Clemson University
TigerPrints

Publications | University Libraries

Winter 12-2014

The Library: Big Data's Boomtown

Rachel Jane Wittmann
Clemson University, rwittma@clemson.edu

Lauren Reinhalter
Pratt Institute, lareinhalter@gmail.com

Follow this and additional works at: https://tigerprints.clemson.edu/lib_pubs

 Part of the Library and Information Science Commons

Recommended Citation
Please use publisher's recommended citation.

This Article is brought to you for free and open access by the University Libraries at TigerPrints. It has been accepted for inclusion in Publications by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

To cite this article: Lauren Reinhalter & Rachel J. Wittmann (2014) The Library: Big Data's Boomtown, The Serials Librarian: From the Printed Page to the Digital Age, 67:4, 363-372, DOI:10.1080/0361526X.2014.915605

To link to this article: http://www.tandfonline.com/doi/pdf/10.1080/0361526X.2014.915605

The Library: Big Data's Boomtown.

LAUREN REINHALTER and RACHEL J. WITTMANN

Pratt Institute, New York, New York, USA

Since 2012, nearly every sector has developed a fascination with the seemingly new discovery of Big Data and its unprecedented capabilities to fuel analytic breakthroughs. It is clear that the use of Big Data as an information resource will continue to become more prevalent as it is employed in academic research and data-driven decision making, and even emerges as a vehicle for government transparency. This article reviews the emergence and potentials of Big Data, describes the policies fueling the current data surge, and discusses the impact on libraries. As libraries evolve to provide more data services, there is an opportunity for librarians to become experts and authorities in the data age.

KEYWORDS: Big Data, data curation, library education, data librarianship, Open Government, skills & training

INTRODUCTION

Starting in 2012, exalts of Big Data became ubiquitous in the headlines of newspapers, proprietary magazines, and academic journals. Nearly every sector developed a fascination with

the seemingly new discovery of Big Data and its unprecedented capabilities to fuel analytic breakthroughs. The concept of Big Data can be applied to all disciplines in widely varying capacities, and its sources and uses are myriad. The data fueling this surge is often collected from human behavior and tied to individuals, thus raising privacy concerns.[1] It is clear that the use of Big Data as an information resource will continue to become more prevalent as it is employed in academic research, data driven decision making, and even emerging as vehicle for government transparency.

SCOPE OF RESEARCH

As Big Data emerges as the basis of modern research, it has fittingly become increasingly present in research libraries, where data centers, equipped with data librarians (sometimes called data curators), are becoming more common. Librarians and information professionals have traditionally served as stewards of information access, research instructors, and privacy advocates. Librarians are poised, and arguably required, to be a vital part of the data era.[2] Within the recent tidal wave of available data resources, the librarian's role remains consistent and even has the potential to expand. To accommodate for the rising data demand, library and information science programs are adapting to include new curricula along with the fundamental skill set. Along with the specialty area of data librarianship, the broader information profession field must be aware of growing data-based research and the issues involved in its storage, processing, and use. In order to inform librarians on the emergence of Big Data and how the data surge is impacting librarianship, we will investigate the following questions:

- How is big data effecting libraries and how are librarians preparing for more data in (or as) information resources?

- How do librarians fit into the new reality of Big Data research? What aspects of Big Data impact the Library and how are librarians preparing for the new role and challenges posed by Big Data?

In order to better understand the environment of Big Data, we will first review the emergence of Big Data, and the potentials of unprecedented and innovative analysis. Next, we will describe the policies fueling the current data surge, including the government lead Open Data Initiative, and finally, discuss the impact on libraries.

LITERATURE REVIEW

1. Big Data Defined

The abundance of Big Data is a product of data being generated constantly, automatically, and rapidly. To credit one source of the data surge, improvements in technology have resulted in affordable computer devices and the prevalence of internet access, connecting more and more people to data-collecting entities. The statistics on the pace of data production are staggering. As of 2012, 90% of the world's data had been generated within the past two years.[3] However, Big Data is much more complex than just massive amounts of data. Big Data can be characterized by the three V's: volume, velocity, and variety.[4] To illustrate the volume of data at hand, one petabyte is equal to the amount of text in 20 million filing cabinets. In the time span of fifteen minutes, humans will have produced about 20 petabytes of data.[5] The rapid speed at which data are created adds to its dynamic capabilities. For example, a poster-child for Big Data projects, the Google Flu Trends website is able to source searches for flu related illnesses, thus tracking geo-specific flu outbreaks in real-time; much faster than the Center for Disease Control. Velocity, however, should not be considered as solely speed, but the constant evolution of a dataset.[6] The variety of data sources, especially those that can be combined for complex analysis,

is another key feature of Big Data. The sources of Big Data are widespread, from weather reports to traffic sensors and electricity grids, and are largely generated by people conducting routine activities. Due to the omnipresence of personal computing devices and smart-phones, individuals generate massive amounts of data through e-commerce, e-communication, social media, online searches, and GPS navigation.

2. Big Potentials

Big Data has been hailed as "humanity's dashboard" by Rick Smolan and Jennifer Erwitt, the primary authors of the illustrative book *The Human Face of Big Data.*[7] Smolan and Erwitt, along with other contributing writers and photographers, review the ways in which humanity is able to measure and analyze the constant flow of data being captured through sensors, satellites, and GPS enabled devices.[8] Big Data has also been referred to as the gold rush of our time. The analogy of prospectors flocking in droves to seek quick fortune, compares to the modern day rush of private and public sectors seeking potential breakthroughs found in Big Data. Innovations in technology have made the massive accrual of data possible, along with more access through Big Data management software tools. During the gold rush, stories of easily found gold nuggets quickly evolved to a more laborious and tactical mining operation. Similarly, Big Data touts progressive jackpots but also requires sifting through quarries of mundane or chaff data.[9] Both of these analogies are accurate descriptors of a new age of data and the widespread eagerness to take advantage of its possibilities. While Big Data lends itself to an array of metaphors, it is rather its very reality that has catapulted it to prominence.

Having taken over headlines in 2012 and 2013, one might ask why the sudden spotlight on Big Data? Data has long been used in research and online marketing but Big Data is not just about targeting customers online with eerily accurate, and often intrusive, advertisements based

on web viewing habits. There are non-commercial benefits to Big Data in the fields of medicine, environmental science, and urban planning, to name a few, which celebrate unprecedented potentials. Urban metropolises have begun by analyzing data on energy and water usage, waste, traffic, pollution, noise complaints and crime statistics, to optimize efficiencies and thus, become "smart cities."[10] New York University's Center for Urban Science and Progress (CUSP), is an academic center founded on using Big Data to make New York City more efficient. Datum is being collected by automated systems such as wireless sensors, noise meters, and the City's 311 non-emergency service. Additionally, the New York City government will be giving CUSP all of its public data in the hopes of improving inadequacies.[11]

The public and private health sectors hold some of the highest potential to be revolutionized by Big Data. The South Asia Institute at Harvard University has deployed a team of doctors and researchers to the religious festival called Kumbh Mela to create the largest public health dataset of a transient population. The goal of the project is to use clinic data to predict outbreaks and epidemics. For instance, with the data they have already collected, certain diseases can be traced to populations near riverbeds.[12] While Big Data's benefits to populations are apparent, the individual can use their own data to monitor themselves. With widespread ownership of smart-phones, self-tracking applications have flourished. Diet, exercise, and schedulers are some of the most common self-trackers.[13] According to Stephen Baker in *The Numerati,* "constant monitoring is bound to change the very nature of healthcare," by shifting the focus from "critical response to prevention."[14] Putting data analytics in the hands of, and for the benefit of, the individual has the potential to educate the public on how similarly, albeit on a different scope and scale, Big Data can improve quality of life.

3. Open Government Data

The rise in publicly available data resources is not a coincidental trend growing in parallel to the massive data phenomenon; it has been lead, in part, by a government mandate. When Barack Obama entered his first term as President, he immediately moved toward a more transparent, participatory, and collaborative government. The memorandum titled "Transparency and Open Government," issued in January 2009, declared the Obama Administration's commitment to an unprecedented level of openness intended to bolster democracy and efficiency of the Government.[15] In the beginning of President Obama's second term, data was targeted to further enable this pledge to openness. A May 2013 Executive Order expanded the government's data policy to default to open government data (OGD) to not only "strengthen democracy," but also "contribute to economic growth" by spurring scientific and business innovation.[16] The Order specified the timeline for various agencies to adopt this new default state, assess and expand data processes, and also ensure that individual privacy is safeguarded. The corresponding memorandum, "Open Data Policy-Managing Information as an Asset," outlines the requirements for federal agencies "to collect or create information in a way that supports downstream information processing and dissemination activities".[17] Under the mandate, every agency must host a www.[agency].gov/data URL listing the available datasets in both human-readable and machine-readable formats.

The coming flood of open and accessible government data holds immense potential for researchers, public sector industries, lay citizens, and also for librarians. Shortly after the "Open Data Policy" was released, data.gov launched a data catalog, powered by open source software CKAN, to provide access to the immense array of data available. The CKAN catalog is integral to the success of the Open Data Policy as it goes beyond organization and access to also harvest data inventories from federal agencies to increase availability.[18] At the time of writing, there are

over 88,400 datasets available on data.gov and this number will only continue to expand. In fact, citizens can directly request a specific dataset to be released. In an attempt to research the economic potential of OGD, the Governance Lab (GovLab) research institute at New York University recently developed Open Data 500, the first initiative to study U.S. companies that use OGD in business and product development.[19] Researchers, too, are discovering a wealth of newly available science and engineering data, collected on a scale that only the federal government could possibly fund.[20]

4. Big Data Policies

While the trove of OGD is a welcome development, librarians are especially aware of and uniquely poised to cope with the demanding policy concerns facing a data catalog as large as data.gov. The Office of Management and Budget, along with the Office of Science and Technology Policy, launched Project Open Data, an online resource of standards, best practices, and schemas to guide agencies in adopting the new framework.[21] Without proper funding and data curation expertise, however, agencies may not meet data management requirement specifications. Government information librarians, such as James A. Jacobs of The University of California, San Diego and James R. Jacobs of Stanford University, cite concerns about the long-term preservation of and access to OGD, pointing out that data.gov does not comply with Open Archival Information System requirements and is not yet a Trusted Digital Repository. Libraries have already been grappling with data curation and, according to James R. Jacobs, "could be vital partners in terms of metadata standards, metadata creation, preservation, and managing the whole information life cycle."[22]

In addition to librarians' concerns with the curation and preservation of OGD, is the burden of protecting individual privacy during analysis and use of this data. Librarians have long

been advocates for privacy— a role that increased with the adoption of the Patriot Act in 2001— educating users about informed consent and adapting practices to protect the privacy of patrons. While the new Open Data Policy requires agencies to incorporate privacy analysis in every aspect of data collection and dissemination, according to Jeannette Woodward, "not only are the procedures for collecting and using data flawed, but the technology itself often fails."[23] In fact, as the size of a dataset increases, privacy decreases such that identities can be triangulated to a point of incredible probability.[24] OGD privacy management concerns, in conjunction with the June 2013 realization of the NSA PRISM data collection program, are pushing librarians to become even stronger defenders of privacy.[25] The current level of government data collection is unprecedented, but for librarians already concerned with privacy, navigating this brave new world is simply an extension of the work that has been going on for decades.

5. Big Data in the Library

The trend of data-fueled research is ubiquitous in all sectors, creating the opportunity for librarians to collaborate with other disciplines to fill a service gap. Libraries, the bastions of information, are adapting to accommodate for the growth of data resources and provide data literacy instruction. Academic and research libraries have been most impacted by this evolution; many federally funded research grants have recently required data management plans (DMP) as part of the application process. It is largely the librarians' role to create and provide data management services that meet this new grant specification. In addition to providing DMP guidance, libraries have entered into the data lifecycle by taking on the task of housing and preserving the data generated from student and faculty research projects in data repositories.[26]

The development of data resources has led some libraries to incorporate specialized data departments. In the article "The Emerging Role of Libraries in Data Curation and E-Science,"

Byron Heirdron points out that libraries must become curators of digital data in order to adhere to their core mission: to protect and disseminate information.[27] Because grantmaking agencies require both DMPs and plans for data maintenance post-research, university libraries are quickly adapting. In 2011, the National Science Foundation (NSF) began to require DMPs with all grant proposals and many other funding institutions have followed suit.[28] With competition for funding increasing, a well formulated DMP could increase award chances. Many scholars, however, are unprepared to create such proposals, manage, and preserve the data accumulated through their research. To meet this demand, major research libraries are creating data service departments, such as New York University (NYU), Cornell University, Massachusetts Institute of Technology (MIT), and John Hopkins University, to name a few.[29] With the growing incorporation of data services within libraries, there is a need for data librarians to manage these centers. Laura Gordon-Murname specifically addresses the rise in big data and the emergence of data librarians in her article "Big Data: A Big Opportunity for Librarians." Gordon-Murname is optimistic that the data explosion means a new job market for librarians. As she points out, librarians are needed in four key areas: (1) organization, (2) search and access of internal datasets, (3) awareness of external data sources, and (4) to serve as authorities on copyright and intellectual property issues.[30] To equip librarians with these essential data-specific skills, "data curation" programs are becoming common in library and information schools.[31] Data librarians serve as resources for instruction on data discovery, data management plans, data analysis, and visualization tools such as geographic information systems (GIS).

Some academic libraries have elected to take a more active role in data management. The case study of Purdue University's development of data the repository, Purdue University Research Repository (PURR), demonstrates the libraries involvement in creating a solution to

the data needs of its researchers. The library joined forces with information technology and research departments to create PURR. The library, being the forefront of data instruction and reference, while also being knowledgeable on metadata standards, was a critical leader in the development process.[32] The data repository is an exemplary way for libraries to provide data service to its patrons while also exhibiting control over data produced by the academy.

The burgeoning field for data services in libraries, along with expanded career options for information professionals, indicates that new skills and education are required for such data specific roles. While these new roles demand improved technical intelligence, it is also imperative to view data as a collection to be incorporated it into the library's cadre of resources. Noting this challenge, Dr. Deborah Rabina, from Pratt Institute's School of Information and Library Sciences states, "I would like to see LIS education graduate students who have not only the technological skills, but, more important, the policy perspective that views data as a collection."[33] It is in the middle ground between programmers, statisticians, and data scientists, where the librarian's skills must be developed in data reference and data curation. Librarians must take a lead in data science education to prevent data specialty positions within libraries from being sourced from other professional backgrounds.[34] As data is information at its most basic, this is not that far of a stretch. In seeking further data education, librarians will be poised to succeed in the data age.

Terence Huwe, Director of Library and Information Resources at University of California, Berkeley's Institute for Research on Labor and Employment, suggests that not only should librarians embrace managing Big Data for career opportunities, but also use it to benefit our own institutions. While many are gleaning ground-breaking analysis from data, so, too, can libraries, especially in harnessing predictive awareness for library-managed areas of collection

development, instruction, and reference. Information professionals can solidify their importance by becoming active developers within the academy by providing predictive data relating to education trends in the ever changing landscape of future universities.[35] Circulation and collections data can be analyzed to aid in highly accurate weeding and serials cancellation decisions.[36] Increasingly, academic libraries are adopting the business world's return on investment (ROI) techniques to illustrate their value through metrics and quantify their achievements.[37]

CONCLUSION

While the capabilities of Big Data are just being realized, its possibilities have captured the attention of the library world. When it comes to scientific research, librarians can fill a service gap by enforcing standards and best practices and providing guidance on the creation of DMPs. Another possible point of collaboration is librarians' ability to create trustworthy data repositories. The proliferation of data in research are undoubtedly influencing the information profession and providing career opportunities. Librarians will rise to these new challenges, as to all others, by adapting to new technology and staying abreast of the latest trends in research. In the near future "data reference" will be included in general reference and datasets not given special attention. Adding to this forecast, data curation curriculum will not only be provided as an option, but one day become core requirements to library and information science degrees. After all, data is the rawest form of information; the data scientist is the closest relative to the information scientist.

References

[1] Adam McAfee and Erik Brynjolfsson, "Big Data: The Management Revolution," *Harvard Business Review* 90 no.10 (2012): 59-68.

[2] Laura Gordon-Murnane, "Big Data: A Big Opportunity for Librarians," *Online* 36 no. 5 (2012): 34.

[3] MaryAnne Gobble, "Big Data: The Next Big Thing in Innovation*" Research-Technology Management* 56 no.1 (2013, January-February): 64. Retrieved from: www.gale.cengage.com/AcademicOneFile/.

[4] McAfee and Brynjolfsson, "Big Data: The Management Revolution," 59-68.

[5] Juan Enriquez, "Reflections in a Digital Mirror," *The Human Face of Big Data* (Sausalito: Against All Odds Productions, 2012), 18.

[6] Mayer- Schönberger and Cukier, *Big Data: A Revolution that Will Transform How We Live, Work, and Think.*

[7] Steve Lohr, "The Age of Big Data," *The New York Times*, February 11, 2012, http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

[8] Rick Smolan and Jennifer Erwitt, *The Human Face of Big Data* (Sausalito: Against All Odds Productions, 2012).

[9] Brad Peters, "The Big Data Gold Rush." *Forbes Magazine*, June 21, 2012, http://www.forbes.com/sites/bradpeters/2012/06/21/the-big-data-gold-rush/.

[10] Steve Lohr, "SimCity, For real: Measuring an Untidy Metropolis." *The New York Times* (2013, February 23). Retrieved from: http://www.nytimes.com/2013/02/24/technology/nyu-center-develops-a-science-of-cities.html

[11] Steven E. Koonin, "Center for Urban Science and Progress: The Promise of Urban Informatics," CUSP, May 30, 2013. http://cusp.nyu.edu/wp-content/uploads/2013/07/CUSP-overview-May-30-2013.pdf, 15.

[12] Somini Sengupta, "Can big data from epic Indian pilgrimage help save lives?" *The New York Times*, February 8, 2013, http://bits.blogs.nytimes.com/2013/02/08/can-big-data-from-epic-indian-pilgrimage-help-save-lives/.

[13] A.J. Jacobs, "Quantifying myself," *The Human Face of Big Data*, (Sausalito: Against All Odds Productions, 2013), 54-57.

[14] Steven Baker, *The Numerati* (New York: Houghton Mifflin Company, 2008), 160.

[15] Transparency and Open Government [memorandum], 47 Fed. Reg. 4685 (Jan.21, 2009).

[16] Exec. Order 13642, 78 Fed. Reg. 28111 (May 9, 2013).

[17] Office of Management and Budget, Executive Office of the President "Open Data Policy-Managing Information as an Asset," (Washington, DC, May 9, 2013), 1.

[18] Meredith Schwartz, "What Government big Data May Mean For Libraries," *Library Journal*, May 30, 2013, http://lj.libraryjournal.com/2013/05/oa/what-governmental-big-data-may-mean-for-libraries/.

[19] Nick Sinai and Erie Meyer, "Open Data in Action," White House.gov, January 8, 2014, http://www.whitehouse.gov/blog/2013/12/17/open-data-action.

[20] Data.gov, "Research.gov," May 23, 2013, https://www.data.gov/research/research-gov/.

[21] Office of Management and Budget, "Open Data Policy-Managing Information as an Asset," 5.

[22] Meredith Schwartz, "What Government big Data May Mean For Libraries."

[23] Jeannette Woodward, *What Every Librarian Should Know About Electronic Privacy,* (Westport: Libraries Unlimited, 2007), 88.
[24] Mayer-Schönberger and Cukier, *Big Data: A Revolution that Will Transform How We Live, Work, and Think,* 153.
[25] Glenn Greenwald and Ewen MacAskill, "NSA Prism program taps in to user data of Apple, Google and others," *The Guardian*, June 6, 2013, http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data.
[26] Laura Krier and Carly A. Strasser, *Data Management for Libraries* (Chicago: ALA TechSource, 2014), 10.
[27] Bryan Heirdorn, "The Emerging Role of Libraries in Data Curation and E-Science," *Journal of Library Administration* 51 issue 7-8 (2013): 662-667.
[28] Krier and Strasser, *Data Management for Libraries,* 4.
[29] Heirdorn, "The Emerging Role of Libraries in Data Curation and E-Science," 662-667.
[30] Gordon-Murnane, "Big Data: A Big Opportunity for Librarians," 32.
[31] Ibid., 33.
[32] Michael Witt, "Co-designing, Co-developing, and Co-implementing an Institutional Repository Service," *Journal of Library Administration* 52, no. 2 (2012, March 1): 172-188.
[33] Meredith Schwartz, "What Government big Data May Mean For Libraries," 1.
[34] Jeffrey Stanton, "Big Data and the Library Professional." *Journal of the Library Administration and Management Section* 8 no 2. (2012, May): 22-24.
[35] Terence K. Huwe, "Big Data, Big Future," *Computers in Libraries* 32 no.5 (2012, June): 20-22. *Library Literature & Information Science Full Text (H.W. Wilson)*, EBSCO*host.*
[36] Jeanne M. Brown and Eva D. Stowers, "Use of Data in Collections Work: An Exploratory Survey," *Collection Management* 38 no. 2 (March 12, 2013): 143-145. DOI: 10.1080/01462679.2013.763742.
[37] Betsy Kelly, Claire Hamasu and Barbara Jones, "Applying Return on Investment (ROI) in Libraries," *Journal of Library Administration* 52 no. 8 (January 15, 2014): 656-658. DOI:10.1080/01930826.2012.747383.