COMMENT

A. JOHANSSON/SHUTTERSTOCK.COM

Don't let copyright block data mining

Matthew L. Jockers, **Matthew Sag** and **Jason Schultz** explain why humanities scholars have pitched in to the *Authors Guild* v. *Google* lawsuit.

Advances in computer technology combined with the availability of digital archives are allowing humanities scholars to do what biologists, physicists and economists have been doing for decades — analyse massive amounts of data. A far richer understanding of literature promises to emerge. For instance, large-scale quantitative projects are forcing scholars to reconsider how literary canons are formed and are showing the extent to which authors' works are shaped by factors outside their own creative control, such as the period in which they lived, their gender and their nationality.

Yet in the United States, legal action pursued by the Authors Guild, an advocacy group for writers, could bar scholars from studying as much as two-thirds of the literary record. A small group of humanities scholars (ourselves included) is fighting back.

CASE HISTORY

In 2004, Google began scanning and digitizing books held in prominent US academic libraries such as those at Stanford University in California and the University of Michigan in Ann Arbor, to make these collections fully searchable. Currently, more than 20 million books, most of which are out of print, can be searched at Google Books (books.google.com). Unless a book's copyright protection has expired, or the copyright owner has agreed to make the content freely available, the search engine displays just three-line 'snippets' from each book — enough to tell the searcher that the listed item is indeed what they are looking for. With the right tools, however, data from the full text can, in principle, be mined and used in large-scale analyses.

In 2005, the Authors Guild, based in New York, with some 8,500 members including published authors, literary agents and lawyers, filed a class-action lawsuit claiming that Google's scanning activity was a "massive copyright infringement". Google, the Authors Guild and a group of publishers agreed to a class-action settlement in 2008. This gave Google permission to continue scanning and to sell electronic books individually or as part of a subscription service. In return, ▶

Google agreed to share the advertising revenue from Google Books with authors and publishers, and to make one-off payments to copyright owners amounting to a minimum of US$125 million.

The settlement was strongly opposed by foreign governments, the US Department of Justice, the US Copyright Office, authors, academics and rival technology companies for various reasons. Many feared that it would create an unfair monopoly, with Google having the sole right to publish millions of 'orphan' works — books whose copyright owners cannot easily be located. In 2009, the settlement was revised to try to address these concerns. But the court rejected the revised settlement in 2011, and the legal controversy continues.

In September last year, in a separate case, the Authors Guild sued several universities for participating in Google's book-scanning project. As part of this case, known as *Authors Guild* v. *HathiTrust*, it is also pursing legal action against the HathiTrust Digital Library, a service that enables a large consortium of universities and research libraries to store, secure and search their digital collections using a shared infrastructure.

Among the issues at the heart of this dispute is what researchers in the emerging field of digital humanities will be allowed to analyse: only public-domain books (mostly those published before 1923 in the United States), or all known literary works. The answer may define the future of the field.

TO THE BARRICADES

On 3 August, the Association for Computers and the Humanities and a group of 64 scholars (that includes us), from disciplines ranging from law and computer science to linguistics, history and literature, filed an *amicus curiae* brief on behalf of the digital humanities. We are urging the court in *Authors Guild* v. *Google* to grant a summary judgment in favour of Google, a step that will effectively end the litigation[1]. We filed a similar brief in the HathiTrust case on 7 July. The judge in the HathiTrust case is currently considering our submission, and a decision is expected imminently. The court in *Authors Guild* v. *Google* will consider our argument as soon as the appeals court deals with certain procedural issues.

We feel that if the Authors Guild wins the cases against Google and the HathiTrust, the ruling could set a dangerous precedent — that copyright gives authors and publishers the right to control all, even 'non-expressive' uses of their works that involve copying. Copyright law has long recognized the distinction between protecting an author's original expression and the public's right to access the facts and ideas contained within that expression. According to the US Constitution, the purpose of copyright is "To promote the Progress of Science and useful Arts". Preventing authors from monopolizing facts and ideas allows others to explore their own creativity and 'stand on the shoulders of giants'.

We believe that copyright law is not (and should not be) an obstacle to statistical and computational analysis of the millions of books owned by university libraries. We are not talking about republishing them or even quoting from them. We simply want to extract information from and about them to sift out trends and patterns.

As an example, clustering more than 3,000 nineteenth-century novels according to how much they share certain stylistic properties (specific words and punctuation marks) and thematic features (such as groups of commonly co-occurring words) has thrown up findings that would be hard to glean from reading a handful of books individually. One is that books authored by men tend to cluster quite distinctly from books authored by women (see 'Knowing your subject'). This illustrates the degree to which gender determines the choices made by writers, but also flags up outliers. For instance, within this clustering, the works of George Eliot (real name Mary Anne Evans) sit firmly among those of male writers. In other words, such 'macroanalytic' methodology gives researchers a way to see individual authors and publications within the context of a much larger system.

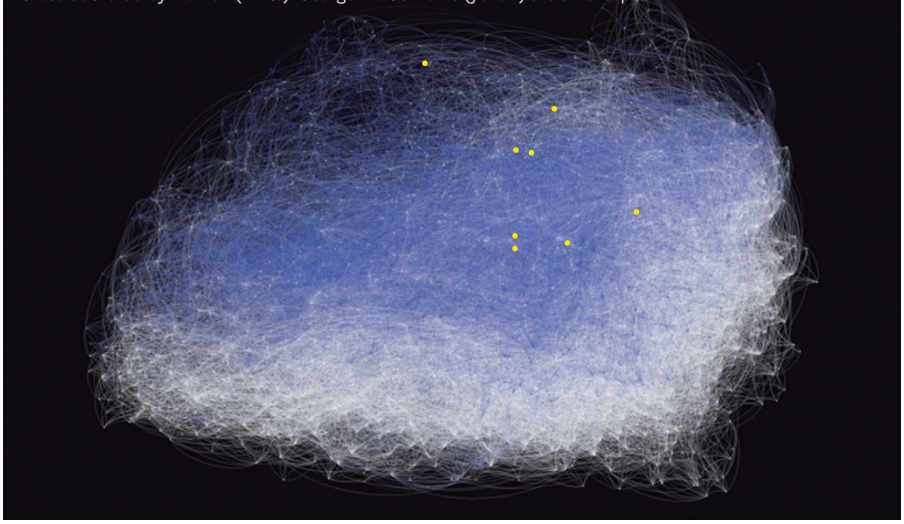
Authors' rights deserve protection. And governments and the various stakeholders involved may eventually work out how to achieve the full potential of digital libraries in a way that is fair to writers, readers and providers. But digitizing books for 'non-expressive' uses, such as basic searching and text mining, is a separate issue and should not be barred on the basis of concerns over copyright. An independent review last year of intellectual property and growth commissioned by the British government came to a similar conclusion[2]. Unauthorized music-file sharing can infringe copyright because humans ultimately experience those files as musical works. Scanning words from library books to make a search index, or to compile a list of word frequencies, does not interfere with the rights of the author. These uses simply convert masses of text into metadata.

It is time for the US courts to recognize explicitly that, in the digital age, copying books for non-expressive purposes is not infringement. Courts have already applied this logic in analogous cases: Google, Microsoft and others copy web pages to feed into their Internet search engines; the online service Turnitin copies exam papers and other sources so that plagiarism can be detected. These practices have been challenged and found to be legal under copyright law.

It is crucial for future research that the right precedent be set. We hope that the judges decide that digitization for text mining and other forms of computational analysis is, unequivocally, fair use. ■

KNOWING YOUR SUBJECT

In a network of more than 3,000 nineteenth-century novels, arranged according to how much they share certain stylistic and thematic properties, books authored by men (blue) tend to cluster separately from those authored by women (white). George Eliot's works (yellow) are an exception.



SOURCE: M. L. JOCKERS

Matthew L. Jockers *is assistant professor of English at the University of Nebraska, Lincoln, USA.* **Matthew Sag** *is associate professor of law at Loyola University, Chicago, Illinois, USA.* **Jason Schultz** *is assistant clinical professor of law at the University of California, Berkeley, USA.*
e-mail: mjockers@unl.edu

1. Jockers, M. L., Sag, M. & Schultz, J. preprint at Social Science Research Network (2012); available at http://ssrn.com/abstract=2102542.
2. Hargreaves, I. *Digital Opportunity: A Review of Intellectual Property and Growth* (Intellectual Property Office, 2011).