Big Data or Big Brother? Data, Ethics, and Academic Libraries

Barbara Fister

Published in *Library Issues: Briefings for Faculty and Administrators* 35.4 (March 2015)

We live in an era of Big Data, in which we are able to collect and analyze data at a speed and scale that is unprecedented. Indeed, much of the what we think of as our information infrastructure today – from our favorite social media platforms to the website of *The New York Times* – is built on a business model that depends on collecting data from users and monetizing that data through targeted advertising, sale of bundled data to third parties, or both. The U.S. federal government has also revolutionized its use of data, relying on post-9/11 legislation and a Reagan-era executive order to build a data-gathering system so vast that William Binney, a former NSA official, has described it as being close to being a "turnkey totalitarian state."

Currently, Google is the most prominent corporation to develop and use deep data-gathering practices for its massively successful advertising business, with Facebook following close behind. It's hard to imagine doing research today without Google products. Even though many users understand that they are paying for the "free" service by providing the company with volumes of personal information, the tradeoff seems either worthwhile or unavoidable. Though there are privacy-friendly browsers such as Tor (www.torproject.org/) and alternatives to Google's search platform, such as Duck Duck Go (duckduckgo.com) and Startpage (startpage.com), they are far less popular and arguably less effective than Google's Chrome browser and search engine. Though the revelations about the National Security Agency's use of data made by Edward Snowden's leaks have sparked discussion about privacy and security, citizens often feel privacy is a thing of the past.

In 2010 Mark Zuckerberg, founder of Facebook, told an audience of technologists that privacy was no longer a social norm, that people preferred sharing their lives publicly. That didn't stop millions of Facebook users protesting changes in the platform's privacy settings, but membership in social networks remains high. Because of social media's popularity, many assume that Zuckerberg is right, that young people no longer care about personal privacy. This is not true, as researcher danah boyd explains in detail in her groundbreaking study of youth culture online, *It's Complicated: The Social Lives of Networked Teens* (Yale University Press, 2014).

Most Americans are aware that their privacy is at risk. A Pew Internet and American Life poll conducted in November 2014 found that over 90 percent of Americans feel they have lost control over the way their personal information is collected and used by corporations and 80 percent believe citizens should be concerned about government monitoring of online communications.

So What Does This Mean for Libraries?

Privacy is one of the core values of librarianship, but in a data-saturated world, this value is growing difficult to interpret and uphold. According to the American Library Association, protecting user privacy is "necessary for intellectual freedom and fundamental to the ethics and practice of librarianship." Yet the benefits of Big Data challenge that position. Academic librarians need to consider how important privacy is to them when the tools they use in everyday practice – search tools, licensed content, Facebook and the other social media platforms they use for outreach, their own websites – gather and mine data about people and their connections.

Librarians need to think about whether their patron records are secure as data breaches become common. In analyzing library catalog platforms, Marshall Breeding recently reported that while some patron data is generally secured, almost no catalog vendors use end-to-end encryption throughout their platforms. This is problematic as it grows increasingly easy to identify individuals by combining anonymized data sets. In an article in the January 2015 issue of *Science,* researchers examined three months of anonymized credit card data from over a million Americans. They found it was possible to identify 90 percent of the card holders with as few as four pieces of data.

Academic libraries face many new challenges in an era of Big Data. They will be called upon to support the use and preservation of data as an increasingly valuable piece of our knowledge ecosystem, which will require developing new library programs and skill sets. They will also need to think through how new uses of data might benefit the library's mission and how (or if) those new uses can be reconciled with privacy.

The Big Picture on Big Data

According to data scholar Rob Kitchen, there are several characteristics that define Big Data:

- volume (it's called "big" for a reason),
- velocity (it's created in real-time or close to it),
- variety (capturing many kinds of data, both structured and unstructured),
- exhaustive (trying to capture entire populations or systems),
- fine-grained (extremely detailed),
- relational (connectable to other datasets), and
- flexible.

Consider that in 2012 WalMart reportedly collected 2.5 petabytes of data about customer transactions *every hour*. This is a far cry from previous data collection programs such as the decennial U.S. Census of Population and Housing and the General Social Survey. Having the ability to gather, link together, and analyze vast amounts of real-time information offers promising new ways to study our world. Chris Anderson (former editor of *Wired* magazine) has predicted that Big Data is so transformatively powerful that it makes the scientific method obsolete. Data itself will reveal findings without anyone having to pose a hypothesis or develop a theory.

Kitchen and others have countered that data are not pure and unbiased, they are created within social processes that shape it. Algorithms are not *sui generis*; they are created by people, and people have to interpret what the results actually mean. Many critics have also pointed out that ubiquitous data collection often harms those who are marginalized or who challenge those in power.

When the president claimed that the NSA's collection of phone data in bulk is harmless because it's "only metadata," sociologist Kieran Healy imagined how such an approach during the colonial era could have allowed the British to identify Paul Revere with geospatial data about his famous ride.

A more somber essay by historian Ivaro M. Bedoya reminds us how information was used to identify Japanese-Americans for removal to internment camps and capture fugitive slaves, arguing that we need to balance the benefits of ubiquitous data collection with the understanding that we cannot always trust those who might use it.

Zeynep Tufekci, a sociologist with a joint appointment in the School of Information and Library Science at the University of North Carolina, Chapel Hill, notes that there are additional features of Big Data that we must consider. Not only is a large amount of data available for research, it is increasingly being tied directly to individuals. It's being used for persuasion (particularly in the commercial sector). It changes dynamically in ways that are opaque to its subjects,, and it has given rise to new power brokers who control the data and algorithms as valuable but hidden trade secrets.

A glimpse into these issues and how people feel about their privacy was provided when an article in the prestigious *Proceedings of the National Academy of Sciences* reported on an experiment in which Facebook users were unknowingly used as subjects in an experiment that involved manipulating their timelines to see how it affected their mood. Though such manipulation is standard operating procedure for Facebook, many Facebook users found it creepy and invasive. Further, the research design didn't meet expected standards of informed consent typically used by scholars. The journal, facing criticism, finally issued a statement of concern, a step short of retraction.

Beyond issues surrounding the collection and use of Big Data, some critics feel we rely too much on its gee-wizardry. Morten Jerven, an economic historian, boils this down to four points:

- Not everything that counts can be counted.
- Data is not the same thing as statistics.
- More data doesn't mean better decisions will be made.
- There are other methods for finding things out than by counting.

All of these issues have an impact on what libraries do, how they might use data themselves, and how the profession will strike a balance between the benefits of data gathering and the threat it poses to privacy.

Data in Library Collections and Services

With so much data being collected and analyzed in the course of research, one of the many challenges academic libraries currently face is managing and preserving locally-generated data. Sharing data publicly is an expectation that is increasingly built into grant funding and often required by publishers, particularly in STEM fields. Though increasingly scholars expect research data to be discoverable and publicly available, not all researchers have the resources or training to manage and maintain data.

Many academic libraries are developing data repositories and helping researchers write data management plans, create documentation, and learn how to use standardized metadata. Other libraries without sufficient staff resources are beginning to explore how they might provide these kinds of support without additional funding or in-house expertise.

Further, as collecting and analyzing data becomes part of the curriculum, academic libraries will have to fold data literacy into their instructional programs. This will entail helping students find or create datasets and providing instruction in the use of analytical tools, from geographic information systems to statistical packages to emerging visualization tools. Because these needs are developing so quickly and have features that make them distinct from traditional library instruction and repository programs, they present a significant challenge to academic libraries of all sizes and will likely require making some difficult decisions about staffing and resource allocation.

Using Data in Collection Development

Librarians need to stretch acquisition dollars, so naturally want to know how often a resource is used. With print collections, circulation records provide a limited snapshot of use, often omitting consultation at the shelf or other in-library uses. Electronic resources offer more fine-grained data collection. Databases offer various analytics, though harmonizing the different kinds of information they provide and finding ways to analyze it effectively can be a challenge.

In the case of books, electronic packages collect information we never had before: which books have been browsed, which chapters downloaded, and even (in some cases) which pages were examined. Vendors have their own uses for this data, though the details of how these data are used are not always readily available.

While these data are useful, its collection can seem worryingly intrusive. A news report on an ebook vendor in the UK recently reported that only 28 percent of ebook consumers finished reading *Twelve Years a Slave,* a 19th-century memoir that landed on the bestseller list after being made into a film. Contemporary novelist Donna Tartt didn't fare much better. Less than half of those who purchased the Pulitzer prize-winning novel *The Goldfinch* finished reading it.

Though consumers may be disturbed to realize their reading patterns are being monitored, publishers stand to learn a lot by looking over the shoulders of readers. Information about the use of resources is often gathered in the background by vendors and software platforms, and that information is in some cases, shared with third parties.

That privacy challenge is compounded when it's shared insecurely. Google encrypts its searches, but, as Gary Price recently demonstrated at a CNI meeting, the details about the actual use of Google Books by searchers are not encrypted. When Adobe Digital Editions, used widely by academic libraries, recently launched a new version, librarians learned the company collected information about what users were reading in astonishing detail – what books they read, when they were read, where they were read, even which pages were looked at, all linked to specific user IDs.

That was disturbing enough, but what was truly alarming was that Adobe transmitted that information over the Internet in plain text, meaning it could be easily intercepted and read by others. Though the company quickly acknowledged that mistakes were made and assured consumers that this information would be encrypted in future, it was an illustration of how hollow libraries' commitment to privacy is when we aren't paying attention to what our vendors are doing.

In Andromeda Yelton's estimation, we haven't given this nearly enough thought:

> *If you have chosen, whether actively or by default, to trust that the technical affordances of your software match both your contracts and your values, you have chosen to let privacy burn. If you're content with that choice, have the decency to stand up and say it: to say that playing nice with your vendors matters more to you than this part or professional ethics, that protecting patron privacy is not on your list of priorities.*

Rick Anderson has also argued that we need to confront these issues head-on, but while we value privacy, we also have a duty to provide information. He believes we need to consider whether it might be better to let patrons decide which is more important to them: access or privacy. In any case, by

simply asserting that we value privacy without conducting audits of our own privacy practices we are ignoring a clear and present need to find a workable balance.

Learning Analytics

Academic librarians have long been involved in assessing the role the library plays in student learning. (See *Library Issues,* Jan. 2003, Sept. 2008, Nov. 2013.) In recent years, librarians have also used data to provide institutional decision-makers with strategic information about the library's return on investment. These two trends come together with the Big Data movement to pose a question for academic libraries: Do the new capabilities of data mining offer libraries valuable opportunities to peer into individual students' experiences to improve learning and demonstrate value – or does it pose a significant challenge to one of librarians' most revered values, patron privacy?

This debate is likely to heat up as the practice of using learning analytics becomes embedded in the discourse around higher education in the United States and as learning analytics become an entrenched feature of K-12 education.

Learning analytics is a move to adopt Big Data technologies to compile data from various sources – potentially including the classroom, the finance office, student services, and the library – so that problems can be predicted and a collective solution applied. Purdue University's Course Signals has codified this approach for the past decade. Some universities are adopting commercial services that encourage students to use swipe cards to track their visits to the career center, student leadership programs, and other co-curricular programs to encourage engagement. Libraries are developing ways to peer into multiple data sets and run analyses to see whether library use correlates positively with student success. Some would argue that not doing so because of abstract privacy concerns would be an injustice to at-risk students.

Learning analytics are not just Big Data, they are big business. Critics, such as ed-tech journalist Audrey Watters, question whether the money spent on corporate data-driven schemes would be better used for improving the working conditions of faculty or reducing the financial stressors that hinder student success.

Academic libraries face similar decisions. How much time and energy should we put into analyzing data about our students? Are there other, less invasive, perhaps more effective methods available? Should we moderate our traditional defense of privacy to enable data-driven processes that might help students succeed and may provide evidence that the library's budget is well-spent? Or should we play a more active role in defending privacy in a digital age?

This clash of values is neatly illustrated in the awards made by the Knight Foundation for its Knight News Challenge on libraries, announced in January 2015. Among the projects chosen were several that would help libraries use tracking technologies to expose resources to library users or help libraries understand how their spaces are being used by patrons. Two projects will provide educational programs about privacy for library patrons.

We cannot avoid facing the tension between the potential new data-gathering technologies offer and the risks they pose to privacy given that our values are already in conflict. As Hugh Rundle, Australian librarian and technology pioneer, points out, we've already made a choice. We just don't always realize it.

Currently most libraries seem to be (accidentally) providing a huge hoard of private user data to virtually anyone who wants it, but not actually using any of it themselves. If we are to credibly claim to be defenders of intellectual freedom and *responsive to our communities, we need to use data more cleverly - and protect member privacy while we do so.*

References

Alvaro M. Bedoya. "Big Data and the Underground Railroad." *Slate.* Nov. 7, 2012. http://www.slate.com/articles/technology/future_tense/2014/11/big_data_underground_railroad_history_says_unfettered_collection_of_data.html

Kieren Healy. "Using Metadata to Find Paul Revere." http://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/

Morten Jerven. "The Problem With the Data Revolution in Four Venn Diagrams." *The Guardian.* Dec. 22, 2014. http://www.theguardian.com/global-development-professionals-network/2014/dec/17/data-revolution-limitations-in-images

Rob Kitchen. "Big Data, New Epistemologies, and Paradigm Shifts." *Big Data and Society.* April 2014. http://bds.sagepub.com/content/1/1/2053951714528481

Hugh Rundle. "A Measured Approach." https://www.hughrundle.net/2015/02/01/measured-approach/

Zeynep Tufekci. "Engineering the Public: Big Data, Surveillance and Computational Politics." *First Monday.* July 2014. http://firstmonday.org/ojs/index.php/fm/article/view/4901

Audrey Watters, *Hack Education*. http://hackeducation.com/

Andromeda Yelton, "Ebook Choices and the Missing Soul of Librarianship." http://andromedayelton.com/blog/2014/10/08/ebooks-choices-and-the-missing-soul-of-librarianship/

Additional links to sources for this article can be found at https://www.zotero.org/bfister/items/collectionKey/WMDE2V42

Resources

American Library Association. *Privacy Toolkit.* http://www.ala.org/advocacy/privacyconfidentiality/toolkitsprivacy/privacy

Association of College and Research Libraries. *Assessment in Action.* http://www.ala.org/acrl/AiA

Carol Tenopir, Ben Birch, and Suzie Allard. *Academic Libraries and Research Data Services.* June 2012. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf