



Interview with Jay Gale, co-founder, Inkubate

**For podcast release
Monday, May 13, 2019**

KENNEALLY: Agents and publisher want to find authors. Authors want to find readers. And readers, they want to find books. Helping make these discoveries happen is a powerful digital tool that evaluates writing styles and matches a work with books just like it.

Welcome to Copyright Clearance Center's podcast series, I'm Christopher Kenneally for Beyond the Book.

Discovery is a persistent challenge for many types of online commerce. In the digital environment, where inventory is nearly unlimited, consumers need help to discover what will interest them. Companies like Netflix and Amazon have invested many millions to address the issue, developing so-called recommendation engines that push forward products.

Inkubate is a data analytics platform expressly designed for authors to reach audiences with pinpoint accurate marketing. Research has shown that readers respond more to writing style than either genre or subject matter. The digital service of Inkubate reads a manuscript to find writing characteristics it has in common with other works on social media platforms and elsewhere. Authors then match their books with readers already inclined to like them. Computers, not so surprisingly, know more about readers than most novelists do.

To explain how it's all done, Inkubate cofounder Jay Gale joins me from his office in seaside Portsmouth, New Hampshire. Welcome to Beyond the Book, Jay.

GALE: Chris, great, thanks for having me.

KENNEALLY: Well, we're looking forward to exploring this with you. It's another example of how authors and algorithms are interacting in the digital age. And I guess the place to start is to explain for the listeners exactly how Inkubate works. If I'm an author, I upload a manuscript, what happens? What does Inkubate service actually do?

GALE: So what we do is we take that uploaded manuscript and we dash it up to the cloud, where we convert it to a bunch of ones and zeros, and we use several



hundred thousand very powerful algorithms to look for the patterns within that manuscript that are really diagnostic of the individual. It allows us to hone in on how a individual uses the underpinnings of the construction of language to communicate. We then compare that instantly to the pre-computed what we call fingerprints of hundreds and thousands of manuscripts that have been previously published and placed in the marketplace. Most of them have been on a book scan best seller list at one time or another, and we're able to measure the similarities between all of those thousands of works to find the closest matches.

And, in fact, what we've learned is that writers who write alike using our definition of writing style, which are these underpinnings, they actually have – their audiences actually will have affinity. So if I am a independent author and I discover that I write a lot like James Patterson, for example – out of thousands of authors, he's my closest writing match – we know that the readers of James Patterson have a predisposition to like not simply the content that James Patterson writes – that's important – but even more importantly, how he uses language to communicate.

KENNEALLY: Well, that's the important point there is the use of language. And as I understand it, the algorithms that you work with take advantage of a certain field of data processing called natural language processing. Can you help us better understand what natural language processing is?

GALE: Well, natural language processing is actually a very large field which defines the ability for computers to actually ingest the written word and then to turn that into a machine-readable data that can be analyzed. But there are many subsets of that particular field, and one of the subsets, the one that we lead, really, the world in expertise in, is known as stylometry – S-T-Y-L-E-O-M-E-R-T-Y, (sic) stylometry. And that is the study of forensic language.

KENNEALLY: Well, then, tell us what exactly is forensic language? It sounds a little spooky. And we are not talking here about critical evaluation, such as the critics for *The New York Times* might make, but the computer is making evaluations, it's matching, it's seeing patterns, it's identifying common uses.

GALE: That's correct. So a good example of how forensic analysis of language can be used and has been used by our data scientists in the past is the very famous example of Robert Golthwaite (sic) the author, and the book *A (sic) Cuckoo's Calling*, which was a book that was published in the UK and quickly raced to the top of the charts. And the author of that book could not be determined by following any of the usual cues in the author biography, the CV, any of that. And our team was asked to opine on the likely authorship. In fact, our science team



determined that it was no other than J. K. Rowling, and it was quite a significant news story at the time. In fact, there's a *Scientific American* article about this. So matching an author by being able to understand the way that they write, which is virtually inalterable, forms the basis of being able to not only identify an author, but make comparisons between authors.

KENNEALLY: And so for the authors, though, this is an important tool. Many authors – they know their own work, but as I mentioned in the introduction, they might not really understand what readers are looking for, and how well it matches up with other books out there. But by seeing these common patterns in style, you can then tell the author, well, you ought to be thinking about reaching readers who, as you said in your example, have a certain passion for James Patterson. You may not even notice that yourself, or be aware of it, but the algorithms are telling you this is true.

GALE: That's right. So once we've identified who you write like on that atomic level, the next step is helping the author actually identify ways, or to find ways to get in front of that audience of that particular author. So that's where another piece of our technology comes into play where we actually, in real time, scour the web for positive reader reviews for those particular books that were matched by our independent author, and look for the positive reader reviews across many different websites, but some of the best known are the Goodreads site, for example, Amazon book reviews, these are all publicly published reviews. We are able to then ingest to those reviews, many hundreds of them at a time, and collate and correlate what's in common – phrases, key words, themes that these independent readers are using to talk about the books that they're enjoying and reading online for this particular – in this case, James Patterson.

What we know is that readers use those same sorts of terms and those same sorts of phrases when they're doing searches through search engines, whether on the retail side, or Google or Yahoo or your favorite search engine. They're using the same sorts of phrases and terms to search for new content. So this list of words that we provide, and phrases that we provide back to those authors becomes extremely valuable in their search engine optimization or in their Google ads or perhaps in Amazon ads, where they're bidding on key words and phrases to trigger their display advertising.

KENNEALLY: Well, it's fascinating. It's sort of taking the book-selling world at its particular moment. We no longer wander into a bookstore. We may still do that from time to time, but so many people today are looking for something to read online, and it's just accepting that as a reality, and helping the authors reach those



particular readers as quickly as possible. It sorts things out for the authors in advance and really allows them to pinpoint their marketing.

GALE: That's right, and we have numerous writers that have reported back to us that when they run the score product and then taken a look at the suggestions made for the key words and phrases that are relevant to attracting that audience that they've been matched into, that they're getting upwards of double the number of impressions on their display advertising. That's pretty significant.

KENNEALLY: And I wonder whether the authors are sometimes surprised by what your algorithms tell them about their work.

GALE: Well, sometimes they are because you don't necessarily write the same content or even the same subject matter as the writer whom you most closely resemble in terms of communication style. That can be a little off-putting at first, but it can also be very valuable because, as you're aware, a tremendous amount of content, particularly content from the self-published community tends not to be properly tagged with its metadata in the first place. Everything from genre and subject matter are frequently misinterpreted by independent writers who may not understand all the nuances of the many thousands of slices of data that go into determining what is the actual genre.

So we might have someone who's writing romance perhaps matched into a mystery author. Or someone who's writing science fiction being matched into fantasy. That also shouldn't be surprising because fiction very often contains a lot of those elements. It's sometimes very hard to classify a specific genre. So it is insightful.

KENNEALLY: And does the service work for nonfiction, as well?

GALE: It does, although our nonfiction component for this is an offline component. We don't have that product deployed publicly in the marketplace. We only have the fiction product out there. There's a lot of reasons for that, some of which have to do with the complexities of building a corpus. Just let me tell you that a corpus are those collections of books that we preprocess to analyze these bestselling authors' elemental writing style. It's a much small segment of the marketplace, so we tend to deploy that directly to publishers. However, it does work equally well.

KENNEALLY: Well, we are talking today about a fascinating application of authors and algorithms with cofounder of Inkubate, Jay Gale. And Jay, I wonder whether other publishing professionals can make use of the service, too.



GALE: Well, they do, and they can. It's a great tool for publishing professionals who are interested in examining their slush piles. It can help them quickly identify the cream from what is really a sort of a mixture. It's also very effective for publishers who are interested in having an objective tool to help them evaluate what are commonly known as the comps, C-O-M-P-S, in the marketplace for books that they considering publishing. Until now that's been very much purely a subjective determination. Having some objective data to help with that decision can have a significant impact on how those publishers strategically place and market those books.

KENNEALLY: Well, to that last point, Jay Gale, I wonder if there's any resistance to the idea that computers can tell people in publishing more about books and their catalogs than they think they know themselves. It seems to me that publishing has, for a very long time, been a business that's all about gut, and this is about data.

GALE: Right, and that's been true for a long time, but most publishing companies realize today that their books are being sold into an environment that is entirely dominated by big data. A book does not rise to the top of a Amazon or Barnes & Noble or retailer du jour's search engine solely based upon co-oping (sp?) fees or floor planning like it used to be in the old days at a brick and mortar store. Folks are using key words and phrases to search for content within these search engines. If the data isn't significant and if the data isn't detailed enough, you run the risk of not getting the same sorts of responses that you did in the past.

So most of the big publishers realize that data is an important part, and they're trying to incorporate analysis of actual data surrounding the content and the underlying style to make these books relevant in the marketplace. It's not unlike what's happened with the music industry, for example. Spotify and Pandora, as examples, make book recommendations based upon actually analyzing the underlying data surrounding sonograms as well as utilizing information about what other people may have bought, what other people might have liked. So it's very important.

KENNEALLY: The relationship of data analytics to the new world of publishing is a story we are following closely here at Beyond the Book and we appreciate learning more about it from Inkubate cofounder Jay Gale. Jay thank you so much for joining me on Beyond the Book.

GALE: Thank you very much.

KENNEALLY: Beyond the Book is produced by Copyright Clearance Center. Our co-producer and recording engineer is Jeremy Brieske of Burst Marketing. Subscribe



to the program wherever you go for podcasts, and follow us on Twitter and Facebook. The complete Beyond the Book podcast archive is available at beyondthebook.com. I'm Christopher Kenneally. Thanks for listing and join us again soon on CCC's Beyond the Book.

END OF FILE