Aabha Pandit, Heather Charlotte Owen, and Alois Romanowski

Meeting Secondary Data Needs through an Open Data Internship

The Core10 Data Collection

Data literacy has emerged as a critical topic in education and research within the past decade.[1] Basic data skills are taught at precollegiate levels to prepare students for increasing data needs, even outside of computer science and mathematics classrooms.[2] Despite teaching these skills earlier than ever before to meet increasing demand, evidence shows the US educational system is not meeting the demands of the data age.[3]

This isn't a single point of failure. Accelerating data education to meet demand is easier said than done. There are significant hurdles in teaching data literacy: It is poorly defined, rarely standardized, and thus difficult to assess and measure.[4] The ambiguity of what constitutes data literacy can lead to some skills falling through the cracks. One of these skills is the discovery of secondary data sources—data created by someone other than the user. Secondary data are extremely important, as researchers often need to find multiple data types as well as data from other disciplines, and available datasets are often very specialized.[5] The authors have observed that educators assume users are able to easily discover and reuse other shared data sets.

Discovering secondary data is more difficult than it seems. In a 2020 study, more than 1,500 researchers surveyed across 105 countries reported that finding secondary data was challenging or outright difficult; 33 percent of all respondents had trouble with searching.[6] Of note, the respondents to this survey were not students—they were seasoned, published researchers. They pointed to a lack of training in discovery and search techniques and also to the dispersed nature of available datasets across the Internet, making it difficult to find a suitable dataset. Search success can often come down to luck.[7] A heavy investment in open data resources would assist in data discovery.

Open data is "research data that is freely available on the internet permitting any user to download, copy, analyze, re-process, pass to software or use for any other purpose without financial, legal or technical barriers other than those inseparable from gaining access to the internet itself."[8] Various governments, research agencies, and journals across the world have noted the importance of open data when it comes to reproducing research. As open data policies become more common, an organically grown open data ecosystem may arise.

Aabha Pandit is research assistant – data services at the University of Rochester, email: apandit@u.rochester.edu, ORCID: https://orcid.org/0009-0001-2052-7235
Heather Charlotte Owen is data librarian at the University of Rochester, email: howen@library.rochester.edu, ORCID: https://orcid.org/0009-0001-1771-366X
Alois Romanowski is STEM liaison librarian at the University of Rochester, email: arjay.romanowski@rochester.edu, ORCID: https://orcid.org/0000-0002-6078-368X.

However, until a robust infrastructure eases the discovery of secondary open data, patrons will still call for easy access to datasets.

The authors experienced these systemic issues firsthand. University of Rochester (UR) librarians receive frequent requests from students seeking help with finding data for coursework. Student consultations span multiple classes and disciplines and require considerable amounts of time and effort from librarians. Our team envisioned addressing these issues with a curated data collection featuring open, multidisciplinary datasets. This led to the proposal of the Core10 Data Collection, an open data collection designed to meet community needs and operated through a summer experiential learning internship.

Experiential Learning Internship

The UR Libraries Summer Internship Program is a paid, ten-week, full-time experiential learning internship first piloted in 2023 and currently on its third iteration. The Core10 internship took place during the 2024 internship program.[9] It was conceived as a joint venture by library leadership and the student association government to fulfill the experiential learning objective of the university's strategic plan.[10] The student association officers had noticed a rising need for job opportunities aimed at second- and third-year students—the group most in need of developing their basic job skills.

The internships have two major focuses: each position's unique project and the cohort experience. Each cohort of seven to ten students is taught so-called "soft" job skills, such as resume building, leadership exercises, crafting emails, and basic workplace skills as a unified group by specialists across campus. This approach fostered camaraderie within the intern cohort, and it helped acclimate students to working full time in a professional setting.

Over the past two years, twenty-four interns have participated in the internship, and we have received almost 300 applications. Each cohort has a new set of project-based positions designed by library staff, who also manage the interns. Most of the positions have been for user experience design, instructional/curricular design, acquiring assessment data, and extended reality development. The Core10 Data Collection position was unusual for having a heavy data focus and received the most applicants of the 2024 cohort, drawing twenty-six total applicants. This speaks to an emerging number of data professionals that libraries can tap into to fulfill data projects, matched with undergraduate students' heightened desire for meaningful and unique data opportunities.

Heather Owen and Arjay Romanowski served as librarian supervisors in 2024, and they hired an undergraduate data science student, Aabha Pandit, as the experiential learning intern. The librarians were in charge of writing the project proposal and plan, the hiring process, and supervising Pandit, who received training in topics such as data literacy and ethics, needs assessment, institutional repositories, and project management. After analyzing the basic project plan, Pandit had autonomy to develop a workflow and present it to the librarians for approval.

Core10 Workflow

The Core10 Data Collection is a curated list of ten diverse datasets available in our institutional repository, the University of Rochester Research Repository (URRR).[11] Designed with students and faculty in mind, Core10 supports research-based coursework,

independent learning, and skill development in areas like data analysis, visualization, and machine learning. The collection aims to ease access to well-documented, high-quality datasets across a range of subjects.

The creation of Core10 was an iterative, collaborative effort involving students, faculty, and library staff. The intern began with a needs assessment starting by interviewing subject librarians to better understand data-related questions and student requests within their disciplines. These conversations gave key insights into the kinds of data students typically seek in addition to existing data resources, such as a subject-specific LibGuide. Next, the intern reached out to faculty whose courses already included data-related assignments. Faculty provided valuable feedback on the characteristics of an "ideal" dataset for their courses, including considerations such as file format, subject relevance, and clarity of documentation. Concurrently, the intern surveyed students about their data needs and preferences. Students expressed interest in datasets that were relevant to health, economics, machine learning, and social sciences. They also shared challenges in finding clean, publicly available data suitable for course projects and their own exploratory work.

Based on the identified community needs, we moved to the task of finding and selecting ideal datasets. Because Core10 was intended to serve a broad audience—including beginner-level data enthusiasts, students in general courses, and faculty seeking versatile teaching materials—our focus was on identifying open datasets that were both accessible and adaptable for a variety of educational purposes. This would distinguish Core10 from other available library resources.

Finding the ideal datasets for the Core10 Data Collection was a thoughtful and extensive process. The intern focused on publicly available datasets from sources such as government agencies and research institutions (e.g., NASA, data.gov). Although these platforms offered many options, much of the data were either outdated or highly specific to research projects. We also explored librarian-curated databases on our library's website, but these were often too specialized to meet Core10's goal of broad usability.

The key challenge was sorting through datasets to identify those with educational potential—data that could be understood by non-experts yet still offer enough complexity to support skill development. The intern found suitable options through online searches of the university's collection and outside sources. Each dataset was individually evaluated based on completeness (little/no missing data), size, subject area (no highly specific research data), trustworthiness, and application areas such as data cleaning, statistical modeling, mapping, or machine learning. The chosen datasets span a range of topics and disciplines, including public health (e.g., cancer, zoonotic diseases, COVID-19 behaviors), economics and workforce trends (e.g., workers' income levels, business revenue growth, postgraduation employment), education and technology (e.g., modeling user knowledge), and popular data science examples (e.g., Titanic, restaurant reviews, meteorite landings).

The intern developed ReadMe files for each of the selected final ten datasets (Figure 1), providing clear metadata and documentation to help users understand the contents, context, and structure of each dataset. The completed collection was published in our institutional repository and made available for coursework and research.

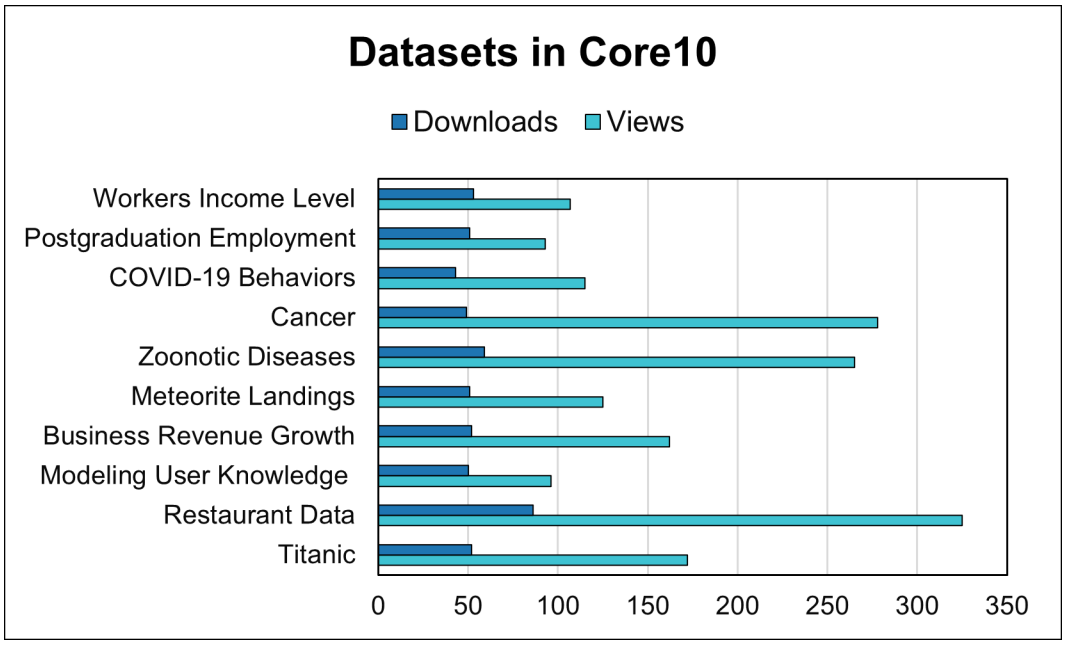

Figure 1: Downloads and view counts of different topics and disciplines covered by the Core10 Collection. Data collected January 2025.[12]

Next Steps

The Core10 collection was created to assist students and learners with finding datasets for classroom assignments and practice. It is important, therefore, that the collection not remain stagnant as it can quickly become obsolete. As the Core10 Collection was a valuable learning activity for a data science student intern, it is fitting and imperative that student employees remain involved in its marketing, upkeep, and growth. The library data services team will hire student employees outside of the experiential learning internship to assist with a wide variety of tasks, including Core10 maintenance and growth. Pandit will develop further documentation (e.g., a formalized data evaluation rubric, favored sources) to ease the transition of the Core10 Collection to future student workers.

Our first goal is to increase marketing for the collection and to encourage the use of datasets in classroom activities. We created a wordmark for the Core10 Data Collection (Figure 2), and we plan to create flyers and other advertising materials to allow us to market it via social media, newsletters, and electronic screens. Pandit is also reaching out to faculty members to advertise the Core10 Collection and discuss strategies for embedding it into curricula. Library staff regularly use Core10 datasets in workshops they design, which also provides opportunities to advertise the collection to the university community.

Our second goal is to continue to add datasets to the Core10 collection. To ensure the Core10 collection continues to be relevant and up-to-date into the future, we will assign a data services student employee to maintain it. We will also create an advertising strategy each semester and will add datasets to the collection as needed.

Finally, after practicing data analysis skills on the open data provided, it is imperative that students develop the ability to discover their own secondary data. As we mentioned, finding secondary data is a skill most students are not taught in the classroom, yet it is crucial for students to learn. Teaching these skills early in a student's career will prove beneficial

Figure 2: Wordmark developed for the Core10 Data Collection.

as students continue their education or enter the workforce, especially as sharing open data becomes the norm within research. Therefore, our third goal is to create instructional materials that will assist students in developing these skills.

Conclusion

Discovering secondary data is a challenge for both students and researchers. The difficulty in finding these resources led the library to launch an open data set project as a stopgap measure to ease data discoverability. A preexisting experiential student internship was an ideal choice for this project, allowing us to hire a student who could effectively tap into the community needs while simultaneously developing unique skills in their field. The collection was then designed with an emphasis on open data for ease of use. The Core10 Data Collection still requires maintenance, expansion to meet future needs, effective marketing, and embedding in coursework and library events to reach its full potential.

Notes

1. Bahareh Ghodoosi, Tracey West, Qinyi Li, Geraldine Torrisi-Steele, and Sharmistha Dey, "A Systematic Literature Review of Data Literacy Education," *Journal of Business & Finance Librarianship* 28, no. 2 (2023): 112–27, https://doi.org/10.1080/08963568.2023.2171552.

2. Joshua Rosenberg, Elizabeth H. Schultheis, Melissa K. Kjelvik, Aaron Reedy, and Omiya Sultana, "Big Data, Big Changes? The Technologies and Sources of Data Used in Science Classrooms," *British Journal of Educational Technology* 53 (2022): 1179–201, https://doi.org/10.1111/bjet.13245.

3. Forester Research Inc., *The Great Data Literacy Gap: Demand for Data Skills Exceeds Supply* (Forrester Research Inc., 2021), https://www.tableau.com/sites/default/files/2021-06/Tableau_Data_Literacy_Report.pdf.

4. Jeonghun Kim, Lingzi Hong, and Sarah Evans, "Toward Measuring Data Literacy for Higher Education: Developing and Validating a Data Literacy Self-Efficacy Scale," *Journal of the Association for Information Science and Technology* 75, no. 8 (2024): 916–31, https://doi.org/10.1002/asi.24934.

5. Kathleen Gregory, Paul Groth, Andrea Scharnhorst, and Sally Wyatt, "Lost or Found? Discovering Data Needed for Research," *Harvard Data Science Review* 2, no. 2 (Spring 2020), https://doi.org/10.1162/99608f92.e38165eb.

6. Gregory et al., "Lost or Found?".

7. Kathleen Gregory, Siri Jodha Khalsa, William K. Michener, Fotis E. Psomopoulos, Anita de Waard, and Mingang Wu, "Eleven Quick Tips for Finding Research Data," *PLoS Computational Biology* 14, no. 4 (2018): e1006038, https://doi.org/10.1371/journal.pcbi.1006038.

8. "Open Data," SPARC, accessed July 11, 2025, https://sparcopen.org/open-data/.

9. Emmely Eli Texcucano, "Summer Interns Take on the Library," July 8, 2024, https://www.library.rochester.edu/about/news/summer-interns-take-library.

10. "Boundless Possibility: 2030 Strategic Plan," University of Rochester, accessed July 15, 2015, https://boundless.rochester.edu/strategic-plan/education/.

11. Aabha Pandit, "Core10 Data Collection," University of Rochester, April 15, 2025, https://doi.org/10.60593/ur.d.c.7384456.

12. Aabha Pandit, Heather Owen, and Alois Romanowski. "Data Discovery Meets Experiential Learning: The Core10 Project" (poster, RDAP Association Virtual Summit, March 11–13, 2025), https://doi.org/10.60593/ur.d.28780778.