9

Finding What We Need: Searching and Filtering

How would authors or publishers know which works had market potential for snazzier editions? And how readers which were most worth their attention? Search engines present would-be readers with everything found on a subject. Depending on the topic, that could be an overwhelming amount. For those interested in much-discussed issues, where to start? In conventional publishing, prestige was the guide. Readers began with articles in the most selective journals, books from the most reputable presses. Here, peer review had worked its astringent magic, sorting wheat from chaff.

That still left a mountainous oversupply of content for weary eyes. Reviews in the press then did some postpublication sorting, alerting readers to the pitfalls and promises of recent output. The relationship between conventional peer review and postpublication review is akin to that between public health expert and emergency room medic. The former seeks to keep populations healthy and away from physicians in the first place. The latter deals with the mess afterward. Ideally, flawless peer review would leave postpublication review unnecessary, but of course that is rarely the case. And it raises the question: in a digital world, what is the point of prepublication review?

Peer Review Redux

In the analog era, prepublication peer review sought to reserve scarce resources for works deserving dissemination. But if the cost of posting content on the global bulletin board is negligible, why bother with upfront vetting? To continue with medical analogies, in theory, prevention would not matter if we had perfect and painless cures. Prevention makes most sense for diseases without remedies or that subject victims to avoidable suffering.

Yet, prevention also imposes costs of its own: restrictions on our behavior, pleasures forgone, travels or experiences shunned. A cure after the fact is often preferable to the abnegation of prevention. Our culture is saturated with the alleged virtue of prevention and its attendant moralizing. Prevention is premised on individual responsibility, and guilt when it fails. We are culturally blinded to the advantages of cures. Historically, treatments for venereal diseases have been attacked as promoting sin through sexual libertinage by sparing the promiscuous their ravages.[1] A pill to dodge obesity, permitting us the costless pleasures of gluttony: were that on offer, imagine the ensuing shriek of moralizing censure.

The issues are similar in debates about preventive pre-facto review versus curative post-facto review. Prepublication peer review assumes that the experts know best and can be trusted to weed out unsuitable content. Besides resting on a paternalist attitude toward readers, it works only when properly implemented. The experts can weigh in just as well after publication. By not limiting it to a small group of initiates, a more thorough review may follow. Once resources no longer have to be husbanded, with content posted to the global bulletin board, what is the advantage of reviewing before dissemination? In the old system, works whose qualities were overlooked or misunderstood by reviewers may never have seen the light. In any case, given a constant quantity of content, the weeding effort is the same, whether before or after

publication. Why not instead throw all content against the wall and see what sticks?

Peer review's weaknesses have been minutely scrutinized. It often fails to spot problems, and few published authors have been spared the annoyance of a querulous commentator with pointless nits to pick. Academic presses in the US did not undertake peer review until the 1960s.[2] Trade houses usually do not bother even today. Even at serious European publishers, review is informal at best. In digitality, does conventional peer review still serve a purpose? We have looked at whether publishers are best placed to manage review, since they merely muster the author's scholarly colleagues to pronounce judgment. Even more pertinent: why demand prepublication review at all? Review is needed. It is one way the ocean of content is channeled into rivulets of enlightenment. But when should it be done, and by whom?

In the analog world, peer review had to occur before publication. Once issued, the work was locked in place, barring those rare occasions when demand spoke for updated editions. The manuscript had to be as perfect as possible before printing. Digitality has upended that finality. Texts have become more fluid, protean, and revisable. Posted on the web, not locked onto the page, they can be updated. Suggestions for revision are useful whenever delivered. Evaluation at various stages of a text's life cycle is already common in the hard sciences, part of what is known as open peer review.[3] Open review, with many participating and the author responding, was tried out already in the 1960s and 1970s by journals such as *Current Anthropology* and *Behavioral and Brain Sciences*.[4] Kathleen Fitzpatrick's *Planned Obsolescence: Publishing, Technology, and the Future of the Academy* was posted for comment before being issued on paper.[5]

In the sciences, peer review has evolved from a judgment akin to a jury's pronouncement and is now more like a conversation among colleagues. Instead of submitting a manuscript to reviewers, taking

on board one round of comments, and then publishing once and for all, digitality permits flexibility. A preliminary version is posted on the web, commented on by colleagues. Revisions follow. By this point, almost everyone interested has seen the manuscript, and dissemination has effectually occurred. But only now does publication in a technical sense take place.

In the hard sciences, formal publication increasingly matters only to future historians, not today's practitioners. The point of traditional prepublication review was to spare the cost of issuing works not worth it. Now that such expenses have diminished, this falls away. Assessment can occur before publication, as part of it, or afterward. Articles have been criticized after publication, then modified or taken down.[6] In 2010, an article claimed to have discovered bacteria that used arsenic rather than phosphorus in their DNA. After skeptical blogs and tweets, further articles disputing it were published.[7]

The new model of review, in turn, questions the very nature of publication. In the digital age, does prepublication differ meaningfully from publication? Mathematicians, physicists, and computer scientists already work largely through prepublication texts posted online. This was their custom even before digitality made it easy. Since the mid-1970s, theoretical physicists have sent around preprints via ordinary mail, racking up large photocopying and postage bills.[8] Digitality merely turbocharged existing habits. Now content is posted to the web. As mentioned, arXiv is one of the most successful of such sites, with costs of less than $10 per article to host.[9] Since arXiv does not technically publish articles, a better comparison is Scipost.[10] It provides journal certification as well, but still at a fraction of the cost of traditional subscriptions or conventional gold periodicals.

Computer scientists are yet further along this route. They consider even articles passé and too cumbersome to keep pace. Instead, conference papers are the currency of the realm—posted, commented,

and revised at a rapid clip. Their promotion and tenure procedures have adjusted accordingly.[11] Other fields have prepublication sites, too. For social scientists, the Social Science Research Network.[12] For economists, Research Papers in Economics.[13] For medical research, PubMedCentral.[14]

Even more impressive is how urgent knowledge can now be thrown up on the web for use globally. Gene sequencing of the Covid virus was posted early in 2020 on preprint sites before peer review, available for immediate use. Getting information out quickly via preprints had begun already with previous epidemics. Between the Ebola epidemic (2015–2016) and Zika (2016–2017), the proportion of articles with important data appearing as preprints (most then issued as conventional articles, but only several months later) increased significantly.[15] Nonetheless, the number of articles on Zika and Ebola that had first seen the light as preprints was small, less than 4% in each case. In part, they were hampered because only some journals accept submissions that have already appeared as preprints.[16]

During the Covid pandemic, matters improved. Research output grew even faster, whether appearing openly or not.[17] More information than ever was posted on preprint sites as scientists raced against the clock.[18] By February 2020, early in the pandemic, more articles on Covid had appeared as preprints than in journals. The venerable *New England Journal of Medicine* posted one paper within 48 hours of submission. Preprint servers, some researchers finally realized, promised them credit for discoveries, regardless of where they eventually were published, even as they contributed immediately to the public good.[19]

Preprints rapidly disseminated crucial data, but they also raised the issue of avoiding nonsense that led readers astray or hogged attention. In January 2020, a preprint by Indian scientists on bioRxiv pointed to supposedly "uncanny" similarities between the Covid virus and the HIV. Fuelling conspiracy theories about genetic

engineering, it was widely discussed on Twitter by news outlets. Within 48 hours, the preprint had received over 90 critical comments and was retracted.[20] Postpublication review had demonstrated its chops. Information was both promptly disseminated and quickly reviewed.

Other instances have been more ambiguous. In September 2020, a researcher who had fled China posted an article on a preprint site claiming that the Covid virus had been created in a lab.[21] The report was subjected to warnings on the site and harsh criticisms elsewhere. Some were published in a journal set up to combat scientific fraud and misinformation.[22] Nonetheless, the preprint drew much attention, including over a million views and three-quarters of a million downloads by March 2021. Picked up by social media, the article's author did the rounds of the morning television shows, caught up in the politicization of China's role in the pandemic's origins.[23]

Peer review reports can be interesting works of scholarship in their own right. Perhaps they should be opened up alongside the content they evaluate. True, such detail may interest only a few. Yet, as always in digitality, space and storage are largely costless, therefore irrelevant considerations. If it already exists, it might as well be preserved. Also up for grabs is whether peer review should remain blinded, keeping at least the reviewers' identity anonymous. Would revealing identities afflict peer review with the same punch-pulling pusillanimity that has come to plague book reviewing? How genuine can criticism be in an increasingly collaborative academic world, with everyone reliant on colleagues and peers for constant evaluation and review? Or would non-anonymous reviews soothe vindictive spirits, forcing reviewers to temper their words and address actual problems rather than just venting spleens? Conclusions are unclear, except that few academics favor revealing reviewers' identities.[24]

Anonymous peer review allows competitors to throw spanners in each others' works. Once again, this is not an issue for the humanities, where prizes for priority are paltry and scholars rarely work on precisely the same problem. But among the sciences, it is not uncommon for anonymous reviewers asked to evaluate papers by competing colleagues to suggest extensive further work before publication, thus hobbling other teams in their race to the goal. Such were the problems tackled with the inauguration of *eLife*, an open journal established in 2012 by Max Planck, Wellcome, and Howard Hughes. Choosing only active scientists who reviewed under their own names, it hoped to avoid the inherited system's malaise. In the meantime, this approach has become widely adopted as a new gold standard by *Nature* and *Science*, among others.

The Marriage of Reader and Content

Conventional publication puts author and reader in touch through methods both targeted and imprecise. The supply side includes advertising, reviews, citations in others' work, lectures, book tours, and other means of getting the word out. Readers, in turn, seek material of interest via reviews, bibliographies, asking around, and searching the web. In effect, the blind seek the sightless. Only through hard work, perseverance, and luck can one hope to make contact. With digitality and its ever more sophisticated search engines, finding pertinent content has become easier. As content is digitized and tagged, it becomes searchable and findable. AbeBooks has made tracking down obscure used books the work of seconds. Dating apps help those seeking specialized erotic fulfillment or complicated emotional satisfaction. So, too, are curiosity and thirst for knowledge more easily slaked by digitality's ability to pinpoint where to look.

Conventional publishers work on a supply *and* demand model. They hawk their wares as customers seek their choices. In the market, publishers face an almost impossible situation, vastly more difficult than for other sellers. In 2020, the US offered 43 new car models and some 260 existing ones to choose among.[25] Before the Great Recession, Americans bought a new car 13 times a lifetime; now the number is about 9.[26] Over the average car-buying lifespan, consumers thus choose among some 300 products once every five years or so. When shopping for food weekly, over a year, they select an average of 260 different items from among the 36,000 found in a typical supermarket.[27]

Books are much more of a crapshoot. In the US alone, 300,000 new books appear annually. Most are not the kind stocked even in large bookstores, much less piled on the front tables. Nonetheless, the choices are overwhelming. In 2014, Amazon had 23 million distinct paperbacks for sale (many more if you count hardbacks and other media, but that raises the likelihood of duplication).[28] The average reader in the US gets through a dozen books a year, the median reader, four.[29] Even assuming that those are all purchased, the sheer pickiness of the selection in the book market compared to supermarkets, not to mention cars and other consumer goods, is staggering. Books involve four choices among 23 million possibilities annually, compared to 260 food items out of 36,000.

And that is ignoring the three million scientific articles published annually. A decade ago, under cosmology (a subfield of astrophysics, itself a subfield of physics, which in turn is a significantly smaller field than chemistry or medicine), the Smithsonian/NASA Astrophysics Data System listed five times as many articles as Netflix has films.[30]

That reader and book ever hook up is little short of miraculous. Perhaps many readers are stuck in what amount to bad literary marriages to the wrong content. Some may be reading Dan Brown when they would be happier with Ken Follett, or they are slogging

through Thomas Piketty when a little dalliance with A. B. Atkinson would brighten their lives. From this dilemma springs the infrastructure of choice-making that guides readers. What Tinder and Grindr do for sexual selection, the literary dating services supply for readers—from Oprah to Frederic Raphael, from *Reader's Digest* to the *TLS*.

Publishers' marketing is one aspect of this, too, seeking to alert potential readers to something of interest. Amplification is one role publishers claim to play, but as a rule, it does not work.[31] If everyone in an auditorium stands, no one sees any better than when seated. If all content is amplified, the overall sound level rises, but nothing in particular is heard. And even if the publishers do make themselves noticed, they are so conflicted that no one takes them seriously. Blurbs on book covers must be the most devalued form of speech in the democratic world, perhaps barring letters of recommendation. That is tacitly admitted by how they are banished from the paperback edition in favor of—highly edited—excerpts from reviews. Marketing is less of an issue for periodicals. Journal publishers enter the market only infrequently—when subscriptions come up for renewal or new subscribers sign on. Book publishers, in contrast, enter the market with each new volume, incessantly sounding their claxons and clamoring for attention.

But with digitality, the hunt for perfect content changes. Once material is posted, sophisticated search engines and translation software, aided by ever-better techniques of discoverability, help bring reader and content together efficiently and reliably. We move from matchmakers to dating apps. Rather than readers and works hoping improbably to connect in the dark, grazers will now suss out the tender shoots in the field, honing in on the most delectable.

In digitality, certain inheritances from the old world where physical volumes competed for attention fall away.[32] Book covers and dust jackets may soon be remnants of the past. For centuries, books were published without stiff covers, in the expectation that buyers

would bind them to match their library's furnishing. For generations, French books from Gallimard and other houses were issued in stark, elegant, and uniform covers. The Pléiade editions are equally constant—the cream of French literature and thought, uniformly leather-bound on bible paper.

The care devoted to book covers today reveals an increasingly commodified product vying for consumers' attention, much like cereals in the supermarket. The digital world returns us to the Gallimard tradition. Search engines skip the covers and deliver us directly to the title page or, even better, the passage we seek. The fuss about layout, margins, typeface, and other accouterments of the printed page will be left behind as e-readers tailor the page to personal preference. Not only will we bind our books to suit us, we will typeset them. Authors will instead fret over metadata, permanent identifiers, and discoverability, ensuring their message gets out efficiently. Having become electronic files, books will be less and less physical artifacts.

Filtering and Searching: How to Find What We Seek

Information scarcity is no longer our problem. We are awash in data. Scholars are reading more, paying each work less attention as they run faster just to stay in place on the content treadmill.[33] The new challenge is to tame, control, and use the hyper-quantities that threaten to overwhelm our comprehension. Despite more information, our time and attention remain unchanged. How do we find data pertinent to our purpose? Two basic strategies tackle information surfeit: filtering and searching.

Filtering allows others to determine the most useful information; searching permits us to pinpoint data that concern us. Recommendation is a subset of filtering: suggestions of potential interest made by others. Recommendation is becoming systematized and

automated. At the moment, it still remains more annoying than helpful, as our past online purchases chase us around the internet, begging for an encore. More sophisticated recommendations use our viewing, reading, or listening choices to prompt tips for further similar pleasures. In the future, suggestions will likely improve. The algorithms are getting to know us better than anyone. A decade ago, Target analyzed shoppers' choices to identify pregnancies, even able to calculate due dates.[34] Yet, even without that degree of prediction, the algorithms know better than we do what is out there, and they can more accurately satisfy our yearnings by connecting us with actually possible choices.[35]

Some argue that filtering is the main act of managing content surfeit. The question is merely when to apply it—before or after dissemination? Publish, then filter, not filter, then publish—that is the new mantra.[36] While this sounds attractive, it presumes that someone is willing to do the filtering. Clay Shirky has suggested an analogy with dinner party conversations: No one would demand that prandial comments be screened before spoken—though we all remember occasions when that might have been advisable.[37] Conversations and publications are not equivalent. We require more thought of one than the other. Can we say the same of blogs, tweets, and other online communication—more formal than dinner party conversation, less than conventional publication?

Filtering alone does not put us in touch with the material we seek. It assumes that content is arrayed from good to bad, with our goal being to remove the mediocre. That is one way of taming content superfluity, proceeding along a hierarchy of evaluation. Searching is another activity altogether. When readers seek information on a topic, they hope for high-quality results. But that is secondary to locating material on their subject in the first place. Searching separates not good from bad but pertinent from irrelevant. At the first pass, pertinence or relevance require no judgment of quality. That may come later but is not the primary concern.

Searching should also be distinguished from browsing. Browsing assumes an initial selection made by others, whether in a journal on a certain subject or an arrangement in a library or bookstore by theme. Within that preliminary sorting, would-be consumers then skim to make a secondary cut. While adequate for a simpler era with fewer choices, browsing is decreasingly useful. Most readers of online articles arrive directly at their destination, guided by a search engine, rather than navigating from the journal homepage, let alone the publisher's.[38] That may explain why the range of articles cited has tended to narrow, as researchers less frequently chance across fortuitous adjacent works while browsing.[39] Tunnel vision is the outcome. Yet, it does not explain why browsing cannot just as well happen on screen as with a journal issue in hand or standing before a shelf of books. Nor does it explain whether the lowered friction costs of clicking through to references in footnotes does not vastly expand the relevant literature readers are led to through online works.

Searching for pertinent material is necessary, however large or small the total amount of information. We are swamped regardless. Filtering may give the comforting illusion of separating wheat from chaff, delivering only the good, but the quantities remain unmanageable. "There are enough peer reviewed articles to read without having those that have not been," as one researcher put the argument against preprint repositories.[40] That is the delusion—that sticking to peer-reviewed articles allows us to surmount the sheer volume even of those.

Recall the old joke about experts who know more and more about less and less until they know everything about nothing. Meanwhile, the generalists know less and less about more and more until they know nothing about everything. Review is less necessary the narrower the field tilled. There, readers more quickly become experts themselves, no longer reliant on others' guidance. It is the generalists who most need pointers to navigate through expansive

landscapes of the unknown. Not all scholars are equally dependent upon review, though the overall quantities of information in even small fields keep the specialists busy.

Why review at all? In and of itself, review adds little value. Most basically, it calls attention to something deserving that might otherwise escape notice. Among many works on similar topics, more sophisticated reviewing suggests what we should read first, saving time. Given overwhelming and growing amounts of information, our consumption can never be exhaustive. If authors continue writing while we read, even immortals will never get around to everything. Selecting where to spend our limited attention will always be necessary.

Review itself generates more content, compounding the problem. It explains where a work fits in the historiography, weighs its faults and virtues, suggests improvements, and counsels for or against reading it. A review often satisfies readers, who save the time needed for the work itself. Other times, the review devours attention that could have been devoted to the underlying work. A numerical rating system, Michelin stars for books, requiring just a glance, would sometimes be preferable. But fundamentally, the point of a review is to act as a guide for the hunter, bringing us within striking distance of prey, allowing us the satisfaction of a kill without the bother of the stalk.

Curation is similar. The highlights of an artistic genre, a choice of the best works on a subject, of primary documents illustrating an event—whatever the selection may be, it is pulled together to spare us having to duplicate the curator's efforts. Whether an edited volume, a museum exhibition, or a greatest-hits compilation, curation takes us down a shortcut. If anything, curation might be considered the overarching principle, with reviewing a subsidiary strategy.[41]

The curation accomplished by peer review is but one instance of filtration. The publishers' selection is the first step of a larger sorting process. It is followed after release by press reviews, prizes,

edited collections, and other means of calling attention to the best of insurmountable output.

Readers have long paid for guidance through abundance. Reviewers have been with us almost since books first arrived. Many scholars extensively read the meta-content—*NYRB*, *LRB*, *TLS*, and the journals in their field. Nor are the digesters spring chickens. Eighteenth-century publishers issued volumes selecting nuggets from the press, such as *The Gentleman's Magazine* and *Harper's New Monthly Magazine*.[42] German journals of the period provided abstracts of the newest scientific literature.[43] *Reader's Digest* and *CliffsNotes* (now issued by Wiley, one of the most profitable scientific publishers) have been at work for decades. *Blinkist* and *getAbstract* are more modern versions, along with the *Browser* and *Sensemaker*. Scholarly versions include the *Faculty of 1000* (*F1000*) site, which used to rate biology and medicine papers via postpublication review.[44] *Mathematical Reviews* has been evaluating articles since 1940, when it took over a similar function from the *Zentralblatt für Mathematik*, which shunned work by Jews under the Nazi regime.[45]

Even more overtly curatorial are overlay journals. They select from unreviewed content posted on the web. The JMIRx journals (launched in late 2019), for example, curate content posted in medicine, biology, and psychology preprint repositories. Editors find articles, make offers to authors, and consider self-nominations. They add a layer of peer review and typesetting and publish the results in *PubMedCentral*.[46] *Discrete Analysis*, a mathematics overlay journal, links to papers posted on arXiv, indicating that they "have been peer reviewed and judged to be of suitable standard."[47] Overlay journals need not create a new gathering of data unless they want to raise the level of an article's editing or presentation. They can just point readers to already-posted content that has passed muster. That makes them largely indistinguishable from the guides for readers mentioned above.

Filtering allows experts a say over what gets channeled our way. Searching, in contrast, puts us at the mercy of the algorithms, but at

least we control what we are looking for and act as the last-instance sorters. However impressive, today's search engines are a pale foreshadowing of what they must become if we are to tame content superfluity.

Even more crucial, content must be searchable, findable, and reachable. Gone are the days when the vast resources of JSTOR, HeinOnline, and other journal databases were dark to the search engines. At least their content now shows up, even if it still hides behind paywalls. Tagging and making content discoverable have become among dissemination's most important tasks.

Unaffiliated scholars outside the university bubble are frustrated to discover articles they must read but cannot access. Worse, the cost structure of academic papers mocks those who have no choice but to pay retail. For listeners and watchers, Apple prices its songs reasonably, Amazon, its downloads. Not the academic publishers. A review of a book in an OUP journal can cost more than the book itself.[48] Such market tone-deafness says much about why publishing is in a pickle. If they instead instituted reasonable prices and functioning micropayment systems, publishers might even be able to move their content instead of suckling at the teat of library budgets.

Such incompetence also explains why Sci-Hub, Z-Library, and other pirate sites (guerilla or black open access) enjoy massive followings. Sci-Hub has become a darling of the movement.[49] It is now the largest open-access academic resource in the world. After just six years, it hosted 67 million papers, two-thirds of all published research, available to anyone.[50] It violates every conceivable copyright law and continues only thanks to its location somewhere in Kazakhstan, supported by Russia to poke a stick in the West's eye.

However, Sci-Hub has recently come under attack. Litigation is ongoing in the High Court of New Delhi, and Virgin Media has begun blocking access to it in the UK.[51] Good manners require registering a polite harumph of disapproval of this blackest form of open access. Yet, it is hard to avoid seeing Sci-Hub's success as the publishers reaping what they have sowed. The pirate sites' users are

not just scholars denied legitimate access but also professors and students at reputable and connected institutions who find them more convenient than their own libraries' often labyrinthine digital collections.[52] While China is the largest downloader from Sci-Hub, the US is number two.[53] Publishers' inflated fees and cumbersome procedures keep the pirate sites in clover.

The perfect search engine will one day provide a universal index. It will include every word in every work and a means of identifying and locating every image. Granted, it may not pick up concepts that are not expressed in particular terms nor necessarily collect synonyms under a common heading. Strictly speaking, it will be more a concordance than a subject index.[54] But realistically, indexes compiled by hand rarely do that either. And it will solve the problem of languages, such as Chinese, that cannot be ordered alphabetically, therefore indexed only imperfectly.

Anyone who has compiled an index knows that names, places, facts, and specific substantives are easier to include than vaguer concepts and ideas. In Václav Havel's play *The Memorandum*, a new language, Ptydepe, is invented to add precision and avoid homonyms, words that sound alike. One consequence is a proliferation of extremely specific terms for concepts that might resemble each other in natural language. Such a tongue would be an indexer's delight, at the very least sparing us the need for tens of thousands of Wikipedia disambiguation pages. In its absence, we must rely on more sophisticated searching. As that improves and content is better tagged, search engines will do our bidding more dexterously. Fine-tuning for results by language, format, provenance, or dates will become child's play.

We have touched on Michael McCormick's ability to wrest from Widener library's otherwise mute tomes evidence of trade between the Arab world and Europe in the eighth and ninth centuries. In the late 1990s, it took him and his students a week of shoe leather. Once Widener's contents are fully searchable, a similar investigation

should be a matter of hours and turn up evidence further down the long tail. That will make Widener less a library and more a database. We are almost there. Google Books has digitized a Widener-sized chunk of content. Only copyright law and the publishers' veto hinder its being put to full use.

As the mega-journals approximate smaller versions of a future global bulletin board, they suffer the tension between filtering and searching. With size no longer a constraint, specialization serves no purpose. Journals or edited volumes focused on particular subjects acted as a preliminary filter. Editors did not have to evaluate submissions outside their remit. The narrower the topic, the more manageable their workload. With active searching rather than passive filtering, however, such needs fall away.

While the mega-journals may foreshadow what is to come, they suffer teething problems of their own. *SpringerPlus* was perhaps the closest mega-journals have come to the global bulletin board, publishing indiscriminately across a Noah's ark of different fields. In the meantime, Springer has shut it, concluding that both humanistic and technical scholars prefer journals more tailored to their subjects.[55] So long as specialized and omnivorous venues coexist, the former will have a leg up. Only when specialization confers no advantage in channeling attention will narrowly focused journals go the way of the Victrola. Journal specialization is, as noted, just a first approximation of indexing and searching. For books, the same holds for tables of content and indexes. Improved search engines will end such crude filtering. Indeed, at the logical extreme, neither books nor articles will need titles.

As content is searched across a massive accumulation like Google Books, the works become less important than the whole. The engine delivers the results, and it matters little precisely whence they stem. The source of a fact or an idea remains important to understanding its context and possibly its validity. A danger in this brave new world of commodified memes will be failing to understand what

is meant when the search engine delivers a disembodied snippet of text. Were Irish children really to be eaten, or was Jonathan Swift being ironic? When first developed in the sixteenth and seventeenth centuries, indexes were attacked as diverting readers from the entire work to mere excerpts. Swift coined "index learning" as a term of contempt. Some authors refused to compile indexes lest readers shirk plowing through the entire text.[56]

Books and articles are composed of smaller units, whether assertions, facts, ideas, or memes. What search engines bring us are less broad ideas—by their nature hard to identify, localize, and trace—than compact, discrete units of meaning, some factual, some conceptual, some argumentative or rhetorical. The search algorithms handle "Who was the best-known constitutional lawyer in nineteenth-century Argentina?" better than "How do constitutional differ from civil rights?"

Under the search engine's dispassionate gaze, works will decompose into their constituent memes. Once digitized and searchable, every text's identity dissolves into the mass of all content. Precisely to which conventional work—book, chapter, article, poem, or blog—the search engines deliver us will be less important than its content. Authors working with the omni-searchable mass of global content will pick and choose with little concern for immediate provenance. As Kevin Kelly said about Google Books, "Once text is digital, books seep out of their bindings and weave themselves together." Digitally combined, books will merge into the "collective intelligence of a library." Together, all books become one massive tome, "a single liquid fabric of interconnected works and ideas."[57]

Those who worry that this bodes ill for understanding ideas in their context may be heartened that few such predictions have yet materialized. A decade ago, a company named Citia was in the business of dissolving books into their component memes, the better for users to reconnect and use them as they saw fit. Today, Citia has become a corporate communications software firm.[58] Perhaps

more comforting still: such decomposition has occurred since the subject index was first invented in the early thirteenth century. Robert Grosseteste's *Tabula distinctionum* collected references to subjects (that God exists) across the Bible, the Church Fathers, pagan authors, and Arabic writers.[59] Insofar as the effects are bad, we have long suffered them.

What epistemological effects will search engines have? Delivering facts more readily than ideas, will they focus the future's clever minds more empirically? Theory is a means of spanning the gap between facts. It connects them into causal narratives that explain why this, and not that. The fewer the facts at hand, the broader the gulf theories must bridge, the more explanatory work they must perform. Conversely, the more data points we have, the less arching and ambitious theory can be—at least if it aims both to have causal power and account for myriad facts.[60] Easy availability of endless data, our ability to slide ever further down the long tail—will that sap the appeal of grand theory? Will future theories, strapped ever-tighter to ever more granular factual underpinnings, necessarily be less ambitious? Two points define a line. That is simple, powerful, and appealing—but also based on de minimis data. The best we can hope for from a wealth of data, in contrast, is that it clusters, indicating a trend. If we are lucky, it suggests the likelihood of one possible explanation. Being dogmatic is harder as data multiplies.

Evaluating, Not Publishing, Is the Goal

One solution for open dissemination is a global bulletin board, a vast repository where everything is first posted. The mega-journals have already moved in this direction. The networks and consortia of repositories are close behind—organizations such as OpenAIRE in Europe or LAReferencia in Latin America. Because digitality removes size constraints, subject specialization is unnecessary.

Selection takes place after publication, not before. Any topic is welcome.

Posting and publication should be distinguished. Posted, everything can be read in typescript. After this, improvements can be added depending on customer demand, author wishes, or whatever motivates a closer scrutiny of a text's claims or a jazzing-up of its presentation. More important than finding the grain in the chaff is locating what interests us amid the irrelevant. With a global bulletin board, getting words before readers will no longer be the problem. We will have achieved peak dissemination as a steady state.

In a deafening cacophony of content, how do we decide where to start?[61] The more focused and precise our interests, the smaller the problem. For scholars homed in on a microtopic, the concern is more finding information than judging which to begin with. Those in pursuit of broad issues most need evaluative aid. Peer review in its traditional prepublication sense is but a partial answer. While helpful, nor are postpublication guidance and curation a full solution. The horizons of evaluation must expand. So far, reviewing has followed the Michelin guide model, with experts sampling the wares. A more Zagat-like approach might be equally useful, where everyday consumers pool crowd wisdom. The disadvantages are obvious—amateur reviewers, including bloviators and ranters. If anonymized, will the reviewers' worst instincts well forth? If anyone can evaluate and comment, who reviews the reviewers? But if we insist on expertise, where will we find enough?

Nonetheless, review by the vox populi may have its role. Aggregated and averaged, with outliers lopped off and extremes smoothed, this approach may prove useful, much as stock market movements convey information. As some works go viral, up-voted by readers, the Zagat approach may unearth sleepers overlooked by the mandarins. Conversely, Reddit-style voting may also deliver a comedownance to the overinflated reputations of eminent but now complacent authors.

All this presumes that works are read and rated. How likely is that? Publishers claim that peer review is their most important contribution. The distinction between publication and self-publication hinges largely on such vetting. Self-publishing a book on Amazon can be done very inexpensively, so if getting the word out were the only goal, this would be an obvious route. Yet, what self-publication lacks is a major element of the academic prestige economy.

Before Amazon, authors who self-published did so at so-called vanity presses. Such houses were paid to issue whatever came over the transom. No one who could squeeze their manuscript past the lions guarding the gate at any conventional publisher would have gone this route. Yet, in the nineteenth century, authors still commonly bore the costs and risks of dissemination, akin to self-publishing. Henry David Thoreau convinced his publisher, Ticknor & Fields, to assume the costs of his second book, *Walden*. That took some persuasion, given that his first, *A Week on the Concord and Merrimack Rivers*, had done so poorly that he stored 600 unwanted copies in his attic until he could sell them back to the publisher.[62]

One hurdle open-access publishers contend with is the public's confusion of them with vanity presses. The new houses' claim to scholarly integrity is based on peer review. That, in turn, justifies the higher costs they incur. Yet, in the meantime, the stigma attached to self-publishing has faded. As seen, self-published editions now dwarf conventional books in the US. No more than a stream of selfies is vain and preening is self-publication considered self-regarding. After all, only a small fraction of works emerge from university or other academic presses, having run the gauntlet of peer review. Most conventionally published books do not undergo much prepublication scrutiny. That adds to the reasons why post facto review is crucial.

Jumping through the hoops of peer review is just the first barrier to surmount—the admission ticket for the cosmic raffle of attention gathering. Anointed by its publisher, the work emerges onto

the field of battle that is the marketplace of prestige. Even in scholarly publishing, most vetting occurs later. The reviews, prizes, fellowships, sabbaticals, grants, conferences, invited talks, and other emoluments that the scholarly world bestows on its favorites all follow publication. The priority of postpublication review is much less of a change than peer review's defenders would have us believe.

But is work reviewed after release? Hume knew that a scholar's nightmare is not to be criticized, but ignored. As in a marriage gone bad, even anger is better than the cold shoulder. Scientists rely on work being noticed. Priority is crucial for reputations, prestige, and prizes. The first to discover something enters the history books, others remain also-rans. Priority can be asserted retrospectively against a late-comer who managed to be noticed first, but being seen as first through the door is far better. The gentlemanly accommodation between Darwin and Alfred Russel Wallace was unusual, including a joint paper to the Linnaean Society in 1858. It contrasted with Darwin's conflict with Richard Owen.[63] Establishing priority by posting a manuscript is a great advantage that explains why scientists have readily taken to prepublication repositories. Curiously, scholars also worry that papers in preprint repositories will be plagiarized, even as they cement their priority.[64]

Establishing priority is crucial for each researcher's career but less for a field's overall progress. Contrary to what Romanticism's individual genius theory of creativity would have us believe, intellectual progress is a collective endeavor. Advances happen independently of any one researcher. If professor X is run over by a bus, colleague Y—working on adjacent issues—would soon arrive at similar conclusions anyway. That is apparent as science becomes more collective, carried forward by large teams. No scholar is irreplaceable.

It is revealed most clearly where trivial necessity clamors for attention—where the market demands quick and easy solutions to mundane problems. Once cars had been equipped with tires early in the twentieth century, exchanging them quickly and easily became

pressing. Largely simultaneously, seven different people invented demountable rims for this purpose.[65] The race to the patent office was determinative. Advances in treating osteoarthritis of the knee have arrived in close, almost photo-finish, succession.[66] This holds more broadly, too. In disciplines drilling away at nature's coalface, many investigators are about to make the same breakthrough at any given moment.[67] Newton and Leibniz arrived independently and proximately at calculus. Three teams of two physicists each published papers within weeks of each other in 1964 showing how particle carriers of force, such as photons, could gain mass—part of the work that eventually identified the Higgs boson.[68]

The correspondence between research and reality is less direct in the humanities and social sciences. The fruits of their work are less breakthroughs in understanding something "out there" and more a subjective interpretation of human-centered events that are themselves reciprocally influenced by how they are understood and, in any case, open to legitimately divergent understandings. Yet, broadly the same overarching functionalist logic as in the sciences holds. Creators work within epistemic bubbles that influence what topics seem important and which conceptual tools are useful.

During the 1970s, Saul Kripke was not alone in working on concepts of identity and essential characteristics across different circumstances. Ruth Markus and Alvin Plantinga did too. But when Kripke coined the term "rigid designator" to specify something that holds across all possible worlds, the need to assert priority was less pressing than in the natural sciences—as suggested by his relaxed approach to publishing. Neither patents nor prizes are promised those quickest to publish. Yet scholars here are keen to be associated with breakthrough concepts. "Prisoner's dilemma," "performative utterance," "excluded middle," "inferiority complex," "conspicuous consumption," "creative destruction"—all are seminal ideas attributable to specific thinkers who have expanded our horizons by crafting intellectual tools.

As in the sciences, thought here, too, moves with the herd. As melody and harmony dissolved, at some point, something like John Cage's *4′33″* would have been written. And indeed, Cage was not alone in proposing silence as the best response to music's travails.[69] Yet, it is best not to overstate such claims. The songwriter George M. Cohan, asked by a Senate committee how he came to write "Over There," answered, "It was just a bugle call. If I had not written it Thursday, someone else would have written it Friday. In other words, it had to be written."[70] That was nonsense. Some other song would have been written, and no doubt was. It would have been in much the same style, but it would not have been exactly that one.

That Cohan was exaggerating is the entire premise of copyright law. Authors can monopolize their particular expressions of creativity precisely because they are singular and unique. That distinguishes copyright from patents. A similarly extended hammerlock on ideas is forbidden since it would freeze creativity. Patents are granted only briefly because they monopolize ideas that could have occurred as well to others. Therefore, they can belong only temporarily to the person who first happens to think of them, or at least to register them. Individual expressions, in contrast, can belong to authors for much longer because they are specific and thus less important than a general concept. You can monopolize "My Funny Valentine," but copyrighting jazz would bring civilization to a screeching halt.

Postpublication Review

It is all well and good to recognize the flaws of conventional peer review, acknowledging that postpublication scrutiny is more important. That still leaves the question of whether it takes place enough to be useful.

Posting largely everything that meets minimum quality standards, allowing it to be evaluated subsequently, if ever, in effect already occurs in the mega-journals. *PLOS One* accepts manuscripts that meet certain technical criteria of presentation, language, and methodology.[71] It specifically does not ask about a manuscript's significance. Most submissions vault this hurdle. *PLOS One* and other mega-journals accept between 50% and 70% of submissions. These rates are only somewhat higher than for conventionally peer-reviewed journals, except the most prestigious, and are comparable to those of more specialized open periodicals.[72]

Vast quantities of research thus issue forth. For a while, *PLOS One* was the world's largest journal. Such outpourings are then left to readers to evaluate, sort, and use.[73] Review follows publication. Users determine the work's value, not the publishers. *PLOS One* and other mega-journals are thus arguably as much repositories as journals. Other journals are more overtly curatorial. In 2021, *eLife* began publishing only articles already posted as preprints, issuing the reviews along with the main text. As an overlay journal, that made *eLife* less of a publisher and more of a referee and certifier of content.[74]

Whether filtering or searching, we need better means of getting to pertinent information. It is worth distinguishing among varying levels and qualities of information. Not everything has to be read. Some research seeks to confirm or replicate already-discovered results. It is valuable if it does, and even more if it does not, calling accepted conclusions into question. Though such outcomes are useful for the interested, they do not need to be sorted and filtered for most readers.

Techniques for attention-signaling can be external to the content but also built into the text itself. Harnessing typological conventions, we can distinguish crucial from skimmable and skippable material. We signal passages needing *careful* attention with italics—sometimes just a word or phrase, occasionally a sentence or two.

Capitalization serves a similar function in phone texts, as President Trump reminded us daily. Inexplicably, quotation marks have taken on a similar use for emphasis on signage. The difficulty of italicizing in e-mails is a daily frustration for many, who resort to *bracketing* important words in asterisks and the like. In the nineteenth century, German books sported multiple levels of attention-signaling. Text to be emphasized could be italicized but also double-spaced with extra room between each letter, l i k e t h i s. Oddly, that made reading harder, since the eye grasps the word less easily if letters are spread out. Perhaps, therefore, this convention has not survived. For extra whammy, German authors also combined italicizing and double-spacing.

That amplified text. Conversely, other conventions indicate second-order content. Foot or endnotes customarily provide the source of ideas or quotations, but often they take on a life of their own as a counterpoint to the text above the page's Plimsoll line. Sometimes they are as interesting as the main content. As mentioned, legal scholarship revels in extensive footnotes. They are often used to hash out historiographical or methodological issues that are insufficiently captivating to deserve the foreground. Novels have played with footnotes as a narrative device: *Sartor Resartus*, *Peter Pan*, *Pale Fire*, *Ficciones*, *Tom Jones*, *Tristam Shandy*, *Finnegan's Wake*, and many others.[75]

Here, too, German publishers of the nineteenth century were inventive. Their books had extensive swaths of text—easily half—indented to indicate supplementary material to read for further enlightenment but dispensable for following the argument's thrust. E-books open up new possibilities for such high- or low-lighting of text. Nor are the author's decisions the only ones. Readers, too, could supply pointers on what was worth attention or not.

Chapter 24 of the third book of *Tristam Shandy* is a reversed meta-use of such techniques. The narrator claims to have excised the novel's best chapter so as not to cast the rest in its shade. Also

possible are all manner of annotation—by authors, editors, or readers. Institutionalizing such text hierarchies would be useful. They would help navigate works, supplying readers pointers on how and where to delve to their level of interest. In effect, they enlist the author's help in skimming. Books suitably signposted for rapid reading would not need shorter, article-length versions for wider consumption.

We need a ranking of content in terms of its claims on our attention. Readers should be equal participants. We could take a leaf from Michelin guides—not the red ones for restaurants and hotels, but the green city and regional ones. Their tripartite hierarchy (interesting, worth a detour, worth a trip), suitably modified, could be expanded. Reviewers and readers would contribute their two cents' worth. Reddit's ("the front page of the internet") system of up- and down-voting postings might be a model. The best content, evaluated by different groups of participants, percolates to the top. Wikipedia and Slashdot also enlist participant reviewers, with trust and influence built up by their performance.[76] The Chinese repository, Sciencepaper Online, reviews articles, with readers assigning stars. These are then considered in evaluating academic performance.[77]

Postpublication review takes time and resources. As more works issue forth in a new low-filtering environment, even greater efforts will be needed. In the UK, research funds are allotted after a quinquennial research assessment ranks university departments. In 2008, its costs were £12 million directly and another £47 million for universities to prepare.[78] That itself would pay for a good chunk of research. The assessment establishes each field's pecking order but does not even buy the general reader a list of the best academic works. In any case, given such costs, why pay both for pre- and postpublication evaluation?

How much content is reviewed after publication? If post-facto review is to be a viable alternative, it has to take place. True, prepublication reviews are the opinions of only a few readers, and

editors may not always know of the best experts. But at least works are being drawn through some kind of comb. For postpublication review, there is no guarantee. Once content has been posted/published, no one can promise it will be usefully commented on or even read. In 2013, the NIH launched a pilot service to elicit comments on the 22 million articles in the PubMed database. When only 6,000 drew any attention, the project folded in 2018.[79] The vast majority of *PLOS One*'s articles remain uncommented-on.[80]

What percentage of content should elicit reactions for postpublication review to be considered a success? Many writings were reviewed before publication in the old system—even those we do not know of because they never appeared. But in the brave new world of global research participation and ever more authors and output, 100% review is unlikely. We cannot all be authors and reviewers, too. Hyperprolific researchers overburden the system. Perhaps a horse trade needs to be negotiated. The amount anyone may publish could depend on how much of others' content they also review. Tyler Cowen suggests capping researchers' output, requiring them to review instead, thus supposedly improving the quality of a reduced quantity.[81]

Athena Unbound

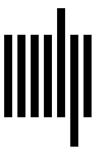
Why and How Scholarly Knowledge Should Be Free for All

By: Peter Baldwin

Citation:

Athena Unbound: Why and How Scholarly Knowledge Should Be Free for All
By: Peter Baldwin
DOI: 10.7551/mitpress/14887.001.0001
ISBN (electronic): 9780262373968
Publisher: The MIT Press
Published: 2023

The open access edition of this book was made possible by generous funding and support from the author



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in ITC Stone Serif Std and ITC Stone Sans Std by New Best-set Typesetters Ltd.

Library of Congress Cataloging-in-Publication Data

Names: Baldwin, Peter, 1956– author.
Title: Athena unbound : why and how scholarly knowledge should be free for all / Peter Baldwin.
Description: Cambridge, Massachusetts : The MIT Press, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022027103 (print) | LCCN 2022027104 (ebook) | ISBN 9780262048002 (hardcover) | ISBN 9780262373951 (epub) | ISBN 9780262373968 (pdf)
Subjects: LCSH: Open access publishing. | Scholarly electronic publishing.
Classification: LCC Z286.O63 B35 2023 (print) | LCC Z286.O63 (ebook) | DDC 070.5/7973—dc23/eng/20220628
LC record available at https://lccn.loc.gov/2022027103
LC ebook record available at https://lccn.loc.gov/2022027104

10 9 8 7 6 5 4 3 2 1