4-2017

Big data's impact on privacy for librarians and information professionals

Lindsey M. Harper
Marshall University, harper166@marshall.edu

Shannon M. Oltmann

Recommended Citation
Harper, M., & Oltmann, S.M. (2017). Big data's impact on privacy for librarians and information professionals. Bulletin of the Association for Information Science & Technology, 43(4), 19-23. doi:10/1002/bul2.2017.1720430406

Feature

Big Data's Impact on Privacy for Librarians and Information Professionals

by Lindsey M. Harper and Shannon M. Oltmann

EDITOR'S SUMMARY

In a digital age, it is very difficult to maintain complete privacy when posting on social media or making purchases. Individual activity on the internet is increasingly collected by corporations, even with the user's knowledge, and can be used to predict future behavior, purchasing choices or other sensitive subjects. This data analysis is often done without a user's consent and in many cases presents unethical behavior and breaches of privacy. In the world of libraries, the privacy of patrons has been tantamount for decades, but trying to keep up with privacy codes and still make use of this big data can be challenging for librarians. Big data can be beneficial to libraries in many ways, and if pointed at library systems, rather than the habits of patrons, can also keep privacy intact.

KEYWORDS

big data
libraries
privacy
user behavior
personal information
data collection

Lindsey M. Harper is a graduate student in the library science master's program at the University of Kentucky. She currently works part-time at the James E. Morrow Library at Marshall University in the special collections department. She can be reached at lindsey.harper<at>uky.edu.

Shannon Oltmann (Ph.D.) is an assistant professor in the School of Information Science at the University of Kentucky. She can be reached at shannon.oltmann<at>uky.edu.

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it…"

– Dan Ariely, 2013

Posting photos, purchasing groceries, location pings on cellular devices – it is almost impossible to go through daily life in 2017 without data about our information behaviors being collected and examined. People may be aware this data is being collected, but most do not understand its actual or intended uses (even when detailed in Terms of Use agreements). Researchers argue that it is one thing for purchasing, posting or searching behaviors to be shared across companies, but inferences based on the obtained data should be a concern as our technology continues to be enhanced in the future [1].

Library and information professionals have a potential role to play here to help patrons and clients protect their privacy. Protecting the privacy of patrons has been one of the guiding principles of librarianship since 1970 [2]. The American Library Association's (ALA) Code of Ethics, section three, emphasizes the importance of maintaining a sense of privacy for patrons utilizing the library's services. This code reinforces the library professional's obligation to "protect each library user's right to privacy and confidentiality with respect to information sought or received and resources consulted, borrowed, acquired, or transmitted" [3]. As another example, the Information Systems Audit and Control Association (ISACA) Code of Professional Ethics indicates that professionals "[m]aintain the privacy and confidentiality of information obtained in the course of their activities unless disclosure is

Basically, big data about us is the process of accumulating past and present information on individuals in various areas of our lives in order to rationally predict future behaviors or needs.

required by legal authority. Such information shall not be used for personal benefit or released to inappropriate parties" [4]. Big data and its relationship to privacy concerns have become prominent issues facing library and information professionals in the 21st century. This article seeks to examine the impacts of big data within the corporate and information profession worlds. It also aims to examine the advantages and disadvantages of big data, including privacy concerns, and ways to better mitigate concerns.

Big Data Defined

Big data is a term that has yet to be operationally defined; however, De Mauro, Greco and Grimmaldi have a solid foundation for the basis of the word. Through the analysis of 1,437 articles that discuss big data, word clouds with four centralized themes were identified. The four emergent themes include information, technology, methods and impact. *Big data* was therefore defined by these authors as "the Information asset characterized by such a High Volume, Velocity, and Variety to require specific Technology and Analytical Methods for its transformation into value" [1, p. 131]. Basically, big data about us is the process of accumulating past and present information on individuals in various areas of our lives in order to rationally predict future behaviors or needs.

Big data has been widely used in the corporate world, and its methods can sometimes be viewed as unethical or as an invasion of privacy. The retail store, Target, has been cited for utilizing big data to gather information about individuals' purchasing behaviors to send them coupons. Target's algorithm for using data uses predictive analytics and has been able to determine whether a shopper is pregnant based on previous purchasing behaviors of the last 25 items [5, 6]. This predictive data has the potential to make false inferences or to alert members of this person's household that she is pregnant before she has had the chance to do so herself. This kind of information, which has not been voluntarily released, can be viewed as an invasion of privacy to individuals. Conversely, big data's use of predictive analytics can be seen as useful or even helpful, because a pregnant woman could gain access to coupons and promotions to products regarding babies and children that she will eventually desire. When applied on a macro level, big data's use of predictive analytics can help shoppers gain access to goods or services they already use or are likely to use, for the benefit of the corporation using this methodology.

Similarly, big data from the popular social networking platform, Facebook, has been used in research. Kramer, Guillory and Hancock studied whether emotional contagion could occur in an online interaction by manipulating status updates for nearly 700,000 Facebook users. The big data algorithm used here created two distinct groups: one group received positively worded Facebook status updates and another group received negatively worded Facebook status updates. With this data manipulation, results indicated that users who were exposed to the negatively worded statuses were more likely to write negatively worded posts. Additionally, users who were exposed to positively worded statuses were more likely to mimic this sentiment and write positively worded posts. Users who were exposed to fewer statuses of either emotion had the tendency to withdraw or make fewer status updates in general [7]. Panger notes the complexities of this particular study and future studies relying on big data – ethical issues such as the research methodology and obtaining consent are difficult to solve. Upon publication of these results, it backfired among Facebook users [8]. The failure to obtain true consent was the main reason [9], but so was manipulating the well-being of social media users [8]. Because of these published results, Facebook took action in laying out a new research-review board and other training for its employees, and they also updated terms of service to indicate that research might be performed on users [9].

Big Data as It Relates to Libraries

Typically, a library patron's circulation logs are deleted upon return of borrowed materials, and there have historically been no long-term logs for materials that patrons received electronically [2]. However, libraries frequently have to justify keeping materials or receiving new materials in a world where budget cuts are eminent. Big data can have a positive effect when looking at the macro level of the library, rather than the individual user, without compromising patron or student privacy.

When examining big data at the macro level in library and information settings, big data can examine patterns of use for materials, such as how often materials are checked out and which materials are the most popular or underutilized. This data can assist library and information professionals to know which materials can be weeded and which materials should stay in circulation [10]. When using big data to interact with a system, rather than individuals, privacy issues can be mitigated.

When users look for information online, they often feel a sense of privacy because of the anonymity of open inquiry. Big data, on the other hand, uses these inquiries to gather information on the user's search behaviors. Big data's principle as it relates to the library setting is to "support incidental discovery rather than deliberate and purposeful information behavior, hence surprising people into recognition rather than enabling them to ask questions and find out answers. And they do so in a fashion that often invades the individual's privacy and reveals their innermost secrets" [11, p. 500].

The effects of big data and its closely related cousin "network data" can often be considered unethical as they are applied to the library or information technology settings. The methods of big data, when applied to the library setting, can inhibit library users from searching for information they may actually need, because they may not want others to know they need it.

Among library users, big data initiatives, when used improperly, can leave members belonging to underrepresented groups the most vulnerable. Campbell & Cowan focus on big data's impact on LGBTQ library users, specifically those who are early in the coming-out process. The authors posit that networked data, which includes all online browsing, posting and search behaviors, has the potential to be exploitative [11]. Although such data is generally in the public domain, it is often used without explicit consent. Even though individuals on social media or search engine platforms are aware data could be used, they often view the information or inquiries as private [10]. For example, researchers have previously used networked data on Facebook to predict sexual orientation of users. If Facebook users chose to hide who they were "interested in" on the social media site, but had an above average amount of openly LGB social media friends, they were predicted to be members of the LGB community [12]. This conclusion could be drawn inferentially, with the use of a few public profiles and without directly contacting social media users, an example of how big data initiatives can be consequential in the 21st century.

Big data can analyze library users' public behaviors, such as liking and commenting on social media, including on a library's social media page. Big data can also analyze less public behaviors, such as searches and browsing behaviors within the catalog. The algorithms employed by big data can often create results for library users that are applicable to the user's tastes (such as with Netflix categories). For example, a big data approach could generate a list of comparable mystery novels for an avid reader. The disadvantage of big data within libraries in this case is that it is measuring each individual user and not necessarily the information they give (whether voluntarily or non-voluntarily). It does not answer a library user's specific questions, but rather, it tracks a user's behavior [11]. Conversely, this disadvantage could be advantageous depending on the information needs of library users. If a library user is looking at African American literature and authors, search results using big data may provide them with additional materials about African American literature or books written by African Americans. With the increasing use of technology and the online environment within the library setting, big data's benefits versus consequences need further examination.

Addressing Privacy-Related Concerns with Big Data

This section explores two ways that library information professionals can address privacy concerns with big data. It explores the datafication model [5] and resource description and access (RDA), as well as linked data technologies [11]. When using big data to interact with a system, rather than individuals, privacy issues can be mitigated.

Big data is a cause for concern regarding individuals' private information, as big data has emerged faster than guidelines concerning privacy can be established.

The datafication model helps address privacy concerns regarding big data because, unlike previous models, it focuses on the information that is created from big data analytics. Thus, returning to the Target example above, the collection of purchasing history was likely not invasive of the user's privacy – the invasion occurred when that information was analyzed and aggregated to produce new information (that the user was likely pregnant). By bringing a new focus to post-collection data analysis, the datafication model helps elucidate new privacy concerns. It gives us a new way to model privacy and consider the privacy implications of big data.

Resource Description and Access (RDA) and linked data can enhance the library's catalogs, although these approaches are currently controversial within the field of library and information science. RDA and linked data, as opposed to big data, are more closely reflective of the more traditional cataloging system. RDA assists catalogers in creating and encoding bibliographic relationships. The bibliographic relationships from RDA offers the ability to "link library catalogs to the standards of linked data" [11, p. 504]. Linked data projects "work to embed meaningful relationships in a purposeful way, thus enabling connections that reflect some level of systematic thought and consensus within and among domains of knowledge" (p. 504). Through using RDA and linked data technologies, the catalogers are able to connect similar pieces of information for the patrons. Unlike the Netflix-like algorithm, which would base suggestions on individual actions, RDA and linked data technologies help library patrons seek materials that are related without compromising privacy. Take for instance the LGBTQ library population mentioned earlier. When a patron searches for an article related to "coming out stories," she could theoretically be linked to other articles or resources with similar themes and authors. Maintaining up-to-date and useful linked data, however, would be an impossible task, as catalogers would have to continuously update the bibliographic relationships among materials [11].

Implications and Future Directions

Big data is a cause for concern regarding individuals' private information, as big data has emerged faster than guidelines concerning privacy can be established [1]. Some argue that big data is necessary for the improvement and development of libraries and other information organizations [1], while others agree that library professionals should find alternative ways to store and analyze data to improve services to library users [2]. The ISACA and the ALA would benefit from redefining the privacy clauses in their Code of Ethics to include information regarding privacy within an evolving technological environment, specifically addressing the datafication model's emphasis on post-collection data created by big data analytics.

Conclusion

This article examines how big data is applied in the corporate world, where predictive analytics are used to predict future needs or resources by individuals. It also highlights big data's usage as it applies to the library system. Predictive analytics and big data could both compromise an individual's privacy, especially when applied to inquiries made within the library setting. However, using big data to analyze systems, not individual library users, could be an effective application of this approach. The datafication model proposes focusing on ways to improve data analysis rather than data collection, and Resource Description and Access (RDA) and linked data propose a new way to examine the already obtained data to provide meaningful results and recommendations to library users. ■

Resources on following page

Resources Mentioned in the Article

[1] De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review, 65*(3), 122-135.

[2] Varnum, K. V. (2015). Editorial board thoughts: Library analytics and patron privacy. *Information Technology & Libraries, 34*(4), 2-4. doi:10.6017/ital.v34i4.9151

[3] ALA (2008). Code of ethics of the American library association. Retrieved from www.ala.org/advocacy/sites/ala.org.advocacy/files/content/proethics/codeofethics/Code%20of%20Ethics%20of%20the%20American%20Library%20Association.pdf

[4] ISACA (n.d.). Code of professional ethics. Retrieved from www.isaca.org/certification/code-of-professional-ethics/pages/default.aspx

[5] Mai, J. (2016). Big data privacy: The datafication of personal information. *Information Society, 32*(3), 192-199.

[6] Herther, N. H. (2014). Algorithms and big data. *Online Searcher, 38*(5), 50-55.

[7] Kramer, A. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *PNAS Proceedings of the National Academy of Sciences of the United States of America, 111*(24), 8788-8790.

[8] Panger, G. (2016). Reassessing the Facebook experiment: Critical thinking about the validity of big data research. *Information, Communication & Society, 19*(8), 1108-1126.

[9] Voosen, P. (2015). After Facebook fiasco, big-data researchers rethink ethics. *Chronicle of Higher Education, 61*(17), A14.

[10] Tattersall, A., & Grant, M. J. (2016). Big data: What is it and why it matters. *Health Information & Libraries Journal, 33*(2), 89-91.

[11] Campbell, D. G. & Cowan, S. R. (2016). The paradox of privacy: Revisiting a core library value in an age of big data and linked data. *Library Trends, 64*(3), 492-511.

[12] Jernigan, C., & Mistree, B. F. T. (2009). Gaydar: Facebook friendships expose sexual orientation. *First Monday 14*(7). Retrieved from http://firstmonday.org/article/view/2611/2302