Reassembling Scholarly Communications

Histories, Infrastructures, and Global Politics of Open Access

OA Funding Provided By:

- Arcadia Fund
- Birkbeck, University of London

The open access edition of this book was made possible by generous funding from Arcadia—a charitable fund of Lisbet Rausing and Peter Baldwin.

The title-level DOI for this work is:

doi:10.7551/mitpress/11885.001.0001

19 Reading Scholarship Digitally

Martin Paul Eve

Scholarship, Labor Power, and Proliferation

In the present moment of 2020, more scholarship and research is published every year than it would be possible to read in a lifetime. The open-access mega-journal *PLOS ONE*, for example, publishes 20,000 papers per year alone.[1] This is not necessarily a bad thing; it may be that high volumes of publication are beneficial to the scientific endeavor and that this volume represents a healthy global research ecosystem. Such a volume does, though, pose a serious challenge for the contemporary researcher, even when one is speaking only of a single, subdisciplinary field.

Namely, the difficulty faced by the contemporary researcher is as follows: how is it possible to keep up to date with the most recent research and scholarship, amid competing demands for time in the saturated life of an academic? How, with a scarce volume of labor time, is it possible to know that one has read all of the most recent and relevant research and scholarship?

The problems of this environment of proliferation are abundantly clear already in academic hiring panels, although the digital solutions that I here pose will not solve this particular case.[2] Faced with hundreds of candidates per post, it becomes near-impossible for panel members to *read* all of the scholarship before them. In the humanities, the prospect of reading 200 monographs to appoint to a junior lectureship is simply beyond the realm of possibility. In the sciences, one could say the same of journal articles or conference proceedings.

It is from this challenge that proxy measures such as the notorious journal impact factor (JIF) sprung. These aggregate and insensitive measures of citation statistics were designed to assign quantitative value to specific

venues. In other words, they moved from the evaluation of the specific article to an evaluation of a scarcity correlation in the container. For, if it can be presumed that only one in 200 papers is admitted to a journal, then that publication outlet can act as a perfect correlation for the scarcity that faces the hiring panel, with 200 applicants for a single job. Since JIF is premised on a scarcity—as it is calculated as citations against volume—this scarcity becomes important.

The problem is that such aggregation to the journal level is deeply flawed on several levels. For one, Brembs et al. have recently contended that the JIF correlates most closely with retractions.[3] For another, such scoring restricts academic choice and freedom in publication venue; if academics and their managers believe that certain journals will be used in their evaluation before hiring, promotion, and tenure committees, they will flock to publish only in such venues and will feel a pressure *not* to publish elsewhere. This can create a set of additional market problems for library budgets in the ever more restricted and almost monopolistic situation that has fueled the serials crisis since the 1980s.[4] Such methods of evaluation are also problematic in their aggregation since every "top" journal has published bad research and every "poor" journal could, in theory, contain brilliant articles.

To avoid these negative situations, the San Francisco Declaration on Research Assessment (DORA) was born, whereby institutional signatories agree to avoid the use of JIF-like proxy measures for their appointment panels.[5] This goes some way toward resolving the unintended consequences of the JIF, but it doesn't then answer the more fundamental question of what lies beneath the development of this measure: how can we know how to spend our reading time, without actually reading the work itself?

One suggestion for how we might fix this is to move to a mode of assessment where candidates for hiring present a research narrative in which they outline the impact, outcomes, and overall arch of their research, referring to a couple of key outputs, to which a hiring panel might turn and read in detail (the kind of "ImpactStory" approach). This sounds good in principle, even with the entirely valid concerns about the Impact agenda in the UK. (In the UK context, "impact" refers to demonstrable behavioral change in response to research and it is measured as part of the Research Excellence Framework (REF). This is controversial because it places an emphasis on translational, rather than early-stage, research. It also seems to demand that research change the world, rather than people's understandings, which can

be hard in the humanities and social sciences—although in the 2014 REF, these disciplines fared well nonetheless in impact assessments.) It reinforces the importance of understanding why we do research and what the work told us, while also moving away from relying solely on the prestige of the venue in which the work appeared.

The problem with this is the onus it puts on candidates. Applying for academic jobs is arduous, unpaid work, with only a slim chance of a payoff. The dilemma then becomes: in implementing initiatives such as DORA through displacing the burden onto researchers/applicants to narrativize their work, the academy achieves some good. It is good that researchers should think more broadly about their work and how they can articulate this to a wide audience. This also gives those with a more quirky, non-prestige-based track record a better chance of employment in academia (at least in theory).

On the other hand, this approach asks candidates to take on more work, in order to spare the work of hiring panels (who are employed members of staff). If candidates have disabilities, (child)care responsibilities, or a host of other life circumstances, this method once more privileges those who can afford to put the most time into a gamble on an academic job. My conclusion from this thinking is that we need new ways to search and appraise scholarship.

Such an approach would not especially help with the problems of evaluation into which I have delved in this introduction; the assessment of the importance and quality of research work without recourse to crude metrics remains a difficult task. But it could help with the rigor of research and scholarship, which frequently does not and cannot cite the secondary literature comprehensively, since discovery has become so hard in an age of open abundance. In other words, while evaluative circumstances are among those where the demands on our reading time are most clear, this is only really a reflection of a broader problem in the general research environment, with which a range of computational approaches could assist.

Distant Reading Methodologies

This problem of abundant material and scarce time is not distinct to scholarship. In the fields of history and English, for instance, various digital methods have been born under the name of "distant reading" to attempt to solve this problem of insufficient reading labor-power.[6] In the sociological

study of social media and the web, the computational solution would be called "text mining." JSTOR Labs has also recently released an example platform that allows for the digital close and distant reading of scholarly material within their database and has been thinking about alternative digital approaches to the monograph.[7] The fundamental premise of such methods, though, is to use digital techniques to scan through hundreds of thousands of papers, articles, or books, and to bring pertinent work or aspects to the attention of the operator.

One prominent group of scientists who are already embedded in such a culture is the Murray-Rust research group at Cambridge University. In 2014, Peter Murray-Rust, a crystallographer by background, was awarded a Shuttleworth Fellowship for his work on a suite of tools for the extraction of facts from the scientific literature: the ContentMine.[8] Working strictly within the bounds of the law—yet exploiting the exemption that facts cannot be placed under copyright, only their expression can—this nonetheless has the potential to revolutionize how we search academic literature at scale.[9]

For Murray-Rust, the benefits of mining the scholarly literature can be summarized as follows:[10]

- Comprehensive coverage of the secondary literature. At present, in all disciplines, work can go unnoticed or uncited, causing problems of repeated work and duplicated argument. A system that could comprehensively search the scholarly literature would avoid this.
- Comprehensive coverage within a paper. Scholars often read only parts of a work, for time, rather than reading the whole piece. This problem could be mitigated by a system such as that proposed by Murray-Rust that would summarize the entire argument of a paper and ensure coverage of the complete work.
- Aggregation and interdomain analytics. The example that Murray-Rust gives here is the fact that we are currently poor at cross-referencing information. For instance, consider the question: "What pesticides are used in what countries where Zika virus is endemic and mosquito control is common?" This is hard for a person to answer, but relatively easy to aggregate computationally when one has related documents.
- Semantically rich entity tags. Connecting terms that are used in the literature to other sources has the potential to greatly accelerate the research process in many domains.

Murray-Rust believes that his activities in mining the scholarly literature in this way are covered by the Hargreaves amendments to UK copyright law in 2014, which cover his development of the software, but he cannot be utterly sure. Indeed, a lot of time at the ContentMine project is clearly dedicated to ensuring the legality of what they do, the majority of which is due to the fact that the copyright to most research material is owned by publishers.[11]

This is also complicated by Technical Protection Measures (TPM) and Digital Rights Management systems, which more publishers are now employing atop research and scholarship. The purpose of these mechanisms is to ensure that the works cannot be put into general circulation. The problem is that TPMs make it impossible to use such papers with any custom software without breaking the law. Indeed, while it is technically trivial to circumvent some of these systems, there are also hefty criminal penalties for so doing. In the EU, this is specified by EU Directive 2001/29/EC and in the US by the Digital Millennium Copyright Act (DMCA). As an example of a nation-specific implementation of these legal frameworks, the UK has Section S296ZE of the Copyright, Designs and Patents Act. This section allows a researcher to appeal a rightsholder's TPMs where the use is noncommercial research. This involves asking a publisher to voluntarily provide a copy that can be used in such a way and, if they will not, then contacting the Secretary of State to ask for a directive to yield a way of benefiting from the copyright exemption for noncommercial academic research purposes.[12] As of 2014, there had been no successful challenges under this legislation.[13]

Machine Learning and Research Literature Classification

On top of the above, a further promising area that has yet to be explored is whether machine learning approaches could provide a future way by which to bring relevant research and scholarly literature to the attention of researchers. As with their biological counterparts, artificial neural networks consist of groups of interrelated processing units, called neurons, that connect together in order to solve problems. For instance, character-based recurrent neural networks are particularly good at generating sentences and words on a probabilistic basis, once trained on a suitable reference corpus.[14]

One of the tasks for which such software systems—and other forms of machine learning—are well-suited is classificatory problems. Given a known

corpus subdivided into groups of desirability, accuracy, or general interest (from "not interested," through to "highly relevant"), one could easily envisage a system that could provide an appraisal on behalf of researchers when fed a new paper or book. One could also imagine the classification of works based on their intersecting bibliographies ("show me works that sit at the center of the citation networks of all these other works"), methodological principles, or any other taxonomographic feature by which scholarship could be clustered.

There are, of course, challenges with such a method. Artificial neural networks tend to replicate existing structures of value. This has even led, in fields of natural language processing, to racist and sexist networks because, unfortunately, these are structural phenomena of our societies at large.[15]

If using machine learning to classify scholarship for personal reading preference, then, the danger is that we simply replicate a list of the works that a scholar would have read anyway; a filter bubble. Instead, we need ways to inject the *unexpected* and *fortuitous* into such systems so that we can still have the experience of chance advancing thought and research, without affecting the classificatory measures too adversely. (Although it is also worth noting that what researchers call serendipity is often actually the result of library classification procedures that bring works into parataxis.) On the other hand, such a system would bring with it the long-sought-after promise of relevant material for reading, reducing the burdening effects of abundance upon the contemporary researcher.

Tempered Possibilities

Such futurological technologies as those upon which I have here speculated are not far off in technical terms; these are no impossible science fiction or utopian dreams, at least in one sense. However, in social and legal terms, we remain some way from such visions. For the ability of these technologies to reach fruition at a viable scale depends upon *access to research works*. There are several routes by which this could become possible. Each of these ways is equally difficult to achieve but some are more desirable than others:

- Total centralization of all research article publication under a large corporate entity. This would allow that corporate entity to develop such systems as those to which I have here gestured. It would also, though, be hugely monopolistic and commercially dangerous.

- A compact between academic publishers to deposit all of their works in centralized repositories upon which mining operations can be performed.
- Total open access to the research literature.

Clearly, despite the promise of amplifying our labor time by reading scholarship with computers, we still have some way to go.

Notes

1. Alison McCook, "PLOS ONE Has Faced a Decline in Submissions—Why? New Editor Speaks," Retraction Watch, March 15, 2017, http://retractionwatch.com/2017/03/15/plos-one-faced-decline-submissions-new-editor-speaks/.

2. For more on this, see Martin Paul Eve, "Scarcity and Abundance," in *The Bloomsbury Handbook of Electronic Literature* (London: Bloomsbury Academic, 2017).

3. Björn Brembs, Katherine Button, and Marcus Munafò, "Deep Impact: Unintended Consequences of Journal Rank," *Frontiers in Human Neuroscience* 7 (2013): 291, https://doi.org/10.3389/fnhum.2013.00291.

4. For a selection of sources on these subjects, see Association of Research Libraries, "ARL Statistics 2009–2011"; George Monbiot, "Academic Publishers Make Murdoch Look like a Socialist," *The Guardian*, August 29, 2011, sec. Comment is free, http://www.guardian.co.uk/commentisfree/2011/aug/29/academic-publishers-murdoch-socialist; Eve, *Open Access and the Humanities: Contexts, Controversies and the Future*, chap. 2; Larivière, Haustein, and Mongeon, "The Oligopoly of Academic Publishers in the Digital Era."

5. "San Francisco Declaration on Research Assessment: Putting Science into the Assessment of Research."

6. For just a selection of such work, see Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London: Verso, 2007); Franco Moretti, "The Slaughterhouse of Literature," *MLQ: Modern Language Quarterly* 61, no. 1 (2000): 207–227; Franco Moretti, *Distant Reading* (London: Verso, 2013); Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013); Ted Underwood, "A Genealogy of Distant Reading," *Digital Humanities Quarterly* 11, no. 2 (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html; Andrew Piper, *Enumerations: Data and Literary Study* (Chicago: University of Chicago Press, 2018); Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: University of Chicago Press, 2019); Martin Paul Eve, *Close Reading With Computers: Textual Scholarship, Computational Formalism, and David Mitchell's* Cloud Atlas (Stanford, CA: Stanford University Press, 2019).

7. JSTOR Labs, "Text Analyzer Beta," 2017, https://www.jstor.org/analyze; Laura Brown et al., "Reimagining the Digital Monograph: Design Thinking to Build New

Tools for Researchers" (JSTOR Labs, 2017), https://hcommons.org/deposits/item/hc:14411/.

8. Note that Murray-Rust uses the term "content mining" instead of the legal terms "text and data mining," because he believes that it has broader connotations for where we might find useful information among multimedia forms, even if these are all, already, technically "data." See Peter Murray-Rust, "What Is TextAndData/ContentMining?," *Petermr's Blog*, July 11, 2017, https://blogs.ch.cam.ac.uk/pmr/2017/07/11/what-is-textanddatacontentmining/.

9. Tom Arrow, Jenny Molloy, and Peter Murray-Rust, "A Day in the Life of a Content Miner and Team," *Insights: The UKSG Journal* 29, no. 2 (2016): 208–211, https://doi.org/10.1629/uksg.310.

10. These bullet points are all taken from Murray-Rust, "What Is TextAndData/ContentMining?" sometimes with the same wording or example questions.

11. Peter Murray-Rust, "Sci-Hub and Legal Aspects of ContentMining 4/n," *Petermr's Blog*, May 6, 2016, https://blogs.ch.cam.ac.uk/pmr/2016/05/06/sci-hub-and-legal-aspects-of-contentmining/.

12. I write more on this in Martin Paul Eve, "Close Reading with Computers: Genre Signals, Parts of Speech, and David Mitchell's *Cloud Atlas*," *SubStance* 46, no. 3 (2017): 76–104.

13. Government of the United Kingdom, "Complaints to Secretary of State under s.296ZE under the Copyright, Designs and Patents Act 1988," August 15, 2014, https://www.gov.uk/government/publications/complaints-to-secretary-of-state-under-s296ze-under-the-copyright-designs-and-patents-act-1988.

14. For more, see Martin Paul Eve, "The Great Automatic Grammatizator: Writing, Labour, Computers," *Critical Quarterly* 59, no. 3 (2017): 39–54.

15. Tolga Bolukbasi et al., "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," *arXiv:1607.06520*, July 21, 2016, http://arxiv.org/abs/1607.06520; Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science* 356, no. 6334 (2017): 183–186, https://doi.org/10.1126/science.aal4230; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: New York University Press, 2018).