of how each field is defined. When ready, the authors plan to mount the join table to Tableau Server and start educating campus partners on how to apply this table to their own projects to consistently filter data using meaningful local taxonomies. We anticipate additional questions concerning how research impact and data visualization services can broadly support organizational performance measurement will result from these meetings. We will also work with partners to develop future tables linking local data to other bibliographic data sources to facilitate benchmarking and comparisons.

Practical implications or value: As research impact and data visualization services evolve, librarians in these roles have an opportunity to work with campus partners to support the university's organizational performance measurement. This study builds the library's existing reputation for providing quality bibliographic research and instruction with the creation of new services that leverage librarian expertise to save researchers time and improve the quality of their analyses. It also provides an example of the creation of a cost-effective data source by harvesting public data and combining it with local datasets which can be replicated by other institutions.

Introduction

Research and development spending by academic institutions totaled $97.8 billion (about $300 per person in the US) in FY 2022.[1] Defined as the "the systems, services, policy, and staffing for the work of the university to generate knowledge" this enterprise encompasses a wide range of organizations, including the academic library.[2] Despite having a role in providing collections and services essential for supporting research, libraries' contributions to the research enterprise are often poorly understood and undervalued. One mechanism to address this issue is to develop partnerships with other essential and under-resourced units required to support university research initiatives.

Librarians provide bibliographic data and help researchers find and understand the appropriateness of scholarly research metrics available through various bibliometric databases. But librarians are not the only campus entity utilizing library-acquired and publicly available bibliographic databases to track the effectiveness of university initiatives and research programs. With a lack of uniform procedures, definitions, and practices for harvesting and cleaning data required for these analyses, inconsistent reports result. Academic librarians can effectively partner with campus units to facilitate the consistent collection of data from various public and proprietary sources and then use existing campus infrastructure to match this data to local discipline groupings, research programs, and other analysis units.

Accurate and timely bibliographic and bibliometric data resources that provide insights at the individual, unit, and institutional level are paramount given the competitive nature of research funding. This study asks how academic libraries can engage campus analysts and use existing taxonomies with unique identifiers to develop tools that save analysts time and improve reporting consistency across units. It shows how integrated library research impact and data visualization services can successfully position the library as a partner for campus organizational performance improvement while simultaneously helping researchers make scholarly work discoverable and administrators understand the breadth, scope, and success of the university's various research programs. Partnerships linking proprietary and public data to local units of analysis may save both librarians and campus analysts time and resources and lead to more effective, meaningful, and consistent reporting.

Literature Review

The type of bibliometric services academic institutions provide varies as well as where the service is situated within the institution. Gutzman et al. investigated the type (informal vs formal) and scale (small, moderate, and robust) of research evaluation support services at seven biomedical libraries in U.S. and Canada finding great variation in the type of services offered, driven by the unique needs of the libraries' constituents.[3] Taking into consideration the availability of expertise in bibliometrics, the level of an institution's investment in bibliometric tools and services, and governance structure of the bibliometric services, Bredahl suggested that the structure of bibliometric services typically falls into one of three representative service

models: collaborative bibliometric services, centralized bibliometric services in the library, and centralized bibliometric services outside of the library.[4]

Bredahl also noted that as bibliometric services mature, the central focus of the service may shift from the individual to the organization, accompanied by a growth of collaborative partnerships across the institution. However, Bredahl cautioned that sustainable relationships between various units, not dependent on an individual, are imperative to establishing collaborative partnerships. Bryant, Dortmund, and Lavoie refer to this as social interoperability.[5] In the area of research analytics, Bryant et al. noted that collaborative partners, such as research administration, academic affairs, communications, IT, with the library as a valued partner, have a growing interest in decision support tools to provide a more "nuanced view of an institution's research strengths, weaknesses, and networks of opportunity."[6]

A few libraries now offer integrated library research impact and data visualization services to support institutional research and analysis teams. Mani, Hayes, Dodd, and Yu and Mani, Cawley, Dodd, and Hayes provide a detailed outline of the University of North Carolina at Chapel Hill's work to partner with various campus units on bibliometrics and information visualization projects, with efforts focused on highlighting the impact of federally funded research programs.[7] Yu and Hayes share the outcomes of a specific research impact project resulting from a partnership with the Cancer Cell Biology Program at the University of North Carolina's Lineberger Comprehensive Cancer Center.[8] The authors assembled citation data from Scopus and the NIH iCite database, analyzed the data and then utilized Tableau and VOSviewer to visualize scholarly activity, co-author, internal, and international collaborative networks, and program researchers' areas of interest. Yu, Van, Patel et al. and Yu, Patel, Carnegie et al. also used bibliometrics and data visualization services to examine the impact of UNC Chapel Hill's Clinical Translational Science Awards program [9] Their projects examined the research output of faculty affiliated with the university's CTSA program, the impact of faculty research using average Scopus Field-Weighted Citation Impact (FWCI) values and average NIH Relative Citation Ratio (RCR) values, researcher co-author networks using network diagrams, and the co-occurrence of Medical Subject Heading (MeSH) terms in select publications addressing health disparities.

This project leverages publicly available data alongside local taxonomies to develop dashboard prototypes showcasing how integrated library research impact and data visualization services can effectively position the library as a key partner in enhancing campus organizational performance at The Ohio State University. Dashboard prototype measures include RCR values, CiteScore, the open-access status of the publication, and the type of NIH funding source.

Design, methodology, or approach

Recognizing an ongoing need to connect funding data with publications, patents, and clinical trials, and to filter this information using local taxonomies, the authors created a list of use cases to share with campus analysts. This was intended to highlight common goals and guide tool development. Given the vast potential of this work, the authors believed a prototype was

essential to spark discussion and focus efforts on linking bibliographic data to local systems. The authors initially developed a series of Tableau dashboards that integrate data from NIH RePORTER, iCites, Scopus, and PubMed. These dashboards not only highlight the richness of NIH data but also demonstrate how to effectively model and visualize both grant and bibliographic data and filter this information using local constructs.

A multifaceted approach was adopted to guide stakeholder conversations and surface mutual goals. With library use cases for bibliographic data tending to focus at a high aggregate level on what journals campus researchers use, each use case written for this project outlined how the assembled bibliographic and local data could support a specific academic unit's assessment needs by

- providing data that illustrates a unit's goal to **increase collaboration** locally and to develop national and international collaborative networks;
- providing evidence of **research productivity** and the relationship of grant funding to research publications; and
- supplying data that supports [the unit's] strategic plan around **interdisciplinary research**.

Because the possibilities of this type of work are infinite, the authors next developed a companion series of Tableau dashboard prototypes to illustrate the concepts presented in the use cases and demonstrate the kinds of questions bibliographic data harvested from various public and proprietary resources answers when paired with local data. The authors simultaneously developed a data dictionary defining each field in the data model, the data type of the field, and the data source.

Finally, the authors scheduled conversations with potential stakeholders to explore partnerships for integrating library research impact and data visualization services into campus data initiatives. To support these conversations and describe the wealth of bibliographic data available to support unit assessment needs, the authors shared

- the data dictionary;
- a detailed PowerPoint presentation including the use cases and a diagram showing table relationships for the data model; and
- the prototype dashboards.

The authors developed questions to guide the conversations:

- What are your thoughts overall?
- Do you think it would be useful to potential users? In what context?
- What additional information would you want to see?
- Do you have any feedback about how the dashboard is organized?

The dashboards integrated NIH RePORTER, iCites, and PubMed data. The NIH RePORTER data included a list of all NIH funds awarded to Ohio State researchers between FY 2010 and

FY 2022, and publications resulting from these awards published between 2018 and 2022. The larger number of award years was selected to match more publications to award details and allow future project to benchmark time from award to publication. The PMID numbers included with the NIH RePORTER publications were then searched with the iCite API to gather the Relative Citation Ratio (RCR) value for each article and gather a list of PMIDs for the papers referenced by each article and the papers citing each article.

The PMIDs from the publications, reference papers, and citing papers lists were then gathered, deduplicated, and loaded into a Python script to gather the article title, journal title, journal volume, journal issue, publication year, authors, and author affiliations. Author affiliations were processed to isolate institutions and, wherever possible, match Ohio State academic unit and division. When institution or department affiliation was unclear, the PMID was then used to locate Scopus author identification and affiliation identification numbers. These lists were then merged and reconciled with local data to create a master list of Ohio State University authors and their collaborators. The 2023 CiteScore and open access status for each journal on the original NIH RePORTER publication list was also gathered using the Scopus API and joined back to the list.

The data was last prepared in Tableau Prep to assign start and end nodes required for select visualizations and parse the publication, grant, and citing/referenced paper data into separate tables to improve dashboard efficiency.

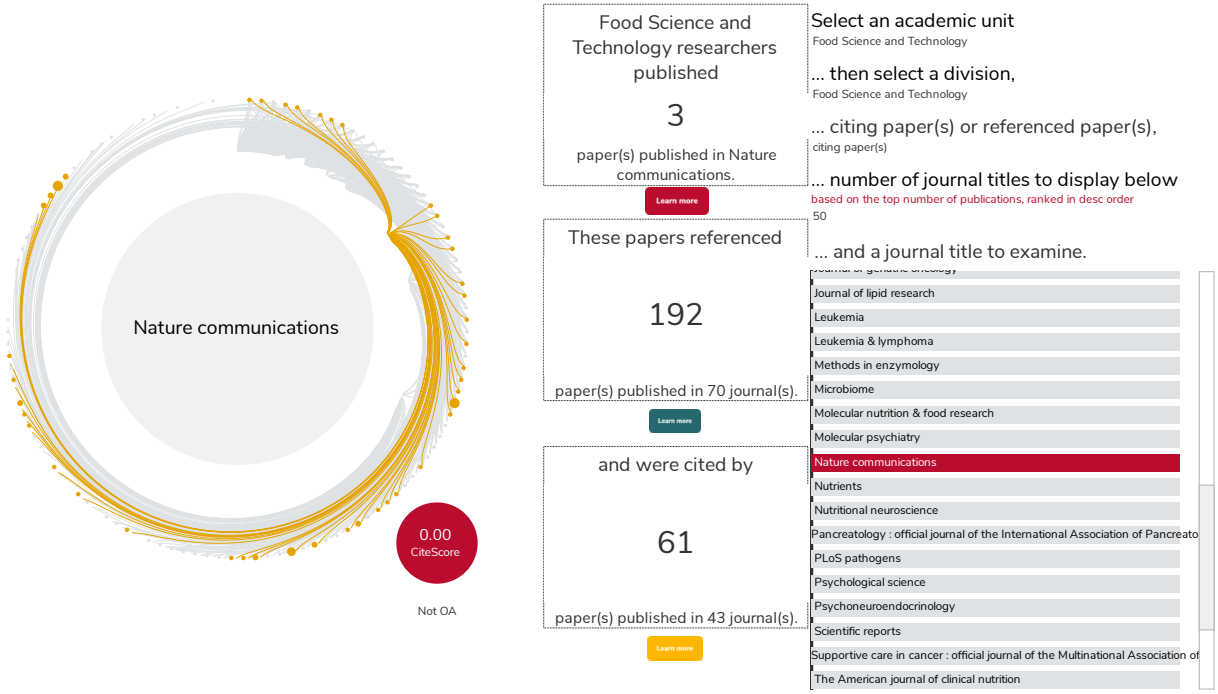
Findings

The authors presented the use cases and dashboards to three potential users to explore partnerships and engage in informal conversations about the project. Feedback from these individuals, comprising both librarians and staff, was highly valued due to their familiarity with the source data, experience in creating and using dashboards, and understanding of the bibliographic data needs at the institution. After demonstrating the prototype, the participants provided valuable feedback on the dashboard and expressed support for the project's concept. They noted that the current prototype can meet the data needs of bibliographic data users at the institution. The participants confirmed the use cases identified by the authors, highlighting that university leadership is asking complex questions of bibliographic data users. For instance, academic leaders are requesting data that describes the volume of scholarship produced by researchers at the institution and its impact on their respective fields. Additionally, data that can be included in grant applications, particularly training grants, is a priority. Currently, there is no institution-wide research information management system that collects this information and enables users to respond to data requests and prepare regular reports.

The participants also discussed the potential for cross-campus collaboration in the creation and deployment of the dashboard, listing units and individuals identified by the authors and new partners who could potentially propose additional use cases. Also, the participants shared that a higher level of collaboration in the form of sponsorship by administrators or thought leaders on campus can influence others to adopt this type of data project for regular use. Nevertheless,

the participants noted that to fully meet the needs of potential users, the authors, and their potential partners, will need to implement training during the deployment phase of the project. Incorporating training and education as the project is disseminated will be essential as users may be unfamiliar with terminology. Finally, given the potential for use of the prototype to be responsive to various use cases by different users at various levels of the organization, the participants suggested usability testing to help the authors adapt the tool to different audiences.

Figure 1. Dashboard created for academic librarians



The dashboard prototypes demonstrated how to show bibliographic data at various levels of granularity. The first dashboard prototype aggregated data by journal title and selected filters (Figure 1). Users change the filters in the right column by selecting:

- an academic unit;
- a division within the academic unit, if appropriate;
- citing paper(s) or referenced paper(s);
- the number of journals to display in the visualizations; and
- one journal title to examine.

Filters are nested so that once an academic unit is selected, only divisions affiliated with the unit will display, and only journals publishing articles authored by OSU researchers affiliated with the academic unit may be examined.

The current example displays data for Food Science and Technology researchers who published articles in *Nature Communications* between 2018 and 2022. The circular Sankey

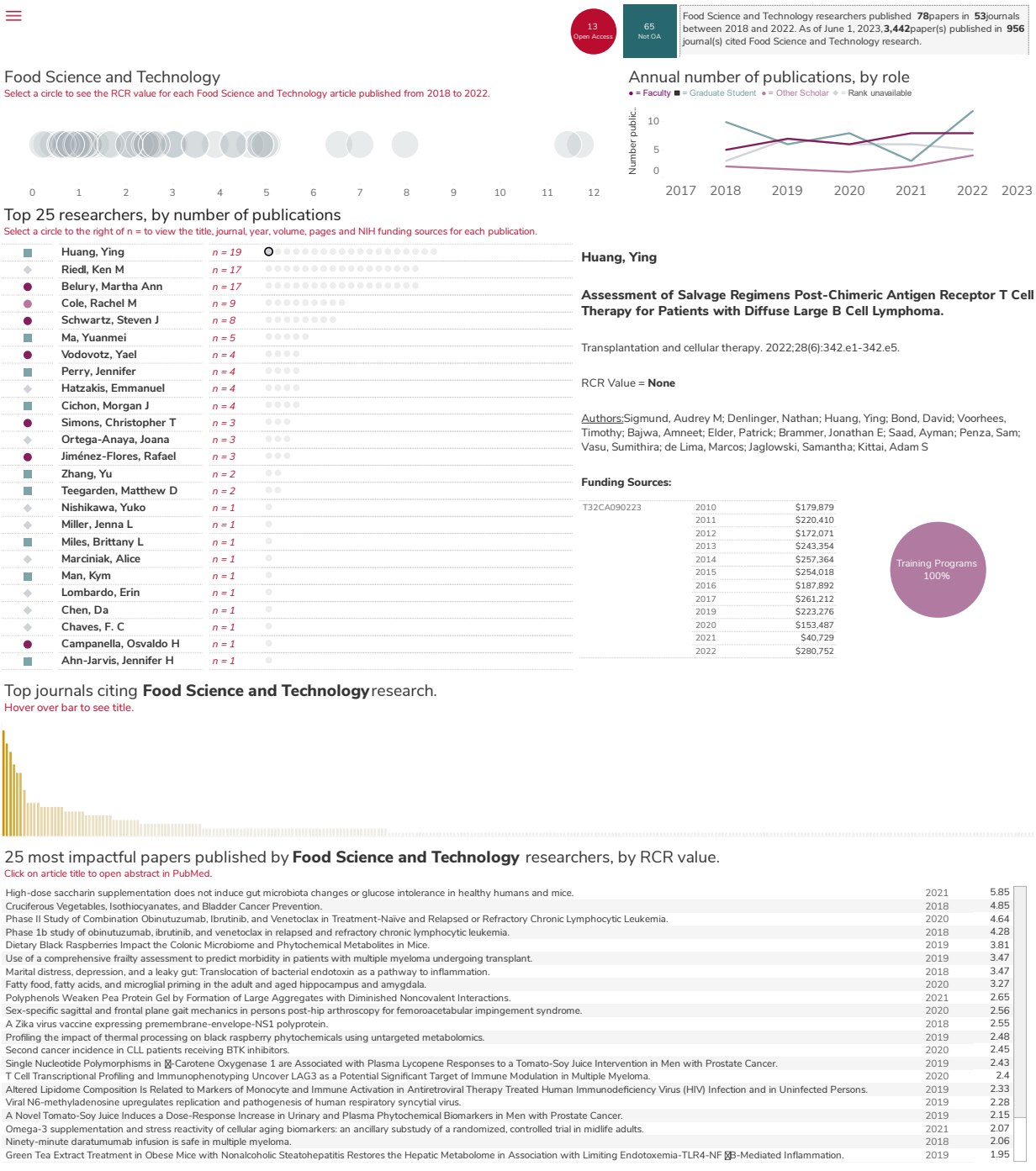
diagram on the left of the dashboard links *Nature Communications* to the journals publishing articles citing this research. The size of the yellow lines and circles represents the number of citing paper(s) in each journal. Users can hover over the circle to view the citing paper(s) journal title, and the total number of citing paper(s).

Call out number on the right summarize the number of articles published by Food Science and Technology researchers in *Nature Communications*, the number of paper(s) Food Science and Technology researchers referenced in these articles, and the number of paper(s) citing these articles. The Scopus CiteScore, if available, is also provided in the visualization, and the publication journal's open access status.

While the circular Sankey diagram is effective for showcasing the complexity of an academic unit's information needs, the use cases and stakeholder conversations confirmed that figure 1 – which summarizes bibliographic data at the journal level – may have limited utility outside of an academic library environment. Campus analysts must report data at a more granular level to confirm the effectiveness of various unit and division strategies, programs, and more.

Figure 2, using the same data, provides a more granular analysis of Food Science and Technology researcher's publication activity. A narrative box appears on the top right summarizing the overall number of papers supported by NIH funded published by Food Science and Technology researchers between 2018 and 2022, the number of paper(s) referenced, and the number of paper(s) citing this research as of January 21, 2023. An interactive strip plot on the top left of the dashboard, then shows the range of RCR values assigned to each paper. Each overlapping circle represents one article and selecting a circle reveals the citation and assigned RCR value.

Figure 2. Dashboard created for campus analysts



Immediately below the strip plot is another interactive table showing the top 25 researchers within the department by number of publications. Each dot to the right of the researcher's name represents one publication. Clicking on a dot opens a window to the right of the table that includes the citation, the funding sources for the article, and a summary of the funding sources broken down by NIH activity code. This allows analysts to see if the research was supported by training grants, research grants, or other NIH funding categories. The top journals citing the academic unit's NIH funded research are then listed in an interactive bar chart and a

table showing the top 25 most impactful papers published by the unit's researchers is provided. Each paper is linked to a PubMed abstract and impact is determined by the individual paper's RCR value.

Conclusions

Discussions about linking bibliographic data to local constructs can be challenging because of the inherent complexity of the data and the vast possibilities this type of work presents. Equipped with relevant use cases and dashboard prototypes, libraries can effectively foster partnerships with campus analysts, achieving mutual goals while applying bibliographic tools and measures to assess research impact ethically and effectively. This project demonstrated how integrated research impact and data visualization services can respond to the needs of campus data analysts and share their expertise, skills, and experience to facilitate consistent harvesting, cleaning, packaging, and reporting of bibliographic data.

Using data from NIH RePORTER, iCites, PubMed, Scopus, and local systems, the project demonstrated to campus analysts how they could visualize and explore publications resulting from federal funding at various levels of granularity, using local taxonomies to filter by academic unit and more. The synergistic collaboration between the authors led to the successful design of the prototypes and the development of use cases to focus and guide conversations with potential partners. Inspired by the prototypes and use cases, potential stakeholders shared additional ways the data could be useful for their own needs, confirming the effectiveness of the authors' approach. These conversations also helped identify other potential partnerships, including access to a local data source that will facilitate author authentication, further supporting filtering using local taxonomies.

Throughout the project, the authors developed new skills and gained a deeper understanding of using bibliographic data from publicly available resources and connecting it to local data structures. The prototype dashboards include embedded instructions for using the views and document the sources of the assembled data. A data dictionary is also available to ensure a shared understanding of each field's definition. The authors plan to mount the join table to Tableau Server and start educating campus partners on how to apply this table to their own projects, enabling consistent data filtering using meaningful local taxonomies.

We anticipate that future stakeholder meetings will generate additional questions about how research impact and data visualization services can broadly support organizational performance measurement. While this project initially focused on NIH data, the authors will next integrate data from other sources, including NSF. We will also collaborate with partners to develop future tables that link local data to other bibliographic data sources, facilitating benchmarking and comparisons.

Practical implications or value

Librarians specializing in data visualization and research impact can leverage their existing expertise to develop new bibliographic data services in collaboration with campus partners. This study introduces a prototype designed to save campus analysts time and effort by providing access to public data integrated with local datasets.

[1] Michael T. Gibbons, National Center for Science and Engineering Statistics (NCSES). 2023. R&D Expenditures at U.S. Universities Increased by $8 Billion in FY 2022. NSF 24-307. Alexandria, VA: National Science Foundation, 2023, available at https://ncses.nsf.gov/pubs/nsf24307.
[2] Jane Radecki, Oya Y. Rieger, and Roger C. Schonfeld, "What is the Academic Research Enterprise?" Ithaka S&R. https://sr.ithaka.org/blog/what-is-the-academic-research-enterprise/, accessed 10/11/2024.
[3] Karen Elizabeth Gutzman, Michael E. Bales, Christopher W. Belter, et al., "Research Evaluation Support Services in Biomedical Libraries," Journal of the Medical Library Association 106, no. 1 (2018): https://doi.org/10.5195/jmla.2018.205.
[4] Laura M. Bredahl, "The Current and Evolving Landscape of Bibliometric Tools and Technologies," Library Technology Reports 58, no. 8 (2022): 1-36, https://doi.org/10.5860/ltr.58n8
[5] Rebecca Bryant, Annette Dortmund, and Brian Lavoie, "Social Interoperability in Research Support: Cross-Campus Partnerships and the University Research Enterprise," 2020, https://www.oclc.org/content/dam/research/publications/2020/oclcresearch-social-interoperability-research-support.pdf.
[6] Ibid, p.21.
[7] Nandita S. Mani, Barrie E. Hayes, Adam Dodd, and Fei Yu, "New Data-Related Roles for Librarians: Using Bibliometric Analysis and Visualization to Increase Visibility of Research Impact," in *Handbook of Research on Knowledge and Organization Systems in Library and Information Science*, ed. Barbara Jane Holland (Hershey, PA: Information Science Reference, an imprint of IGI Global, 2021), 317-345, https://doi.org/10.4018/978-1-7998-7258-0.ch017; Nandita S. Mani, Michelle Cawley, Adam Dodd, and Barrie E. Hayes, "Applying Bibliometrics to Examine Research Output and Highlight Collaboration," in *Handbook of Research on Academic Libraries as Partners in Data Science Ecosystems*, ed. Nandita S. Mani and Michelle Cawley (Hershey, PA: IGI Global, Information Science Reference, an imprint of IGI Global, 2022), 75-101, https://doi.org/10.4018/978-1-7998-9702-6.ch005.
[8] Fei Yu and Barrie E. Hayes, "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, no.1 (2018): e1123, https://doi.org/10.7191/jeslib.2018.1123.
[9] Fei Yu, Allison Alicia Van, Tanha Patel, et al., "Bibliometrics Approach to Evaluating the Research Impact of CTSAs: A Pilot Study," Journal of Clinical and Translational Science 4 (2020): 336-344, doi:10.1017/cts.2020.29; Fei Yu, Tanha Patel, Andrea Carnegie, and Gaurav Dave, "Evaluating the Impact of a CTSA Program from 2008 to 2021 through Bibliometrics, Social Network Analysis, and Altmetrics," Journal of Clinical and Translational Science e44 (2022): 1-10, https://doi.org/10.1017/cts.2022.530.