Disease Models & Mechanisms 2, 100-102 (2009) doi:10.1242/dmm.002758

Drinking from the firehose of scientific publishing

Chris Patil[1,*] and Vivian Siegel[2,*]

The fundamental question is this: can the wisdom of crowds be exploited to post-filter the literature?

Before we tackle the future of scientific communication, a few words about the present might be in order. Let us begin with two simple premises:

Premise 1: the amount of literature is frighteningly large and constantly growing. Within any scientific field of sufficient intellectual depth to deserve the name, there is far too much happening in the literature for any individual to absorb, digest and assimilate it all without making that task their full-time job. (That outcome is to be avoided, since the individual in question would then become useless for whatever purpose or calling initially motivated them to understand the literature in the first place.) Although the review articles and highlights of the literature that appear in various journals make some attempt to look beyond the top journals for gems (e.g. *PLoS ONE* articles are picked up with some regularity), they cannot provide a comprehensive scan of any individual field.

Premise 2: scientists are laughably anachronistic in their approach to information. The internet has completely revolutionized the way that everyone accesses information – everyone except for us academics, that is, who continue to organize and locate information in essentially the same old ways. In the old days, we published papers in journals and located other people's articles using citation indexes and PubMed terminals (or, if we're really old, Medline and Current Contents). We made lots of photocopies. Now, we publish papers in electronic journals and locate articles using citation indexes and PubMed *with web interfaces* (although more and more of us prefer Google). We download lots of PDFs and make lots of printouts. The really technologically savvy among us might eschew the print and make digital folders on our laptops. The 'information age' has left us unchanged, except that we no longer have to walk to the library or buy copy-cards.

Academic scholarship has fallen far behind the times, at least compared with such noble pursuits as online shopping. A few fields, such as physics and mathematics, are making progress; anyone can freely read manuscripts submitted to arXiv. Although this may overturn the pre-filter function of the literature, it does nothing to solve the dire need for a post-filter: how is the individual consumer of information supposed to select among the huge and growing body of literature to identify the articles that best suit his/her needs?

Because we are writing about The Future™, our proposed solution will of course involve Technology®. The devil is in the details: what flavor of technology will best satisfy our hunger for efficient access to the most relevant information? In other words, what should a useful and practical post-filter look like?

During the course of sharing our ideas on this subject, we will apply the typical straw man approach and raise and dispatch a few weaker proposals. Here's one now:

This is the first of a series of editorials by C.P. and V.S., imagining the future of scientific communication and scientific research in the age of Google, Amazon, Wikipedia and Facebook. C.P. blogs about his primary research interest at 'Ouroboros: Research in the biology of aging' (http://ouroboros.wordpress.com).

[1]Chris Patil is at the Buck Institute for Age Research, Novato, California, USA
[2]Vivian Siegel is at the Department of Medicine, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
*Authors for correspondence (e-mail: cpatil@buckinstitute.org; dmmeditor@biologists.com)

Straw man solution 1: really good text mining tools. Software text mining has come a long way and is actually pretty good; if you know that you want an article about subjects X and Y but not Z, modern data mining utilities can reliably retrieve every single article in every available database matching that description. There are ongoing efforts to make these tools cleverer using natural-language processing to evaluate the semantic relationships between the entities mentioned within an article in order to improve the quality of search results. But no matter how finely we hone these tools, they can do little to help us sort through the large number of descriptively similar papers pertaining to our original interests in X and Y but not Z. Which ones should I read? Which ones should I delete? Simply, which ones are 'best'?

The reason why search tools ultimately can't help us with this question is that quality isn't a straightforward function of a document's text content. The quality is in there, certainly, but it's latent until evaluated by a human mind. Unfortunately, we're confronted with a large amount of literature and we're trying to save time, so reading every paper in order to figure out which papers to read is a non-starter. Instead, we want to quickly and easily access this 'latent content' about a paper's quality before we devote a precious aliquot of our scarce time to reading it closely.

One way to discriminate among apparently similar objects is to rely on the experience of others. Search engines like Google make a good start at doing just that: search rank is based on linkage, so it potentially represents a measure of how useful others have found a document in the past. This need not be the case, however; other users may have linked to a document for reasons other than its high quality. Furthermore, as with citations, links take a long time to accumulate. How might we best harness other people's assessments and behavior in something approaching real time? One possibility would be to allow others to spend a good deal of their own time to filter the literature within a field and share their opinions with the world.

Straw man solution 2: blogs. A lioness doesn't bother eating individual blades of grass – she lets the antelopes do that drudgery, and then she eats the antelopes. It is similarly tempting to assign the post-filtering task to hordes of enthusiastic volunteers – intrepid, pajama-clad souls, armed only with keyboards and search engines, who would wade through the jungle of the literature and return to us only the choicest prizes. But this is a fantasy. For bloggers to provide an efficient and efficacious post-filter service, they would have to meet an imposing list of qualifications: sufficiently well-trained to make wise judgments about the papers most worthy of attention; sufficiently idle to have nothing better to do than read papers all day; free of idiosyncrasy or agenda that might bias their choices; and willing to work continuously for free. (In other words, there won't be 'hordes'.) Add to that the need for competition between bloggers – comparative prestige being the coin of that murky realm – and soon we'll find ourselves combing through myriad blogs in order to make sure we're reading the best one. And then we'll write a column about the need to post-filter the blogosphere.

Let's consider another approach.

An Amazon for science publishing? Consider shopping on Amazon.com for a common appliance, for which many nearly equivalent options are available. After a simple query, we can see all the options (at this point in the analogy, we're exactly where our top-of-the-line text mining script would have gotten us), but we can also see the reactions of previous customers. The alternatives are ranked, and the rankings are backed up by prose descriptions of buyer experiences. We're also treated to richer information about other people's behavior, e.g. other items that they browsed or purchased after evaluating the item that we're now pondering for ourselves. The crowd is post-filtering the options; their aggregate actions reveal the latent content about the quality of the

bicycles, lawnmowers or neck massagers that we might purchase. We just skim their words and make a good decision informed by their past experience.

Could we establish such a system for the literature?

Suppose that rich metadata accompanied papers, detailing their classification – questions asked, methods used, model organism used – as well as information about the reactions of every other reader of that paper. A few days after you download a paper, it asks you some questions about itself, perhaps triggered by your own annotations. (Remember, papers are digital entities; they could be so much more than passive sacks of kilobytes.) 'How would you rank me on a scale of 1 to 10?' 'Was I classified correctly?' 'On the basis of prior reader commentary, the original legend to Fig. 4B was confusing. Did you find the new annotated legends clear?' 'Can you recommend other papers that readers of this paper might find interesting?' With your answers in hand, the paper would 'phone home' and communicate your recorded evaluations, comments and annotations to its mother ship, which in turn would communicate those experiences to every other copy of the paper in the world.

Such massively parallel, bottom-up, user-driven annotation would handsomely meet the need for a post-filter of the literature – far better, certainly, than the straw men we introduced above. Different aspects of the user experience could be distilled into metrics of overall quality or usefulness, which could in turn be used as a filter by other end-users; the reader-improved classification could further increase the likelihood that a given paper would find its ideal audience. Because it relies on human rather than machine intelligence, the approach would not suffer from the limitations of literal-minded text mining tools, which are unable to uncover latent content about the usefulness of a document to human users. Because the effort involved in evaluation and sharing is small compared with the overall effort of reading a paper, individual users would experience only a minimal energy barrier to participation. Furthermore, because of the distribution of effort, the impact of individual idiosyncrasy would be minimal and there would be no risk of a bottleneck when the field's bloggers have to take orals or go on holiday.

By this point, we have certainly given you reason to ask, 'are you crazy?' We ask ourselves the same question. The technology required exists now – leaving aside the science fiction of the solicitous document, a slightly less pushy version of the approach described above could be implemented using a web browser – so there is no technological barrier to adoption. Someone would have to design the system, but as Amazon and other online shopping giants have demonstrated, this is a soluble problem with a lot of the kinks already worked out in the private sector. Granted, individual publishers would have to cooperate, which may seem an insurmountable barrier at first; but, they have already demonstrated their willingness to do so with features like CrossRef, which allows you to follow a chain of references from paper to paper, even across competing presses.

The fundamental question is this: can the wisdom of crowds be exploited to post-filter the literature? If not, is the barrier at the level of the product, or the end-users, or both? Is there something qualitatively different about academic papers, or academicians themselves, that would make it impossible to adapt the ways we have come to signify and communicate the quality of commercial products?

Finally: if we scientists fail to adapt, and to efficiently use the tools that have become available, what will it mean for the future of science communication?