JEFFERSON POOLEY

Surveillance Publishing

25 March 2022 | doi:10.5281/zenodo.6384605 | No Comments

Jefferson Pooley on Surveillance Publishing, its history in modern societies during the last couple of decades, and the potential costs of these practices for both service providers and their users.

AUTHOR INFO

Dr. Jefferson Pooley is professor of media & communication at Muhlenberg College in Pennsylvania, USA. His research interests center on the history of media research within the context of the social sciences, with special focus on the early Cold War behavioral sciences. He also writes frequently on scholarly communication topics as well as social media and the self. He is author of *James W. Carey and Communication*

In April 1998, two Stanford graduate students, Sergey Brin and Larry Page, flew across the world to deliver a paper on their nascent search engine, Google. Speaking at the Seventh International World Wide Web conference (WWW 98) in Brisbane, Australia, Brin and Page described how their approach—taking the web's existing link "graph" as a proxy for quality and relevance—improved on the classified-by-hand indexes of Yahoo!, Lycos, and the like. Six months later, they took their idea commercial, with the pair working out of a nearby garage. Within two years Brin and Page had dispatched their search engine rivals, on the way to building the largest advertising business in the history of capitalism.

Google's dorm-to-garage origin story is well-known. Less famous is the debt that Brin and Page owed to library science and the field of bibliometrics. As the pair acknowledged in Brisbane, their key idea—to use the web's link structure as a plebiscite for search relevance—was borrowed from citation analysis. "The citation (link) graph of the Web," they said, "is an important resource that has largely gone unused in existing search engines." A given webpage's "PageRank," they explained, is a measure of its "citation importance," which turns out to match, with uncanny consistency, what searchers want to find. Their approach, they continued, is an extension of the "[a]cademic citation literature."

The Google founders had taken the core insight of bibliometrics, a field that emerged in the 1960s to study (among other things) the web of academic citations. As the historian of science Derek de Solla Price put it in a seminal 1965 paper, citations furnish

Research: Reputation at the University's Margins (Peter Lang, 2016), and co-editor of *Society on the Edge: Social Science and Public Policy in the Postwar United States* (Cambridge, 2021), The History of Media and Communication Research (Peter Lang, 2008) and *Media and Social Justice* (Palgrave, 2011). Pooley is director of mediastudies.press, a scholar-led open access publisher of books and journals.

CITE AS

Pooley, J. (2022). Surveillance Publishing. *Elephant in the Lab*. https://doi.org/10.5281/zenodo.6384605

a “total world network of scientific papers.” By the early 1970s, on the strength of computing advances, full-fledged citation analysis was being used to measure journal impacts, scientific productivity, and the structure of academic influence.

Two decades later in Brisbane, Brin and Page positioned Google as the academic antidote to ad-driven search engines. They complained that companies like Yahoo! wouldn’t make their methods public, with the result that search technology remains “largely a black art.” With Google, they said, “we have a strong goal to push more development and understanding into the academic realm.” In a now notorious appendix to their published talk, the two graduate students decried the ad-driven business model of their commercial rivals. “We expect,” Brin and Page wrote, “advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers”—a “particularly insidious bias,” they added, since it’s so hard to detect.

They changed their minds. In the face of the 2001 dot-com meltdown and investor demands, Brin and Page—to borrow the Silicon Valley verb—pivoted. As Shoshana Zuboff has documented, the company went all in with ads: targeted ads, informed by the user data trove the company had laying about. By 2004, the company had gone public, valued at $27 billion. Harnessing its search-and-services-derived user data, Google went on to capture almost 30% of worldwide digital ad revenue. Today the market value of Alphabet, Google’s parent company, hovers around $2 trillion. Built up from academic citation analysis, the company is the defining example of what Zuboff calls “surveillance capitalism.”

There is another irony. The field of bibliometrics, all

the way back to its early-1960s emergence, was already enmeshed in data capitalism. Here again, the story is well-known: Eugene Garfield, a would-be chemist turned science entrepreneur, established his science-indexing business, the Institute for Scientific Information (ISI), in the mid-1950s. In 1964 Garfield's ISI produced the first Science Citation Index, a database of published papers and their citations. Bibliometrics pioneers such as de Solla Price partnered with Garfield to mine the service's database—hence de Solla Price's total world network of scientific papers. Other ISI indexes for the social sciences and for the arts followed in the 1970s, when Garfield's firm also began publishing its Journal Citation Reports. In 1992, with the World Wide Web in its infancy, Garfield sold ISI to Thomson, the Canadian information giant. The business traded hands again in 2016, in a private equity spinoff called Clarivate. Garfield's citation index—now called the Web of Science—stood at the center of the $3.5 billion deal.

From the Web of Science back to the web: In fundamental ways Clarivate's business resembles Alphabet's. Clarivate, of course, doesn't feed from the advertising firehouse like Google. But both companies mine behavior for data, which they process into prediction products. In Google's case, we're all in on the action, with every search and email; once refined, the data is sold to the company's customer-advertisers for targeted display. Clarivate's behavioral data is harvested from a much smaller public—working academics—who, in another difference from Google, are the company's main customers too. But the core business strategy is the same: extract data from behavior to feed predictive models that, in turn, get refined and sold to customers. In one case it's search terms and in the

other abstracts and citations, but either way the point is to mint money from the by-products of (consumer or scholarly) behavior. In place of Google's propensity to buy, Clarivate is selling bets on future scholarly productivity and impact, among other academic prediction products.

This article lingers on a prediction too: Clarivate's business model is coming for scholarly publishing. Google is one peer, but the company's real competitors are Elsevier, Springer Nature, Wiley, Taylor & Francis, and SAGE. Elsevier, in particular, has been moving into predictive analytics for years now. Of course the publishing giants have long profited off of academics and our university employers—by packaging scholars' unpaid writing-and-editing labor only to sell it back to us as usuriously priced subscriptions or article processing charges (APCs). That's a lucrative business that Elsevier and the others won't give up. But they're layering another business on top of their legacy publishing operations, in the Clarivate mold. The data trove that publishers are sitting on is, if anything, far richer than the citation graph alone.

Why worry about surveillance publishing? One reason is the balance sheet, since the companies' trading in academic futures will further pad profits at the expense of taxpayers and students. The bigger reason is that our behavior—once alienated from us and abstracted into predictive metrics—will double back onto our work lives. Existing biases, like male academics' propensity for self-citation, will receive a fresh coat of algorithmic legitimacy. More broadly, the academic reward system is already distorted by metrics. To the extent that publishers' tallies and indices get folded into grant-making, tenure-and-promotion, and other evaluative decisions, the metric tide will gain power. The biggest risk is that

scholars will internalize an analytics mindset, one already encouraged by citation counts and impact factors.

SURVEILLANCE AS A SERVICE

Useful as it is, Shoshanna Zuboff's notion of "surveillance capitalism" is too tightly drawn around a relatively small pocket of the economy, digital advertising. That same narrowed aperture led Zuboff, in *The Age of Surveillance Capitalism*, to over-emphasize the novelty of the behavioral futures business she attributes to Google. The insurance and credit-rating industries, to mention two, have hitched data to predictive profit for well over a hundred years. As we have seen, Garfield's ISI was in the data business before Larry Page and Sergey Brin were born.

To get at the publishers' kinship with Google or, for that matter, the Hartford, we need a broader descriptor. The legal scholars Mariano-Florentino Cuéllar and Aziz Huq have proposed a pluralized alternative, "surveillance economies," to refer to the range of business models that seek to monetize behavioral data. "As more industries find ways to incorporate behavioral surpluses into their business models," they write, "the share of the economy that falls under this term will increase, perhaps dramatically." Cuéllar and Huq foreground the pluralism: The specific contours of any given surveillance economy will vary, based on sector-specific norms and regulations. There is, in other words, no need to take the analogy to Google too far. Data businesses based on academics' citations and downloads are unlikely to emulate Google's ad-driven model. The big publishers, along with Clarivate and other potential players, are more likely to piggyback on their existing subscription strategy,

with data products licensed to university and other research clients. Either way, they'll be lapping up the behavioral surplus that scholars produce. As CUNY law professor Sarah Lamdan put it in a recent talk, "your journals are spying on you."

The publishers are in an enviable position, since researchers generate data with every article engagement or peer review report. Some of that data gets folded into the publishers' core products, by way of download counts and article recommendations. But we have every reason to believe, based on existing data products alone, that publishers are skimming scholars' behavioral residue on the prospect of monetization to come. In an important recent paper, STS scholar Jathan Sadowski took issue with the commonplace that data is the "new oil." On the commodity view that he challenges, data is raw material for other products, easy to exchange for cash. Data is often a commodity like this, Sadowski concedes; the sprawling data brokerage industry is an illustration in point. But it's also useful to think about data as *capital*, in the specific sense of "capital" developed by the late French sociologist Pierre Bourdieu. Data capital resembles in form something like Bourdieu's cultural capital: Though a learned appreciation for abstract art can, in certain conditions, lead to a lucrative job, the value of that cultural capital isn't merely, or even mainly, monetary. Data capital, likewise, can be converted into dollars in some contexts. But its value to owners may lie elsewhere. Firms may use data to guide strategy, refine workflows, or train models, among other things. Like social or cultural capital, there is a prospective quality to data accumulation—an incentive to hoard on the expectation of future value.

Scholarly publishing is its own, emerging surveillance economy. We can call a company a *surveillance*

publisher if it derives a substantial proportion of its revenue from prediction products, fueled by data extracted from researcher behavior. On that definition, we already have surveillance publishers in our midst.

THE FULL-STACK PUBLISHER

Consider Elsevier. The Dutch publishing house was founded in the late nineteenth century, but it wasn't until the 1970s that the firm began to launch and acquire journal titles at a frenzied pace. Elsevier's model was Pergamon, the postwar science publishing venture established by the brash Czech-born Robert Maxwell. By 1965, around the time that Garfield's Science Citation Index first appeared, Pergamon was publishing 150 journals. Elsevier followed Maxwell's lead, growing at a rate of 35 titles a year by the late 1970s. Both firms hiked their subscription prices aggressively, making huge profits off the prestige signaling of Garfield's Journal Impact Factor. Maxwell sold Pergamon to Elsevier in 1991, months before his lurid death.

Elsevier was just getting started. The firm acquired *The Lancet* the same year, when the company piloted what would become ScienceDirect, its Web-based journal delivery platform. In 1993 the Dutch publisher merged with Reed International, a UK paper-maker turned media conglomerate. In 2015, the firm changed its name to RELX Group, after two decades of acquisitions, divestitures, and product launches—including Scopus in 2004, Elsevier's answer to ISI's Web of Science. The "shorter, more modern name," RELX explained, is a nod to the company's "transformation" from publisher to a "technology, content and analytics driven business." RELX's strategy? The "organic development of increasingly sophisticated information-based analytics and

decisions tools." Elsevier, in other words, was to become a surveillance publisher.

Since then, by acquisition and product launch, Elsevier has moved to make good on its self-description. By moving up and down the research lifecycle, the company has positioned itself to harvest behavioral surplus at every stage. Tracking lab results? Elsevier has Hivebench, acquired in 2016. Citation and data-sharing software? Mendeley, purchased in 2013. Posting your working paper or preprint? SSRN and bepress, 2016 and 2017, respectively.

Elsevier's "solutions" for the post-publication phase of the scholarly workflow are anchored by Scopus and its 81 million records. Curious about impact? Plum Analytics, an altmetrics company, acquired in 2017. Want to track your university's researchers and their work? There's the Pure "research information management system," acquired in 2012. Measure researcher performance? SciVal, spun off from Scopus in 2009, which incorporates the media monitoring service Newsflo, acquired in 2015.

Elsevier, to repurpose a computer science phrase, is now a full-stack publisher. Its products span the research lifecycle, from the lab bench through to impact scoring, and even—by way of Pure's grant-searching tools—back to the bench, to begin anew. Some of its products are, you might say, services with benefits: Mendeley, for example, or even the ScienceDirect journal-delivery platform, provide reference management or journal access for customers *and* give off behavioral data to Elsevier. Products like SciVal and Pure, up the data chain, sell the processed data back to researchers and their employers, in the form of "research intelligence." Even the company's PDF viewer, built into

ScienceDirect and other products, is extracting granular details about readers.

It's a good business for Elsevier. Facebook, Google, and ByteDance have to give away their consumer-facing services to attract data-producing users. If you're not paying for it, the Silicon Valley adage has it, then you're the product. For Elsevier and its peers, we're the product *and* we're paying (a lot) for it. Indeed, it's likely that windfall subscription-and-APC profits in Elsevier's "legacy" publishing business have financed its decade-long acquisition binge in analytics. As Björn Brembs recently Tweeted: "massive over-payment of academic publishers has enabled them to buy surveillance technology covering the entire workflow that can be used not only to be combined with our private data and sold, but also to make algorithmic (aka. 'evidence-led') employment decisions." This is insult piled on injury: Fleece us once only to fleece us all over again, first in the library and then in the assessment office.

Elsevier's prediction products sort and process mined data in a variety of ways. The company touts what it calls its Fingerprint® Engine, which applies machine learning techniques to an ocean's worth of scholarly texts—article abstracts, yes, but also patents, funding announcements, and proposals. Presumably trained on human-coded examples (scholar-designated article keywords?), the model assigns keywords (e.g., "Drug Resistance") to documents, together with what amounts to a weighted score (e.g., 73%). The list of terms and scores is, the company says, a "Fingerprint." The Engine is used in a variety of products, including Expert Lookup (to find reviewers), the company's JournalFinder, and its Pure university-level research-management software. In the latter case, it's scholars who get Fingerprinted:

Pure applies semantic technology and 10 different research-specific keyword vocabularies to analyze a researcher's publications and grant awards and transform them into a unique Fingerprint™—a distinct visual index of concepts and a weighted list of structured terms.

The machine learning techniques that Elsevier is using are of a piece with RELX's other predictive analytics businesses aimed at corporate and legal customers, including LexisNexis Risk Solutions. Though RELX doesn't provide specific revenue figures for its academic prediction products, the company's 2020 SEC disclosures indicate that over a third of Elsevier's revenue come from databases and electronic reference products—a business, the company states, in which "we continued to drive good growth through content development and enhanced machine learning and natural language processing based functionality."

Many of Elsevier's rivals appear to be rushing into the analytics market, too, with a similar full research-stack data harvesting strategy. Taylor & Francis, for example, is a unit of Informa, a UK-based conglomerate whose roots can be traced to Lloyd's List, the eighteenth-century maritime-intelligence journal. In its 2020 annual report, the company wrote that it intends to "more deeply use and analyze the first party data" sitting in Taylor & Francis and other divisions, to "develop new services based on hard data and behavioral data insights." Last year Informa acquired the Faculty of 1000, together with its OA F1000Research publishing platform. Not to be outdone, Wiley bought Hindawi, a large independent OA publisher, along with its Phenom platform. The Hindawi purchase followed Wiley's 2016 acquisition of Atypon, a researcher-facing software firm whose

online platform, Literatum, Wiley recently adopted across its journal portfolio. “Know thy reader,” Atypon writes of Literatum. “Construct reports on the fly and get visualization of content usage and users’ site behavior in real time.” Springer Nature, to cite a third example, sits under the same Holtzbrinck corporate umbrella as Digital Science, which incubates startups and launches products across the research lifecycle, including the Web of Science/Scopus competitor Dimensions, data repository figshare, impact tracker Altmetric, and many others. There was, last month, a fateful convergence: Elsevier announced a pilot program to incorporate some Wiley and Taylor & Francis journals into Elsevier’s ScienceDirect. If the pilot leads to something lasting, we’ll be one step closer to what Leslie Chan has called the “platformization of scholarly infrastructure.”

The big publishing oligopolists aren’t the only firms looking to profit from researcher behavior. There is, of course, Clarivate itself, whose $5.3 billion purchase of ProQuest closed in late 2021, the same day that Wiley announced its purchase of stealth for-profit Knowledge Unlatched. The two venture-backed academic social networks, Academia and ResearchGate, re-package researchers’ activity on the sites via user analytics; observers have speculated for years that the companies will build analytics products based on their data troves. ResearchGate is already selling a jobs-search tool as well as targeted advertising (“Upgrade your targeting options with sophisticated Sequential Ads”). Surveillance businesses parasitic on other facets of nonprofit higher ed—student life, for example, or the classroom—are growing too. Online program management (OPMs) firms, a business Wiley is in too, are going public with multi-billion dollar valuations

predicated, according to reports, on the value of their tens of millions of "learner" profiles. Likewise with venture-funded EAB, which touts its data-driven academic advising software as the first enterprise-level "student management system." Even Google itself could, at any moment, decide to monetize its Google Scholar search engine—in what would be a return, a fitting one, to its bibliometrics roots.

The scholarly community is beginning to fight back. There is the Stop Tracking Science petition site, with over a thousand signatures at last count. SPARC North America, the OA advocacy group, has issued an alarm. The German national research foundation, Deutsche Forschungsgemeinschaft (DFG), released its own report-cum-warning in October—"industrialization of knowledge through tracking," in the report's words. A 2020 read-and-publish agreement between Elsevier and Dutch universities sparked an outcry, largely because the company had baked its prediction products into the deal.

Sociologist David Murakami Wood warned us all the way back in 2009: Publishers are becoming, if not Big Brother, then at least several little ones. The chorus, in the last several years, has grown louder, with alerts sounded by Alejandro Posada, George Chen, Lisa Hinchliffe, Leslie Chan, Richard Poynder, Björn Brembs, and—in *Elephant in the Lab* last April—Renke Siems. The problem is that most working academics have no idea they're being packaged and sold in the first place.

LOOPING EFFECTS

Siphoning taxpayer, tuition, and endowment dollars to access our own behavior is a financial and moral indignity. That we are paying the sellers a second

time, after budget-draining subscription and APC outlays, is a scandal. Elsevier made $1.4 billion in profit last year, on $3.6 billion in revenue—a profit margin of 38%. That lucrative business is built on scholars' unpaid labor, as subsidized by our university-employers. The typeset product of that labor, in a longstanding complaint, is sold back to us at extortionate prices. Now Elsevier is skimming the behavioral cream, and selling that too. If anything, profits from the first business have financed the build-up of the second.

Consider, too, the intended use of these surveillance products. The customers for many of the predictive analytics sold by Elsevier and others are university administrators and national research offices. The products' purpose is to streamline the top-down assessment and evaluation practices that have taken hold in recent decades, especially across the Anglophone academy. Some of the practices, and most of the mindset, are borrowed from the business sector. To varying extent, the zeal for measurement is driven by the idea that the university's main purpose is to grow regional and national economies. Products like Pure and SciVal are, or will be, among the quantified tools by which economic and engineering values shape what we mean by higher education. At the very least, their dashboard tabulations will be deployed to justify "program prioritization" and other budgetary re-allocations. As Ted Porter has observed, quantification is a way of making decisions without seeming to decide.

In that sense, the "decision tools" peddled by surveillance publishers are laundering machines—context-erasing abstractions of our messy academic realities. It's true that the standard research article, and even its underlying datasets, are already abstracted. But black box researcher productivity

scores, to take one example, are at another remove from our knowledge-making practices. One reason this matters is that algorithmic scores and indices can camouflage the biases that structure academic life. Consider center-periphery dynamics along North-South and native-English-speaking lines: Gaps traceable to geopolitical history, including the legacy of European colonialism, may be buried still deeper under the weight of proprietary metrics.

The problem isn't merely camouflage. With all the authority granted quantitative measure, up to and including funding and hiring decisions, predictive scoring might make smuggled-in biases worse. As a number of scholars have shown, metrics and rankings help enact the world that they purport to merely describe. Thus native English speakers might appear more likely to produce impactful papers, based on past citation data used to train a predictive algorithm—a measure that could, in turn, justify a grant award. Such dynamics of cumulative advantage would serve to widen existing disparities—a Matthew effect on the scale of Scopus.

The looping effects of algorithmic scoring may include playing to the measure. As Goodhart's Law has it, when a measure becomes a target, it ceases to be a good measure. Scholars, like other subjects of ranked measurement, may "optimize" their papers to appeal to the algorithm. If grants, promotion, and recognition follow, such behavior will reinforce an already metricized reward system. We may tweak our work to be, in Tarleton Gillespie's phrase, algorithmically recognizable—or even to see ourselves through the prism of Elsevier's predictive analytics.

An earlier version of this article will be published in the Journal of Electronic Publishing.