Re-examining ‘Library Futures’: Data, Big Data, and Ethical Innovation

Andrew Weiss

California State University, Northridge

Abstract

I look at the malleable concept of data, explore big data and its impact on patrons and librarians, and examine the impact of surveillance on privacy. It's a wide-open topic that touches on numerous aspects of "New" librarianship, including dealing with future changes to libraries' infrastructures, their foundational ethical philosophies, and their potential possibilities as hubs of innovation (i.e. as maker spaces, research incubators, future open access publishers, information brokers, etc.).

Keywords: *data; big data; privacy; surveillance; library assessment; student learning; open access; innovation; information overload; digitization*

Introduction: The 'Datalyzation' of the World

"*We need the languages of both science and poetry to save us from merely stockpiling endless 'information' that fails to inform our ignorance or our irresponsibility." -- Ursula K. Le Guin*

What is data? It's a simple question with a simple answer, until you start looking at it more closely. When most of us in the library world hear the word data, we likely envision digital information, perhaps in the form of endless strings of zeroes and ones, massive sets of real numbers, or texts that provide endless possibilities for mining information. When repository managers talk about archiving data they often mean primarily the digital-numerical type gathered in sets or spreadsheets or zip folders associated with an experiment or grant-funded research project. Sometimes there's a computer program included to help run through the 'raw' numbers, but it's still limited to digital information. This is the kind of data that you can 'crunch' or 'parse' or 'search' or 'mine'. It's easy to conceptualize data like this for most disciplines – and assume that this is the bulk of what comprises data – especially now that computational studies are all the rage, including but not limited to the STEM fields, digital humanities, sociology, and, of course, computer science.

From this perspective, it seems that any discipline can be '*datalyzed*' (i.e. subject to data analysis), even those such as literature that have historically resisted attempts at quantification (cf. Google's *Ngram Viewer* in image 1 below). Many of us working in libraries for the past decade have also had a direct hand in the digitization of books, photographs, and archival collections. The *datalyzed* world is in some ways the product of our own ambitions and visions for world-wide access and open information, the 21st century's version of the universal library. (For a good early discussion of the pitfalls of this ambition read Jean-Noel Jeanneney's 2007 book *Google and the Myth of Universal Knowledge*.)

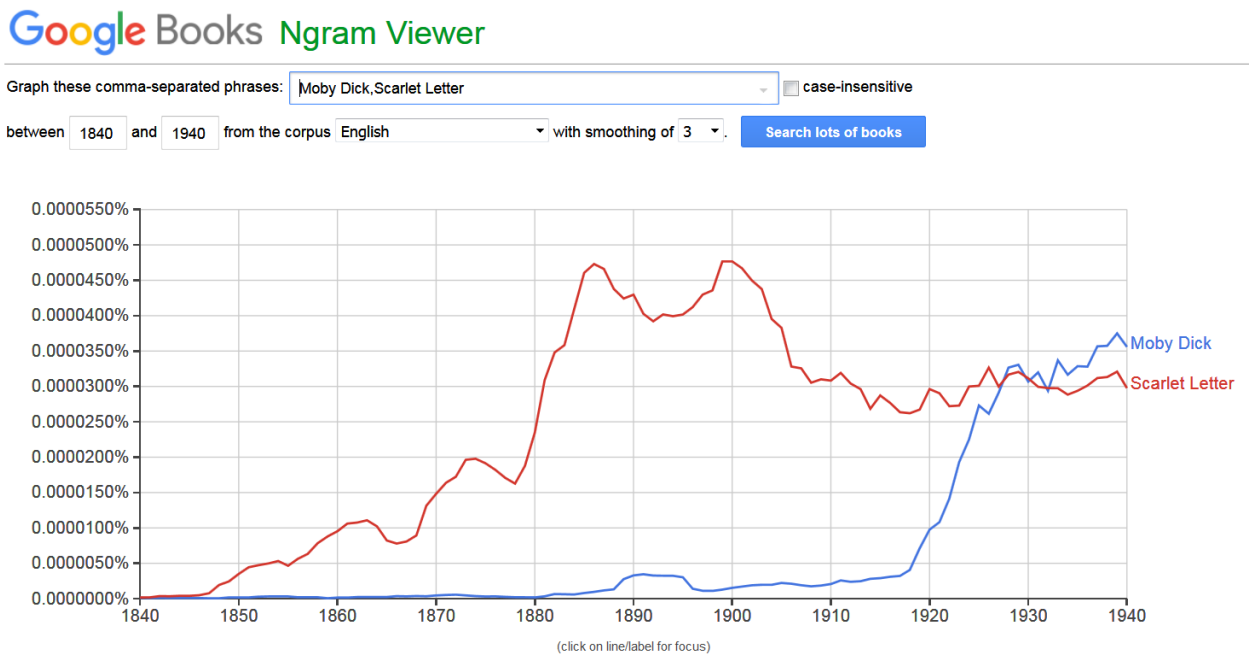

Image 1: Google's Ngram viewer showing the frequency of terms <Scarlet Letter> compared to <Moby Dick> between the years 1840-1940 found in the massive digital library Google Books.

This transformation spans not only the breadth of all types of disciplines, but renders their intellectual depths and their temporal limits available to data analysis as well. "Big data," as it is also known, contributes to this sense that information and data are primarily comprised of digitized numbers persistently available to anyone at any time in any format regardless of laws, regulations, or concerns for personal privacy and individual limits. The concept is further defined by its rapidly accelerating and expanding '5Vs': *Volume, Variety, Velocity, Variability, Veracity*. With Facebook, Google, Twitter, Snapchat and any number of lesser-known IT and social media companies tracking your every move online (cf. Surveillance Capitalism), it's hard to see data as anything but massive amounts of quantified information coming at us at lightning speed used to predict or nudge human behavior in the name of increased profits. Platform hipsters touting Foursquare, QR codes, Snapchat,

or the latest social media craze beware: the speed of change outstrips us all. Prognosticate at your own peril.

Welcome to the datalyzation of the world, we might say. As our lives have moved to the online environment, where everything is by definition converted into a digital format, this narrow conception of data begins to elbow out all others. And it only seems to be the beginning. Wait until your toaster is online, too, and it is able to tell its manufacturer how many bagels you ate last week or, worse, all of last year. Hopefully they won't share this information with your health care provider. (But I get ahead of myself.)

On the positive side, however, this move toward the datalyzation of scholarly disciplines and life in general has provided us with an *explosion* of innovation. The potential to harness large amounts of data for good *does* exist in a large swathe of our society. Martin Hilbert at UC Davis provides us with a clear roadmap toward e-democracy, where these digital tools, methods, and conceptual frameworks are used to improve lives through open science, open access to information, and an overall informed citizenry. (Hilbert, 2018) Big data tracking can provide us with predictive models that curb crime or alleviate traffic jams, allow us to prepare for and avoid pandemics, and help us create more effective urban spaces. The Internet of Things (IoT) promises to combine data analysis with everyday human activities for more accurate predictions of individual and collective behaviors; that approach might, for example, improve education while potentially making it less labor-intensive and hence more cost-effective. All this from the collection of *zeroes* and *ones*!

Here Comes the Sun

There's more to it, of course. And here's where our definition of data gets a little slippery.

Data, as we all know, is more than numbers. Qualitative data gathers the words, expressions, and actions of people, their stated beliefs and customs, and their conscious (or unconscious)

behaviors. Evidence gathered during an experiment through personal observations or expressions of subjective feeling become data used for drawing up theories and testing hypotheses. Yet even scientific observation doesn't rely on raw numbers alone. And neither can those numbers describe all things. One Center on my campus at California State University, Northridge (CSUN) studies, well, the *sun*, taking pictures of it daily and gathering information from their instruments attuned to it. The photographs date back about a dozen years, and as digital information -- as un-rendered zeroes and ones -- they're rather useless. But taken together as a series of *human-readable* photographs and synchronized with other calculations, observations, and instrumentation, they become essential documents about the sun at certain time in its history. These are as essential as the observations originally hand-drawn by Galileo through the lens of a telescope 400 years ago. Though conceivably just one step removed from artistic expression, in the right context photographs, drawings, and diagrams are data.

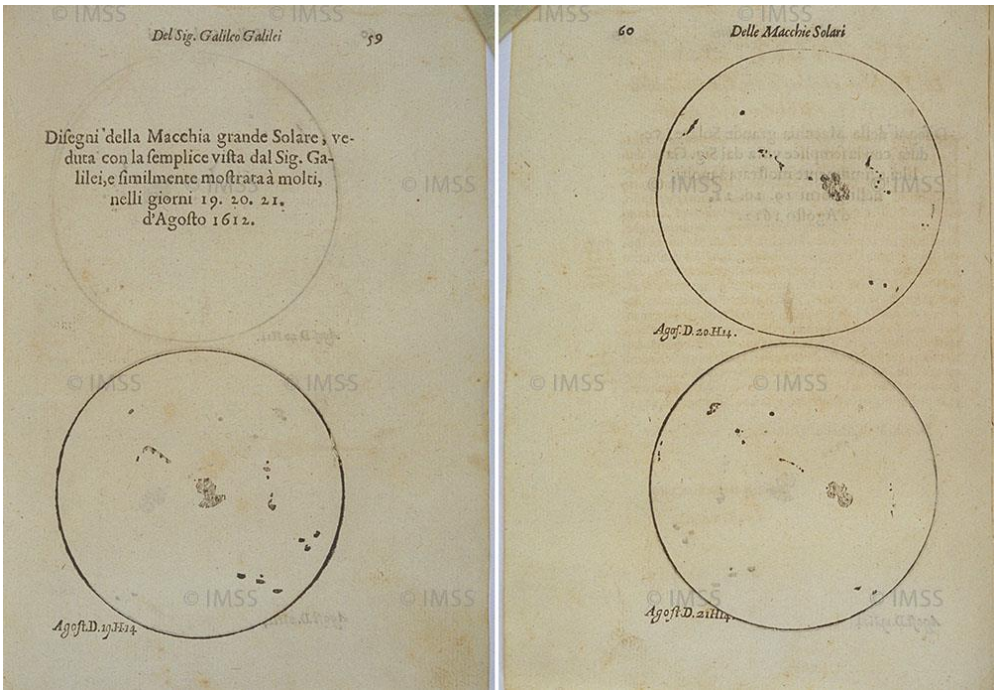

Image 2: Galileo's drawings depicting the observations of sun-spots. Istoria e dimostrazioni intorno alle macchie solari, Roma, 1613. *Available at this url: https://brunelleschi.imss.fi.it/galileopalazzostrozzi/oggetto/GalileoGalileiIstoriaDimostrazioniIntornoMacchieSolari.html*

And that's where the concept of data starts to bleed into something different than just numbers. To stretch this idea even further, is the Polaroid photography that Robert Mapplethorpe took of *New York's S&M subculture* in the 1970s just art or is it also something more? Though documentary in nature, viewers nevertheless experience these photographs in terms of artistic expression. But researchers and scholars find invaluable evidence that helps them describe and understand the time period and the subculture. To them, the photograph becomes more than just a minor curiosity or footnote in a well-known artist's *oeuvre*, it *becomes* important data, capable of explaining theories and informing historical analyses.

Boxing and Babbling

It's quite difficult to resolve this conundrum until, that is, we begin to think of data in a much different way. Back to our original question, but modified: *What is data, then, if it's so malleable, so protean, so slippery, and so inconclusive?* Christine Borgman in her 2015 book *Big Data, Little Data, No Data* argues that "conceptualizing something as data…is a scholarly act." Devoid of an essence of its own, data is instead context-dependent and changeable depending on the purpose and needs of the researcher working within the confines of a specific discipline and scholarly tradition. In other words, it is an assertion about our constructed reality, an assumption that the phenomena of our world (and beyond) can be described, analyzed and ultimately understood in ways that transcend gut-feelings, intuitions, or beliefs.

Why is this important? When we begin to discuss methods of data collection and data management, invariably the discussion becomes dominated by data of the quantitative, digital kind. But if data is often something imprecise whose lines of distinction can easily bleed into others, then as information managers using an *incomplete* conception of data, we may be fighting this fight with one hand tied behind our backs.

Image 3: Library of Babel online. "At present it contains all possible pages of 3200 characters, about 10^{4677} *books." The above page can found here:* https://libraryofbabel.info/anglishize.cgi?1-w1-s2-v18:1

And watch out for that left hook. It's a *doozy.* For as the tide of digital data rises, threatening to drown us in a (sometimes disorganized) glut of bits and bytes or weaponized falseness (including fake

news, misinformation, and disinformation), we are almost in nearly the same dire straits as the librarians in Borges' Library of Babel. *Spoiler alert*: It doesn't end well for most of them, driven, as they are, to despair, madness, and suicide at the incomprehensibility of the thing they are tasked with somehow managing.

Some of this is a natural response to information overload, one of the major downsides of the online data-driven world that has surprisingly been with us for millennia. Anyone interested in the long-documented impact that information overload has had on people would be well advised to check out Ann Blair's research on the history of the concept, which she traces back to classical antiquity. Of course, the online world's mantra of "all data, all the time" exacerbates this ancient problem. Ironically, she also sees the establishment of libraries as a symptom of and contributor to overload, asserting that libraries themselves instill the desire to collect everything. (Blair, 2010)

This may be quite true to a certain extent. As information professionals, we deal with problems of information overload daily, especially when advising students to "narrow down" their research, but not -- I repeat -- **not** to rely on the first few hits in a Google search. Yet, when we do this we inadvertently acknowledge the impact that too much information is having on people, while simultaneously ignoring our own implicit roles in the process. Students stop at the first few hits in Google (or in our catalogs) because to go on further in the face of seemingly-infinite information glut is a form of madness. They know this, instinctively. They are coping with overload in a natural – and quite reasonable – way. *Go for good enough and get out,* they are thinking. It should be noted, too, that only fighters head *toward* their opponents. Most of us learn to get out of the way: *Duck*. *Cover*. *Get out of the ring*. By the same token, it's really only librarians that crave the headlong search into the information jungle. Perhaps we want to show off our hard-won skills in a nod to professional pride. Our users, though, prefer to *find* something, anything, and then move on – preferably *unscathed*.

Private Eyes are Watching Me: "Data is Destiny"

Can't escape the '80s these days. From pop culture references in film to the persistence of discredited economic ideas, the 1980s seem to be back in full force. You can't escape the prying eyes of the internet, either. 1984 was a pretty good year for Eddie Van Halen, I suppose; not so great for those in Orwell's dystopia. The main distinction, though, between the surveillance and privacy invasion in Orwell's book and the one in real life is that most of us have run toward it with open arms, actively ignoring boilerplate user agreements or tome-length privacy policies (and, for that matter, weakening net neutrality and privacy regulations) that have allowed the tech companies to surveil you or to sell your information to third parties. It didn't have to be inevitable or destined, but it sure seems that way in retrospect.

Glen Greenwald's book *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State* paints a damning picture of surveillance gone amok. Aside from the frightening scale and ambition of these surveillance programs, one of the most striking aspects of the revelations in his book is the following statistic: in 2013, 75% of all U.S Internet traffic had the potential to be monitored by the NSA. (Greenwald, 2014, p. 99) I don't know about the remaining 25%, but I suspect someone else was probably watching that too. In the five years since, it is hard to imagine that number decreasing.

Greenwald's book also portrays a tech industry cooperating quite readily with or working as third-party contractors for these Federal agencies. This includes the big names: Apple, Microsoft, and Google, and some lesser-known but major-player private contractors like Stratfor and Booz Allen. In this context, where then-NSA director Keith B. Alexander has stated that he wants to *literally* "Collect it all" (Greenwald, 2014, p.95), the recent developments about Facebook's mishandling of data with *Cambridge Analytica* seem downright miniscule, despite the shocking headlines. 50 million surveilled Facebook fans couldn't possibly be wrong, could they? Only, I'd argue, if they're outweighed by 7

billion more. Or, as Peter Waldman, Lizette Chapman, and Jordan Robertson state in their excellent piece on *Palantir Technologies*, another big data company that has recently implemented a preventative crime database with the LAPD, "when whole communities…are algorithmically scraped for pre-crime suspects," as is being done in Los Angeles, "*data is destiny*." There may be no exit from your fated judgement.

Whither Assessment? The FANGS are Out

It all points to this: *we must tread lightly*. If big data and the Facebooks / Amazons / Netflixes / Googles (FANGs) of the world are the titans smashing whole industries to pieces and recreating them in their own images, where does that leave libraries? We can partner with these online content and social media providers, but that means we run the risk of breaking our codes of ethics and privacy policies. We may also not yet be aware of how pervasively we have been compromised, making us unwitting accomplices to the surveillance state. We may easily find ourselves snake-bitten. Libraries can't escape their associations with the dominant cultures and power structures that fund them. That makes them targets, as Rebecca Knuth's (2003) excellent book *Libricide: The Regime-Sponsored Destruction of Books and Libraries in the Twentieth Century* so pointedly demonstrates.

Student/patron assessment, though, is still necessary for educational purposes, and the "holy grail," as it were, of academic library assessment is finding direct links between a specific student's library use and his or her actual grades (hopefully they're *good* ones and correlations of the *positive* kind!). Of course, that's likely impossible in current conditions and limitations (pre-*IoT*). It's too invasive, requiring a much more significant amount of tracking student behaviors and activities. At least to my taste. But back in 2005 when Google Maps first arrived (and later Google Street View) I was similarly disturbed by it. Now I use it all the time, conveniently ignoring the fact that the archive of information Google has amassed on me includes all manner of locations, destinations, login times, and

personal inquiries, likely pinpointing my existence between two points *on any given day* in the past 10-12 years. (I'm nothing if not predictable.) In this climate of near-total surveillance, though, libraries may suffer a real hit to their reputations. At the same time, if we do not take advantage of the available data out there about students, we may also fall into irrelevance. Some days it's hard to decide which is worse.

But the point of the post is not to dwell *too much* on the negatives. They are huge, yes. But there are some bright spots too. Libraries are hubs of innovation. Universities are quantified engines of economic growth. Open access to articles and data sets increases citations. Open access to information spurs new ideas; it also burnishes faculty and scholar reputations. ORCID and other linked data projects help make the world a more distinct, yet smaller and more civilized place. Data and assessment help us to better serve our users, and the more 'granular' it is, the better we can tailor those services.

Librarians have additionally transformed their collections and moved away from glorified document dumps and book warehouses, "Just in case," to become stewards and servers of communities, which grow wider and *on demand* as they link up together. Providing research commons, maker spaces, or equipment checkouts shows us that libraries can still meet the needs of a changing technological world. Providing new services informed by assessment and measurement shows us that libraries can effectively improve the lives of individual users both one at a time *and* at scale. We are nothing if not adaptable. Like data itself, we too change to meet the needs and demands of our contexts.

Conclusions: Data doesn't have to be Destiny

So, yeah, it's not all pretty. But it's not supposed to be all pretty when you're fighting for your principles, or when you're trying to discover something new. As librarians we are constantly on the

edge, balancing the privacy of users with the demands of our service and organizations. We need vigilance. We need new and stronger regulations of privacy that withstand the onslaught of big data and the coming *Internet of Things*. Libraries are only as good as the trust they generate in their users. If libraries are seen as just another cog in the surveillance state, libraries *will* wither and die. People will go underground for their information needs.

But it's not all bad, either. If libraries have proven anything in the past few hundred years, it is that they can handle information responsibly. More recently, they can also manage data and privacy responsibly. Clear innovation in organizing our most complex societal systems is happening with the gathering and combining of data that to this point had remained isolated. Mash-ups are incredibly effective, providing new contexts for ideas and new avenues of creativity for people. Libraries can facilitate this. But maintaining these institutions, and avoiding the fates of past destroyed libraries, will require sustained levels of trust. We must show people in the era of big data that privacy remains a priority. Ultimately, data is many things: a tool, a method of research, an assertion about reality, a method of explication, a means of surveillance and safety, and a means for finding truth and transparency. Whether it's used for good or ill, whether it's applied to all evenly and ethically is not automatic. It is up to *us* to do this.

References

Blair, A. (2010). *Too much to know: Managing scholarly information before the modern age*. New Haven, CT: Yale University Press.

Borgman, C.L. (2015). *Big data, little data, no data: scholarship in the networked world*. Cambridge, MA: The MIT Press.

Greenwald, G. (2014). *No place to hide: Edward Snowden, the NSA, and the U.S. surveillance state.* New York, NY: Metropolitan Books/Henry Holt.

Hilbert, M. (2018). *E-Democracy.* http://www.martinhilbert.net/category/research/e-democracy/

Knuth, R. (2003). *Libricide: The regime-sponsored destruction of books and libraries in the twentieth century.* Westport, CT: Praeger.

Waldman, P., Chapman, L., & Robertson, J. (2018). Palantir knows everything about you. *Bloomberg Businessweek*. https://www.bloomberg.com/features/2018-palantir-peter-thiel/