WHITHER LIBRARIES?

Publishing, Data, and the New Literary Arts

By Ben Johnson

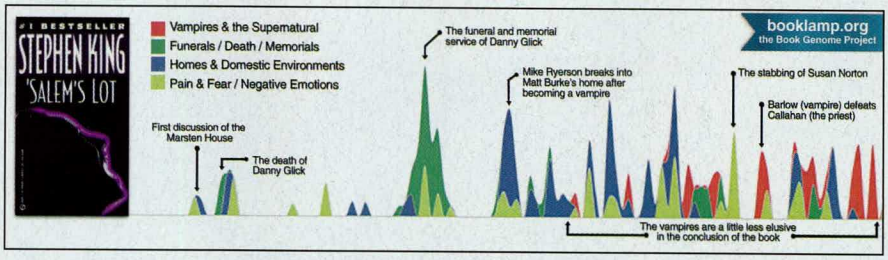
To remain relevant in the long term, libraries will need to leverage the power of data.

Have you ever wondered why you are so addicted to that game on your smartphone? Or why you can't stop binge-watching every episode of that Netflix show? Those addictive qualities are not accidents. They are engineered.

Before that media ever crosses your screen, it is rigorously tested and tweaked. If you are helplessly addicted, it is because the entertainment industry is quite adept at crafting hooking media. Movies are screened by test audiences, and scenes are edited or rewritten based on feedback. Data from video game controls is analyzed—every button pressed and every candy crushed—and used to cultivate those addictive qualities that keep you playing a few seconds longer.

But for the publishing industry, the only measures of effectiveness and reader satisfaction are sales and reviews. Both of these are lagging indicators, generated after a book has been released. Neither metric can help shape nor forecast the next big book—they are tools of dissection, not generation. As Scholastic's publisher David Levithan points out, "With a printed book, there's no such thing as an analytic."

Electronic reading, on the other hand, is Big Data—which is a big deal, and it keeps getting bigger. Yes, we are in the midst of something big, some would say a revolution. But as big as it is, quantity is not the revolutionary part of the Big Data revolution. The truly big deal is that we are finding ways to do things with this data, and in the process, this data is transforming countless professions and industries—everything from science to baseball.



BookLamp's Book Genome Project maps the StoryDNA of millions of popular books.

Yale University's department of economics, argues that winning baseball games is "about getting things down to one number." He also argues that with better data analysis, a team can "find value in players that no one else can see." Sure, the Yankees had the same data as Peter, but its staff lacked the tools for parsing and analyzing the data for meaningful patterns.

While there are 30 professional teams in Major League Baseball, there are only five major players in publishing. But just as a 40-man roster playing 162 games of baseball can generate gobs of statistics, a single publisher—supporting hundreds of authors and thousands of books across multiple formats in different markets—can quickly find itself overwhelmed by data.

> What is fascinating about this new paradigm is that these metrics say nothing at all about the book as a work of art. As far as these algorithms are concerned, books are fungible cogs.

But what about the stodgy old business of publishing books? What about (and this is bordering on blasphemy) that delicate and creative art of writing? Why did Apple spend possibly $15 million to buy the startup BookLamp? Surely literature is the product of an artist, which is outside the purview of data mining, focus groups, and deep analytics.

With the advent and rise of electronic reading, I can tell you that 9,260 Kindle users highlighted the same line in *Pride and Prejudice*, only 8.2% of *The Da Vinci Code* is about art and art galleries, people who read *The Hunger Games* do so at an average rate of 57 pages per hour, and only 6.4% of readers who bought *Infinite Jest* read it to the end.

Players That No One Else Can See

There is a scene in the film *Moneyball* where Peter, a young graduate of

Historically, the book business has been fairly opaque. Publishers looked at their own sales figures and numbers from sources such as Nielsen BookScan. However, that data wasn't generally shared with the authors or the public. In 2012, this began to change when Amazon started sharing sales data with authors, and the big publishers followed suit.

The idea was simple: By providing authors with sales data, publishers would empower authors to better focus their brands and target their messages. Whereas marketing had been an afterthought, it would now permeate the entire process from inception to book tour. Like Peter, the fictionalized assistant general manager of the Oakland Athletics in *Moneyball*, publishers were trying to arrive at a magic number—that elusive variable that would unlock the secret to success.

A Win Is a Win

At the end of the day, sales are what matter most, so they are always taken into consideration. But where do those sales come from? In *Moneyball*, Peter ignores the obvious, traditional indicators of a player's worth (such as RBIs; runs batted in) and focuses on more obscure statistics (such as on-base percentage and slugging percentage). Similar to RBIs, sales are an indicator of past success—a measure of success that, by itself, only communicates "how much" not "why."

In recent months, Next Big Book (a division of Next Big Sound) and its founder, Alexander White, have been winning industry accolades and awards for an innovative new service that pulls together and analyzes vast amounts of disparate data about books and authors—everything from Facebook Likes to Wikipedia searches to television appearances—and looks for correlations that illustrate how these factors affect sales. If publishers can measure where consumers are spending their time and attention, they can better forecast where to spend their money.

When it was announced that Macmillan was partnering with Next Big Book to create an analytics dashboard, White explained what his service has to offer a company such as Macmillan: "What we're doing is combining all the public social data for every author and title, with our customers' private sales numbers—paper and digital—to help

them understand which social signals are leading indicators of their sales numbers and which marketing events were able to reliably drive social and sales numbers."

The ability to compare data from a wide range of sources—anything from Goodreads ratings, ebook sales, and Twitter followers—allows publishers to track a book's trajectory, measure ROI on marketing campaigns, and perform targeted experiments in different markets and track results in real time. And just as the team in *Moneyball* sought to "find value in players that no one else can see," Brittney Kleinfelter, a marketer at Macmillan, hopes to use Next Big Book to "target our marketing campaign where the other author isn't." And, indeed, Next Big Book is already informing how Macmillan markets books and authors.

Next Big Book analyzes social, sales, and marketing signals.

It is not hard to see how data aggregation and analysis tools are useful to corporate types (marketing executives and the like), but why are publishers opening up these tools to authors? Isn't it the job of an author to focus on her craft and the job of the publisher to promote the work and grow sales?

Do Algorithms Dream of Ebooks?

What is fascinating about this new paradigm is that these metrics say nothing at all about the book as a work of art. As far as these algorithms are concerned, books are fungible cogs. These metrics seek to determine a formula for success, but is that success entirely the result of marketing and branding? What about those quaint ideas such as talent and craft?

By sharing data and opening these platforms to authors, the message from publishers is clear: Sales are the only measure of achievement that matters, so write what sells. Like baseball, it doesn't matter how you get there as long as your team has more runs by the end of the game. It will be interesting to see if this *Moneyball* paradigm—when applied to publishing—precludes idiosyncratic works of literary fiction or better enables publishers to identify and cultivate talent as Peter did when he leveraged the talent of marginal, oddball players.

In a *Wall Street Journal* article, Jim Hilt (then the VP of ebooks at Barnes & Noble) shared that deep analysis of Nook user data reveals a lot about reader behavior (for example, that fantasy fans read more books than readers of literary fiction and at a faster rate with a greater rate of completion). Hilt suggests that by determining where readers drop off in certain kinds of books and by sharing that data with publishers they can "help authors create even better books than they create today." Some publishers are already releasing and testing digital excerpts before the print release, studying the analytics, and reporting back to the author.

It seems likely that as authors become students of what sells, some will go beyond looking at winning genre tropes and media markets and start looking at styles of writing, allowing the data (such as average word count, chapter length, number of unique words, and number of words per paragraph), either deliberately or subconsciously, to influence their craft. As tools of analysis become more sophisticated, data will inform what authors write, how they write it, and to whom.

Author Tim Parks, in a recent piece for *The New York Review of Books*, notes, "The conditions in which we read today are not those of fifty or even thirty years ago, and the big question is how contemporary fiction will adapt to these changes, because in the end adapt it will. No art exists independently of the conditions in which it is enjoyed."

These "conditions" in which we enjoy books are part of an increasingly fractured media landscape. That landscape generates lots of data and offers up myriad opportunities for distraction and analysis, and the energies required to take on a substantial work of fiction are stretched in all directions.

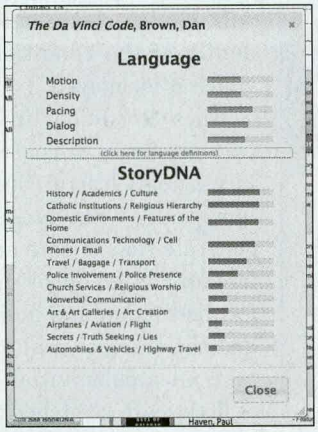
As serious fiction becomes more marginal and cultic, commercial fiction will respond to our changing patterns of attention just as it has responded to market pressures in the past. The idea is that data will now inform the process of writing to adapt to our changing patterns of attention.

For fans of popular fiction, distraction is often the goal. Books are entertainment—welcome distractions. For those of us desirous of distraction, Big Data is going to help authors craft more addictive and enjoyable books just as data from video game testers helps shape content in that industry.

BookLamp Lighting the Way?

Apple has been buying up several small companies in recent months, leading many in the industry to speculate that the company is laying the groundwork for a new media streaming service. The aforementioned BookLamp is one of those companies. It analyzes the pacing and density of a work by tracking 32,160 distinct datapoints per book. For example, users could analyze various writing styles and chart them against sales data.

For retailers, this data enables the creation of better recommendation engines. For authors, this data provides insights into reading habits. But for readers, this same data provides an opportunity to learn more about the content they are reading or discover new content by looking for similarities in the data. Another reference tool, Amazon's X-Ray, allows users to explore data related to a book's contents and "get to its bones."



The Da Vinci Code *is 18.6% Religion and Religious Institutions, 9.4% Police & Murder Investigation, 8.2% Art and Art Galleries, and 6.7% Secret Societies & Communities*

It is my assertion that, albeit on a smaller scale than the video game industry, the publishing industry is taking a step in this data-centric direction, and future projects will be analyzed as video games are now. The value of a book may lie in the subjective judgment of a reader, but that reader's habits and preferences will be analyzed, scrutinized, and quantified. The popular author of the future will be more like a researcher in a lab and less like a writer gazing out of her window contemplating the texture of leaves as they fall from a tree.

While retailers and publishers are engaged in an all-out arms race to engineer data-driven discovery and delivery platforms, libraries are still wrestling with the logistics of basic delivery. As users become accustomed to seamless discovery platforms that use data to deliver tailored suggestions, the library catalog and lending platform become ever more alien and removed from users' lived digital experience.

Libraries Left in the Dark?

These days, "free" nearly always means surrendering privacy. Libraries are perhaps the only exception. For years, the prevailing attitude toward data in Libraryland has been, "Don't ask, don't tell." Librarians do not want anything to do with patrons' borrowing records and routinely purge records in the name of privacy. While this cautious approach is understandable and even admirable in an age of corporate and government spying, it may be out of step with changing user expectations.

A recent Pew Internet & American Life Project study found that many library patrons, about 64% of responders to the survey, would welcome receiving recommendations based on library borrowing activity. As we live our busy digital lives, it seems that we are increasingly willing to sacrifice some measure of privacy if there is sufficient payoff in the form of greater convenience.

Of course, these same tools of data analysis being used by publishers and tech companies could theoretically be leveraged by library systems to analyze trends and inform future purchasing or marketing. But libraries are only a small part of the book-consuming world, which is only a small part of the information-consuming world.

> With the advent and rise of electronic reading, I can tell you that 9,260 Kindle users highlighted the same line in *Pride and Prejudice*. ...

The most relevant institutions in the 21st century will be the ones with the most data and the most sophisticated tools for analyzing it. But Big Data is big money—something that libraries don't have. Libraries will not get far by miming the practices of big players in a macro environment and risk obsolescence if they are unwilling or unable to capture and share detailed data across a multitude of institutions.

Indeed, there has been some movement in this direction. Even as we refuse to capture or share anonymized statistics on patron behavior, readers advisory services marketed to libraries are buying and using sales data from Amazon and elsewhere to improve discovery and reader's advisory platforms. It seems libraries are unwilling to collect the data, but they are eager to leverage the data collected by others.

The American Library Association (ALA) defines privacy as "the right to open inquiry without having the subject of one's interest examined or scrutinized by others." This is certainly a noble sentiment, but scrutinizing individual interests is exactly what Big Data is all about. It is practically a prerequisite for the development of robust discovery and recommendation engines.

With more than 17,000 libraries and 2.5 billion materials circulated annually in the U.S., libraries are throwing away potentially useful data that could help them anticipate and adapt to cascading technological disruptions. To remain relevant in the long term, libraries will need to leverage the power of data. To do that, they need to find a practical balance—one that protects anonymity while gathering and leveraging user data.

As things stand now, libraries have an advantage over subscription ebook services such as Scribd and Amazon Unlimited because publishers have been reluctant to release their best titles to these services. But as these services grow, it seems likely that publishers will come to terms with using these platforms as alternative revenue sources. When these services are able to offer seamless delivery, no waitlists, robust discovery platforms, and comparable content, the public library—with its cumbersome interfaces and long waitlists—will suddenly seem less appealing.

Libraries have remained relevant for decades by adapting in order to deliver the content people want when, where, and how they want it. But changes are coming more quickly now; without the aid of data, we are in the dark. Without Big Data, at best, we are like the Yankees in *Moneyball*: a dynastic and inflexible system driven by instinct and tradition, but largely ignorant of trends that can only be illuminated by sophisticated analysis.

Ben Johnson *(bjohnson@cbpl.lib.ia.us) is the collection development librarian at the Council Bluffs Public Library in Iowa.*