Issues in Information Systems
Volume 16, Issue III, pp. 64-69, 2015

TOO "BIG" TO PROCESS: BIG DATA IN THE BUSINESS LIBRARY: KEY RESOURCES AND IMPLICATIONS FOR THE FUTURE

Kara Gust Rawlins, Combined Arms Research Library, kara.j.rawlins.civ@mail.mil
Breezy Silver, Michigan State University, silverbr@mail.lib.msu.edu

ABSTRACT

As Big Data continues to grow and impact all aspects of industry and academia, libraries and librarians will play key roles in providing access to resources for its analysis. This paper will explore the relationship between Big Data and the academic business library, as well as existing strategic data resources and their usage trends at Michigan State University. It will also explore what librarians need to know about Big Data and why it is essential to the use of library resources and continued open access to information.

Keywords: Libraries, Big Data, Databases, Statistics, Numeric Data

INTRODUCTION

Big Data is a relatively new term, but a concept libraries have been familiar with for decades. As repositories of information as their chief function, libraries have served as data warehouses full of thousands of research databases and statistical information; and they will continue to do so into this new era of Big Data. While the concept of Big Data is growing and trending, the charge of libraries and librarians to provide timely access to information will become increasingly more important as data becomes increasingly difficult to search and discern; especially as the idea of uncompiled raw data becomes more desired by researchers. Business libraries will be at the forefront of this need for not only access to Big Data, but a way to search it as the demand for real-time marketing data, consumer social characteristics, and competitive intelligence becomes increasingly important.

Many highly expensive and more complicated and manipulative data sources have now become commonplace in the academic library and business faculty have come to demand them not only for instructional purposes for the classroom but for their own research needs as well. To help both business librarians and faculty identify the resources essential to business data research now and in the future, this paper will provide an overview of key business databases, their scope and capabilities, and investigate their usage trends at Michigan State University over the past 3-5 years. It will also explore why the business librarian must be knowledgeable and well versed in data sources and increasingly familiar with Big Data.

LIBRARIES AND BIG DATA: THE RELATIONSHIP

Librarians possess, at their core, key skills in information, organization, and management; the ability to connect users with the data or information they desire [9]. Traditionally, this has involved directing users to reference books, article indices and abstracts, journal literature, and statistical sources—all across widely varying subject areas. It was the role of the librarian to analyze and interpret specifically what users wanted to know and what information sources would satisfy those questions. This has changed however in the growing boom of data that is collected and harvested as commercial resources focusing on data, statistics, and numeric datasets has grown rapidly. Librarians must now also be able to provide "the same level of service to data as they already bring to other formats of information" [7].

Historically, libraries have provided access to data and statistical sources in the form of various Census and US Government statistical publications. While "numeric datasets have not always been associated with traditional library collections" [7], the shift to more data-centric resources came in 1962, when the Inter-university Consortium for Political and Social Research (ICPSR) produced its data archive focusing on voting patterns and behaviors. Beginning with punch cards and magnetic tape, the ICPSR now has a "data archive of more than 500,000 files of research in the social sciences" [5], available in raw format such as SAS or ASCII for analysis. The need for more transferrable, manipulative, and actionable data was born and an ever-growing need to fulfill the access to these resources became integral to the work of libraries and librarians.

Today, this need for access to data has grown even further as government, companies, and organizations compile and gather amounts of data that are considered almost too big to process by conventional technology. Big Data is generally defined as "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" [6]. The datasets are considered huge and need to be managed by tools that can handle terabyte and petabyte amounts of data [1]. The size and type of data can also vary widely by industry and organization. It can go beyond the traditional and contain more "unstructured data" [1] such as Facebook posts, Twitter feeds, "email messages, photographs, postings on Internet forums, and even phone transcripts" [1]. In today's world of social media, mobile purchasing, and location tracking, these are all sources of personal data that might impact businesses and their ability to reach key markets and analyze spending habits. Any of which can provide them with a competitive advantage.

Libraries traditionally have been vast warehouses of data, information, and access points to research, and this will continue in the era of Big Data. Clearly, all librarians, but especially business librarians, will have to be aware of current resources for data harvesting and manipulation and key business intelligence to best serve their patrons and academic institutions.

"BIG DATA" DATABASES AND RESOURCES IN THE BUSINESS LIBRARY

Business library databases have continued to emerge and evolve, and will especially do so in this era of massive data curation. There are numerous databases available today, however, that already provide large and extensive datasets, numeric data, and statistics relative to business, entrepreneurs, labor, economics, and especially consumer demographics. They will continue to expand and grow as more Big Data is created and collected for use by businesses, government, and researchers. The key difference for the development of future databases is that they place a significant emphasis on providing not only access to data, but on data manipulation as well. As Starkey writes, "Most products in the data marketplace are not providing static data sources and are moving toward analytics, visualization, and customization" [10]. This involves offering features such as the downloading and importing of large raw data files for secondary analysis, charting and plotting of consumer demographics according to a specific location, and visually mapping specific variables and points of comparison. These features are evident in some of the most dominantly used numeric and manipulative databases in the Gast Business Library at Michigan State University (MSU) today as described in the following overview.

WRDS

Wharton Research Data Services (WRDS) is a platform that contains datasets from various data providers. The benefit is that this platform standardizes the interaction with the data and allows large downloads either through the web platform, SSH/Unix, or within SAS. The most commonly accessed data is financial including stock and company data like Compustat and CRSP but there are some other datasets like Comscore and software like Eventus. The datasets available vary among institutions since each must decide on and purchase the particular datasets that their particular institution wants. At MSU, WRDS is accessible to faculty/staff, graduate students, research assistants, and classes. Users must create an account to access and use WRDS. This is a complicated database to use for the novice student or researcher; those having an understanding of datasets and their variables have a much easier time navigating and using it.

It is worth noting that in 2013, access to WRDS at MSU was opened up to the university community. Previously, it was only accessible to those in the Eli Broad College of Business (COB) and they had to request access to it through the libraries. Librarians noted that there were an increasing number of requests for access outside the COB. One business librarian noted that even someone in family medicine requested access to it showing that there is clearly a much greater and more interdisciplinary demand for data. This is also evident in the usage stats for WRDS over the past several years. From 2009 to 2012, the total number of searches across all datasets increased nearly 60%. The latest data available is from 2014 and shows the total number of searches for that year increased another 35% from 2012; which again reflects WRDS being available university-wide rather than limited to the COB, as well as a greater demand for its content.

Passport (GMID)

Euromonitor's Passport (formerly Global Market Information Database) is a database that has industry, market, company, and country reports and data. It is not as data heavy as WRDS, but it does have a fair amount available. The best asset in this database is that it is international in focus and has quite a bit of unique data including sales, company and brand shares, prices, trade data, market size, outlets, country data, and more. This database has a considerable amount of data in it and is fairly easy to use. However, it contains so much data, that it is often difficult to be aware of and find all that is available. It also provides users options for some data manipulation as users can select from Passport's tables specific types of data and download it. Passport also has an add-on called Survey Dashboard which offers a detailed examination of consumer motivations, opinions, and habits. The use of Passport has continued to rise. From 2012 to 2013 there was a 20% drop in use but from 2013 to 2014 there was a 42% increase in use.

SimplyMap

A highly popular business database available today, in terms of providing complex data for manipulation, is SimplyMap. Produced by Geographic Research Inc., SimplyMap is a web-based mapping tool that allows businesses, researchers, and students to access extensive U.S. business, demographic, and marketing data and create thematic maps based upon their selection of variables. They currently offer more than 75,000 variables from leading sources such as Nielsen Claritas, Simmons, and Mediamark. In addition, it contains historical Census data back to 1980 and health data. Specialty data can be included as added packages. With its incorporation of extensive and valuable consumer demographics and mapping feature, SimplyMap is an excellent example of how users can manipulate data and turn it into valuable information to help make important business and research decisions. Usage data for SimplyMap at MSU was attainable for only the past two years. From 2013 to 2014, there was an increase of over 27% in the total number of sessions recorded, with an increase of nearly 18% in the total number of variables used. As of May 6, 2015, there have already been more than half of the sessions logged in 2014 and so it is well on pace to exceed the number of session from the previous year.

Simmons OneView

Simmons OneView allows companies and researchers to create customized reports of an extensive amount of consumer demographics, product purchases and preferences, and media data provided by the National Consumer Study. Conducted for over 50 years by Simmons Market Research Bureau, now Experian Simmons, the " NCS annually surveys 25,000 U.S. individuals 18 years of age and older as well as households across multiple categories including in-depth demographics, psychographics, lifestyles, attitudes, opinions, and product usage behavior" [4]. It also includes coverage of over 500 product categories and 8000 brands. For years, Experian has provided companies and libraries access to the NCS data through their Simmons Choices3 viewer, which has now been replaced by Simmons OneView. Although it provides access to the same data as Choices3, a great advantage of Simmons OneView is that it is web-based with a much more user-friendly interface. Usage data for Simmons OneView was only available for the past year. During that time, from June 2014 to May 2015, well over 5,500 reports or "crosstabs" as they are known were generated. This is something that will need to be tracked and compared for future research.

Summary

These are just a few examples of databases that are growing in popularity at MSU as the demand and use of "data" and especially manipulative databases increases. Data is being requested increasingly more by faculty and students and the Business Library is there to provide that needed data. Librarians are at the forefront of helping in data research, planning instruction with students, and most definitely, purchasing resources and datasets. While librarians cannot purchase all data requests they receive, these are some of the major data sources most used in the business field and how both students and faculty can use them to their advantage. These have shown over time that they have been and will continue to be important reference sources for scholars, students, and researchers.

BIG DATA: WHAT LIBRARIANS NEED TO KNOW AND WHY

A report by the McKinsey Global Institute states that "Big Data has now reached every sector in the global economy" [6]. Its sheer volume and scope is going to impact every industry across the global landscape including education, healthcare, and most especially government. Already "15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress" [6]. Because of this, librarians will definitely need to have a basic understanding and knowledge of Big Data and how it will impact their libraries and academic research at their institutions. Not only should librarians be aware of Big Data tools to add to their libraries collections but also because the faculty will most certainly be "increasingly incorporat[ing] big data into their research" [1].

For business librarians, they will especially need to educate and familiarize themselves with Big Data as firms gather it and use it to help them find a competitive advantage, and to open or expand markets. The McKinsey Global Institute reports that Big Data will "became a key basis for competition and growth," estimating that a "retailer embracing big data has the potential to increase its operating margin by more than 60 percent" [6]. For business librarians, a key part in providing appropriate and accurate assistance with company and market research is having a solid understanding of overall company and industry trends. Therefore, business librarians will increasingly need to be aware of how Big Data is affecting a firm's income and costs, advertising schemes, and whether or not it is being leveraged successfully. Business librarians will also need to know how Big Data will affect students as more of them will be applying it and analyzing it in their coursework, and potentially in their future employment [1]. An increasing number of universities are incorporating classes and degrees on analyzing big data including Michigan State University whose first class for the Master's in Business Analytics began in January 2013. And most certainly, business faculty will be exploring and incorporating Big Data into their academic research and business librarians will need to be familiar with appropriate resources for them as well.

The Emergence of the Data Librarian

With the increasing proliferation of government statistical data, business data, and growing demand for more datasets and data research, many libraries already have seen the need for a specific person to handle all of this type of reference and research assistance – the "data science" librarian. This is a significant development and addition to the library profession as data science becomes more a part of industry and academia. It is now essential for libraries to have a professional who can successfully discern between all the various forms of data, statistics, and analysis, and can direct users to appropriate resources for each. The concept of responding to a "data" question in the libraries has always been somewhat mystifying and challenging to any librarian. The data librarian can provide a vital service by not only greatly assisting patrons but other librarians as well in deciphering any sort of data related reference question or statistical research problem.

Even more integrative and beneficial to the profession and academic institutions is the successful collaboration of data librarians and business librarians. An excellent example of collaboration between the data librarian and business librarian exists at Michigan State University. They identified that the College of Business faculty were "prolific users of data" and working together, they were able to "create a central list of data purchased, created, or used by faculty" [7]. By such outreach and collaborative initiatives, they were able to identify any areas where data purchased by faculty either duplicated or overlapped with other departments and/or the library. As a result of their efforts, the "COB and library agreed that data purchases would go through the library" [7]. The library also "assists with purchases, examines license agreements, and suggests alternative resources, as necessary." With the great expansion and demand for data and in the coming years the emphasis on Big Data, their collaboration is an excellent example for other business libraries and academic departments to follow as an essential cost-saving and time-saving measure.

Open Access and Privacy

Librarians have always acted emphatically as safeguards for open access to information and protection of privacy. As the collection and harvesting of Big Data in various forms epitomizes the intrusion of privacy and need for open access, it is vital and completely in their nature that librarians serve as Big Data liaisons [9]. This is more critical now than ever before with the existence of both government and industry harvesting excessive amounts of personal data. Prime examples of this are the National Security Agency (NSA) and Target Corporation. Beginning in 2007,

through its PRISM surveillance program, the NSA's wholesale collection of Internet and cell phone communications, for supposed national security concerns, has been a widely debated violation of personal privacy and overreach of their authority as a U.S. government security agency. Target Corporation has also come under scrutiny after successfully using purchasing habits and shopping data to determine which of their shoppers were pregnant and precisely in what trimester [2]. From these examples, the collection and use of Big Data can be seen by some as positive in terms of protecting national security and providing a competitive advantage in industry, however, also severely undermining personal privacy rights. There is a very delicate balance between the gathering and use of Big Data and privacy rights that will certainly continue well into the future.

With so much data being harvested by government, there has also been a call for it to become more transparent and for more attainable to the public. In the past several years, the U.S. Government has responded to these concerns and has opened up astounding amounts of information through such open access initiatives as Data.gov, Project Open Data, and also by executive order making open and machine-readable the new default for government information. When Data.gov was launched in 2009, its vision was to "increase public access to high-value, machine-readable datasets generated by the executive branch of the federal government" [3]. It began with just 47 datasets but as of 2014, it contains "more than 100,000 datasets from 227 local, state and federal agencies and organizations" [3]. Several years after Data.gov was established, Project Open Data was created to provide guidance, cataloging and implementation tools, checklists, and case studies to assist agencies in complying with open data policies. This flood of government information and open data practices could have a potentially large impact on data available to industry and research. As champions of open access, librarians need to be well educated and proactive in such new data policies and initiatives.

Big Data Accessibility and Collection Development

As librarians have to promote traditional library resources and services, they will also need to serve as "ambassadors" of Big Data (and data overall) by acquiring complex datasets and manipulative databases when deemed appropriate. They will need to help make these resources more easily accessible and usable as the demand for them grows. Similar to collection development processes as it relates to traditional library databases, librarians will also have to be extremely vigilant to push database vendors to provide data more easily (i.e., limit the size or increase the ability for users to do more "data dumps" or text mining. Librarians and vendors will have a very important relationship in this process as database vendors anxiously try to catch up with the data craze, and the infrastructure needs to be there to provide it.

Big Data tools and resources will also need to be analyzed very carefully in terms of which ones to purchase. The data librarian can work with library management and administration to help determine if there are new and emerging, and perhaps "massive" datasets that can be purchased that previously were completely beyond their scope and budget [1]. It is a delicate balancing act because the library cannot possibly buy every single specific dataset and tool when only a very few number of people might be using it. This can emerge as an area of conflict when the library is determining whether to buy a product for the entire university or for a specific department because it is for just one or two individual's research.

Librarians will also need to take the lead on Big Data curation, which will involve making "big datasets more useful, visible and accessible by creating taxonomies, designing metadata schemes, and systematizing retrieval methods" [1]. They can assist with this process by working with vendors, developers, and across their own institutions. They can also consider harvesting their own collection of Big Data for their universities such as Arizona State which uses data mining software to help students choose classes, monitors their progress, and analyzes their behavioral patterns [8]. Additionally, it would be especially advantageous to consider building a Big Data repository of their own faculty's research and raw data [1]. This data heavy era could serve as a tremendous opportunity for libraries and librarians to become even more integral to their institutions and faculty research in providing valuable data collections and considerable levels of access.

CONCLUSIONS AND IMPLICATIONS FOR THE FUTURE

It is evident that the overall demand for data in various formats, data mining, and manipulative databases is on the rise. Just as librarians have traditionally provided access to information and identified appropriate resources for

information and research, they will continue to do so in the era of Big Data. The most critical areas on which librarians will need to focus in this era are education and collaboration. With the availability and demand for data resources greatly growing and expanding, librarians will need to continue to educate themselves on Big Data and its implications in terms of access to information, application to industry, personal protections, and academic research. Educated library personnel are going to be key in being able to navigate the web of various data resources and properly direct faculty and students to the most helpful resources available to them.

All librarians will need to be aware of Big Data, and business librarians are no exception. They will need to continue to collaborate with data librarians and be essential conduits between researchers and data sources—between the business faculty and students and datasets/databases. Continued collaboration between business librarians and data librarians and between various departments will be vital in ensuring that all involved are aware of the benefits of using data resources that are available and which are best for answering their research questions. As data sources are increasingly expensive, it is imperative that librarians and faculty work together so datasets and resources are not duplicated unnecessarily or purchased with limited access. Collaboration will be key between librarians and database vendors as well. Librarians will have to remain extreme advocates for more data availability and access and work with vendors in providing useful and suitable data sources to academic institutions.

In addition, future research must continue to properly address data usage needs in the library. Building upon this paper, a survey should be conducted to ask business faculty at Michigan State University how datasets and databases are fulfilling their specific research needs and whether more will be needed (or should be purchased) in the future. It would also be helpful to compare the usage statistics of data resources at MSU across other comparable academic libraries such as those in the Big Ten Conference. Business librarians should also consistently analyze the usage statistics of databases and datasets to ensure they are purchasing the most beneficial ones for their constituents, while making sure they are promoting them as well.

REFERENCES

1. Bieraugel, M. (2013, June). Keeping up with…Big data. Retrieved from http://www.ala.org/acrl/publications/keeping_up_with/big_data.
2. Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times*. Retrieved from http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0.
3. Fretwell, L. (2014, June 9). A brief history of open data. Retrieved from http://fcw.com/Articles/2014/06/09/Exec-Tech-brief-history-of-open-data.aspx?m=1&Page=1.
4. Rawlins, K. (2009). Making sense of Simmons: Understanding and using Simmons market data and other marketing databases. *Issues in Information Systems, 10*(2): 12-21. Retrieved from http://iacis.org/iis/2009/P2009_1228.pdf.
5. Inter-university Consortium for Political and Social Research (2013). About ICPSR. Retrieved from https://www.icpsr.umich.edu/icpsrweb/content/membership/about.html.
6. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011, May). Executive summary. In *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.mckinsey.com/~/media/McKinsey/dotcom/Insights and pubs/MGI/Research/Technology and Innovation/Big Data/MGI_big_data_exec_summary.ashx.
7. Mooney, H., & Silver, B. (2010). Spread the news: Promoting data services. *College & Research Libraries News*, *71*(9): 480-483. Retrieved from http://crln.acrl.org/content/71/9/480.full.
8. Parry, M. (2012, July 18). Big Data on campus. *The New York Times*. Retrieved from http://www.nytimes.com/2012/07/22/education/edlife/colleges-awakening-to-the-opportunities-of-data-mining.html?pagewanted=all&_r=0.
9. Stanton, J. M. (2012, July 16). Data science: What's in it for the new librarian? *Information Space.* Retrieved from http://infospace.ischool.syr.edu/2012/07/16/data-science-whats-in-it-for-the-new-librarian/.
10. Starkey, Jennifer (2015, February 15). Data-Planet. *The Charleston Advisor, 16*(4): 21-25. doi:10.5260/chara.16.4.21.