Ticker: The Academic Business Librarianship Review, 2:2 (2017)

THOUGHT PIECE

Buying the Haystack: New Roles for Academic Business Libraries

MEG TRAUNER
Duke University, Durham, North Carolina, United States

For decades academic libraries have been buying databases to support classroom instruction and faculty research. The library market for these resources is well established, with specialized salespeople who know what libraries want and understand how to deliver the information. Data resources are made widely available for use, usually to the whole campus. Librarians customarily purchase database subscriptions with the intention of renewing them annually for perpetuity. Examples of web-based business databases in the academic market include Mergent Online and Capital IQ. Some, such as CRSP or TAQ, use platforms like WRDS to make their products available to libraries. The largest information producers use a mixed approach. For example, Thomson Reuters and Standard & Poor's both sell an array of information products across platforms.

Academic decision-makers have commonly considered the library's information purchases to be community resources. Librarians purchased data that were widely in demand and readily available to a wide cross-section of users and declined to purchase resources that were available to only a small group or an individual. Even when influential faculty researchers requested specific databases, in most cases librarians made purchase decisions based on overall demand for the resource. Assistance with using specialized data was limited.

But now user needs for data are changing, and use of libraries' web-based and WRDS-accessible databases is declining. With increasing frequency, faculty members are requesting a different kind of resource—stand-alone data sets that are not widely available to the library market and not available through WRDS. The seller often withholds university-wide use, and in many cases is not set up to offer it.

Librarians who recognize the change in faculty data needs and want to stay relevant in the university research ecosystem are now purchasing and licensing these unique data sets to make them available to individual researchers or small groups of faculty. These stand-alone data sets are usually custom extracts, often one-time purchases of specific data that an individual or small group of faculty members needs for their research. These data sets involve a number of licensing, hosting, curating, and funding issues that librarians must resolve if they want to serve their institutional research agenda.

What: Data Sets Defined

While there are commonalities, data set products are tailored to the faculty member's research and, as a group, they are difficult to define or even to name. While characteristics vary, I will refer to these as "one-off data sets." The name "one-off" implies that data sets are purchased only once, or perhaps once every several years. While one-time purchases are common, there are exceptions.

Meg Trauner is the Associate Dean for Library Services at the Ford Library at the Duke University Fuqua School of Business, mtrauner@duke.edu.

A one-off data set is usually, but not always, a large chunk of data. The largest such request at my institution, Duke University, was from a business strategy professor who needed the Web of Science Core Collection in XML format, from 1900-2016—60 million metadata records. Size of data needs vary, however; our smallest request last year was for a 50MB file of four years of Blue Chip Financial Forecasts.

This type of data set usually satisfies a specific research need. The typical faculty member requesting one-off data has planned a project and is working with co-authors at other universities. There are few, if any, alternatives for the requested data.

Sometimes a data set needed for faculty research is only available from a commercial vendor that is not in the library market. PATSTAT (global patent data) and SNL Financial (banking and financial services data) are recent examples. Sometimes the vendor is well represented in the library market, but its web-accessible data are not delineated finely enough to satisfy the research agenda of a faculty member. For example, data from Bureau van Dijk is available through a web interface and through WRDS, yet Duke faculty researchers seeking ownership and financial data from 2002-2012 requested an extract that amounted to over a terabyte of data. The custom extract needed to be created by library staff because faculty members were unable to efficiently extract their needed data points using the BVD ORBIS web product, and the files on WRDS were incomplete.

Once the data is delivered, it must be stored, often locally hosted on Duke's custom instance of Box.com, or saved on a shared network drive. There are exceptions, such as the Nielsen data sets, which are marketed by and stored at the University of Chicago. To obtain or extract remotely hosted data, the vendor may offer an application program interface (API), a set of tools for accessing the data, or they may deliver the data on DVD discs or external hard drives. Even so, effort is still required to make the data usable by researchers. These efforts can include actual data extraction, and "cleaning" or reformatting data to enhance its usability.

How: Duke's Process Outlined

Funding for one-off data purchases and renewals is often shared. Specific data products needed by individual faculty members may be funded by their own research budgets, if sufficient. At most research universities, the Deans' offices have pools of available funds, and libraries also have sizeable database budgets.

After many months of experience, the process for approving and funding a one-off purchase at Duke has evolved into the following steps:

- A faculty member identifies the needed resource.
- At the library, reference librarians verify that the information is not already available in-house. This is necessary because faculty members are sometimes unaware that the data they want is a component in a subscribed data product.
- If a purchase is needed, a librarian identifies how to obtain it and locates the right person at the right company, someone responsible for delivering the information. This step can be surprisingly difficult when the data are a byproduct of a large company's mainstream business. The price and the terms of the license must be negotiated. Difficulties may arise at

this stage when the information supplier is not in the business of selling data to libraries or to researchers. The dealer may also have unrealistic expectations regarding price or usage.

- Appropriate funding must be negotiated internally at the university to cover the purchase price. Since funding is negotiated among a number of sources, including the library, a faculty member, and their department, a close relationship between the lead librarian and a high-level academic officer is key to determining the optimal funding scheme for each data purchase. Professional respect and trust among all players is critical.

Why: Value Proposition Stated

In my opinion, here are several reasons why a library wants to be involved in this activity.

First, librarians have been purchasing and licensing data for decades, developing a deep knowledge base about making resources available to university researchers. Librarians are best positioned to know which resources exist on university campuses. Librarians have expertise in locating, purchasing, and licensing new resources, including an awareness of licensing terms that might restrict use of the data. Librarians' knowledge about vendors and how to communicate with them is invaluable. Librarians also know about data accessibility and storage and are aware of sharing and preservation issues. As university employees, we are compelled to share this knowledge for the benefit of our employer.

Second, the nature of data use on our campuses is changing. At Duke, use of traditional platforms for business data is decreasing. Former must-have products such as OneSource, eMarketer, and Frost & Sullivan have decreased sharply. Use of the WRDS platform is also decreasing. (See Table 1.) While the number of web queries and unique faculty users are both declining fairly consistently, the number of observations and the number of SSH-SAS file touches, while variable, seem to be holding steady. These numbers indicate that fewer people are downloading more data, a key trend in business faculty research today.

Table 1: Duke Faculty WRDS Data Use

Year	Web Queries	Observations	Web Unique Faculty	SSH-SAS File Touches (omits NYSE TAQ)	SSH-SAS Unique Faculty
2011	1,176	685,647,096	30	374	4
2012	908	166,074,720	28	284	6
2013	1,399	945,860,148	30	574	6
2014	560	1,292,014,385	25	2,306	6
2015	432	2,384,246,784	25	162	4
2016	430	733,161,199	24	1,250	5

Faculty members require increasingly specialized data, which is driving them to purchase large custom data sets. The need for larger quantities of data is common. If, at one time, faculty used electronic databases to find a needle in a haystack, now they want to buy the whole haystack—a one-off, customized haystack.

Third, providing customized data sets creates opportunities for professional growth for library staff. Librarians have expertise in locating, purchasing, and licensing data. Providing custom data

extracts enables librarians to develop new skills in collecting, integrating, and processing data. These new skills are highly valued in research organizations, increasing the value of library staff throughout an organization and leading to greater compensation for staff.

The final and most important reason a library would want to provide one-off data sets for faculty is strategic. Libraries play a key role in supplying resources essential to university faculty members in order to maintain their funding and status. Academic librarians have been information providers for research faculty for more than a century and need to continue to play this important role. Administrators readily distribute funds to support their top priorities, and faculty research will always be a research university's top priority.

Call to Action: From Needles to Haystacks

Librarians can create new value for users, leverage existing expertise while building new capabilities, get ahead of the curve in research trends, and remain relevant in the digital age. High impact, high leverage, minimal risk—the haystack is an exceptional opportunity waiting to be harvested. Join the new quest for haystacks rather than needles and up your library's game in obtaining one-off data sets.

Share your experience. What have you seen in your organization and/or library regarding the rapidly growing trend to obtain unique, one-off data sets? What role is your library playing? What processes have you put in place or experimented with? What have you learned? What value are you creating for faculty, and how are they reacting? What other details can you share? Send a brief overview to me at mtrauner@duke.edu. I'll collate your submissions and submit an update to *Ticker*.