Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?

sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

By Jeffrey Brainard, May. 13, 2020 , 12:15 PM

13 mai 2020

SARA GIRONI CARNEVALE

***Science*'s COVID-19 reporting is supported by the Pulitzer Center.**

Timothy Sheahan, a virologist studying COVID-19, wishes he could keep pace with the growing torrent of new scientific papers about the disease and the novel coronavirus that causes it. But there are just too many—more than 4000 alone last week. "I'm not keeping up," says Sheahan, who works at the University of North Carolina, Chapel Hill. "It's impossible."

A loose-knit army of data scientists, software developers, and journal publishers is pressing hard to change that. Backed by large technology firms and the White House, they are racing to create digital collections holding thousands of freely available papers that could be useful to ending the pandemic, and scrambling to build data-mining and search tools that can help researchers quickly find the information they seek. And the urgency is growing: By one estimate, the COVID-19 literature published since January has reached more than 23,000 papers and is doubling every 20 days—among the biggest explosions of scientific literature ever.

Given that volume, "People don't have time to read through entire articles and figure out what is the value added and the bottom line, and what are the limitations," says Kate Grabowski, an infectious disease epidemiologist at Johns Hopkins University's (JHU's)

Bloomberg School of Public Health who is leading an effort to create a curated set of pandemic papers.

It's not clear, however, just how much traction the new efforts, many of them just weeks old, are gaining. A global effort to persuade publishers to make all papers relevant to COVID-19 immediately free and available to all, for example, has hit some obstacles; as many as 20% of new papers are still behind paywalls, and that share could grow to 50%, a recent study found. Some of the new search tools, meanwhile, are little-known outside of the research groups that created them. Sheahan, for example, hadn't heard of several literature-mining algorithms that have been rolled out recently. Other tools have interfaces that aren't particularly user friendly. And many researchers are skeptical that the tools can tell them what they really want to know: What is the quality of the work? "People tend to oversell and put up papers with data that do not support their conclusions," Sheahan says. "It's a mess."

Hundreds of teams are trying to help clean things up by pursuing one of at least two basic strategies: creating easily accessible paper collections, including a few carefully curated collections designed to highlight strong papers; and building automated search tools that use artificial intelligence (AI) technologies to cut through the noise.

A massive literature trove

On 16 March, efforts to create COVID-19 literature troves got a lift from the White House Office of Science and Technology Policy, which worked with publishers and tech firms to launch the CORD-19 data set, considered the largest single collection to date. It holds more than 59,000 published articles and preprints, including studies of coronaviruses dating back to the 1950s.

To create the archive, some of the largest groups active in machine learning got to work. Google, the Chan Zuckerberg Initiative, and the Allen Institute for AI collaborated with the National Institutes of Health and other groups to identify and collect the papers using methods that included natural language processing, which looks beyond the coded keywords in documents for variants of search terms and for related text. Participants also converted PDF files into a form readable by machine learning algorithms. The creators intend CORD-19 to help researchers not only search for relevant literature, but also extract meaningful patterns from findings across papers.

Giovanni Colavizza, a bibliometrics researcher at the University of Amsterdam, calls the creation of CORD-19 "amazing work." But analyses he has conducted with colleagues have found shortcomings. For example, more than 60% of the papers in CORD-19 don't mention the search terms used by the collection's creators—such as "coronavirus" and "SARS-CoV," the virus that causes severe acute respiratory syndrome—in their titles, abstracts, or keywords, the researchers reported in a 17 April preprint study posted on bioRxiv. That means these articles might be related only tangentially to COVID-19, he says.

What's more, the team found only about 40,000 papers in the collection had full text, necessary for comprehensive data mining.

A related issue is that not all pandemic papers are freely available. In response to calls from major science funders and government science advisers, most major publishers have pledged to make all COVID-19–related papers free. But a recent study estimated that about 20% of pandemic studies published this year are still behind paywalls. And the number of paywalled publications is growing faster than the free ones, according to the study, which was led by Nicolas Robinson-Garcia of the Delft University of Technology and posted as a preprint on 26 April on bioRxiv. By 1 June, nearly half of all COVID-19 papers could be behind paywalls if current trends continue, the researchers estimate, potentially limiting the data-mining effort and access by some scientists.

Quality, not quantity

At JHU, Grabowski's team is taking a different approach to creating a useful set of COVID-19 papers, focusing on quality over quantity. To create its curated 2019 Novel Coronavirus Research Compendium, which debuted on 17 April, 40 scientists have combed through the literature and selected more than 80 papers on eight topics, including vaccines and pharmaceutical interventions, that they thought were above the bar. They then wrote short summaries of each.

The effort is focused on studies in humans, Grabowski says, and the intended readers are primarily health care workers and policymakers, as well as researchers. "We are trying to fill a void that we saw existed because there is just so much information, but a lot of the studies are not conducted very well," she says. The team has excluded most of the articles it considered for the compendium because they contained only commentaries, protocols, poor-quality modeling studies, or no original findings, Grabowski adds.

Some of the concern about quality arises because many researchers have posted preprints—which are not peer reviewed—in a bid to get their findings out quickly. But contrary to some perceptions, manuscripts appearing only as preprints have made up only a minority of the pandemic literature gusher, according to work done by Robinson-Garcia's team. As of 14 April, some 80% of the more than 11,000 COVID-19 manuscripts it examined had appeared in refereed journals, some of which originally appeared as preprints.

In part, that number reflects efforts by publishers to accelerate peer-review and publication schedules. Since the pandemic began, for example, 14 medical journals publishing the most COVID-19 content have halved the average time from submission to publication to about 60 days, according to research by Serge Horbach of Radboud University. "Some concerns remain about whether faster dissemination might go at the expense of research quality," he writes in a preprint posted 18 April on bioRxiv.

It's still too soon to measure the quality of papers published during the pandemic rush based on citations or retractions, specialists say. But Robinson-Garcia's group found the papers are having an off-the-charts impact when evaluated by a different measure: mentions on social media. COVID-19 papers published this year are getting 10 times as many mentions per article as all scientific publications during the first 5 months of 2019, according to the Altmetric.com database, which tracks Twitter, Facebook, and other sources and displays a composite score for each paper. What's more, the 12 highest scoring scientific articles ever are about COVID-19. (Mentions by scientists aren't reported separately, but they have actively taken to Twitter to challenge individual studies they consider suspect, an ad hoc form of quality control.)

A tool-building rush

To tame the flood of papers, many teams are turning to advanced computational tools. The White House, for example, has asked data scientists to develop tools to analyze the CORD-19 data set, in a bid to help researchers answer 10 high-priority, pandemic-related research questions identified by the U.S. National Academy of Sciences and the World Health Organization. More than 1500 projects are listed on Kaggle, an online hub for machine learning scientists that is owned by Google Cloud.

Among the early fruits of the data mining work is an "AI-powered literature review." Using algorithms, researchers harvested data points of interest from a subset of 783 papers in CORD-19 grouped in 17 categories, then created a web page for each topic that displays the results. For example, one page shows data from studies about heart disease as a risk factor for death from COVID-19. Users can scan a table showing the risk reported by each paper as an odds ratio—and can click through to each paper's text to learn more, says Tayab Waseem, a Ph.D. immunologist attending Eastern Virginia Medical School who has helped lead the project.

The work is far from fully automated. Algorithms don't always correctly extract the relevant data point for these tables, so medical students and other volunteers idled by the pandemic have been checking each against the manuscripts for accuracy, Waseem says. The tool has attracted about 122,000 page views since it debuted on 10 April, he says.

Another challenge is making the tools more user friendly. Although data scientists have spent more than 20 years building tools to mine other topics in scientific literature, they have lagged in fine-tuning ways to help users explore the content of research articles, says Karin Verspoor, a computational linguist at the University of Melbourne. At the same time, "People on the user side haven't quite realized that they need [these tools], until now," she says. And that could promote greater attention to building helpful interfaces for COVID-19 and, eventually, other research topics.

Jevin West, a data scientist at the University of Washington, Seattle, has worked with colleagues to develop a user-focused tool called SciSight to mine the CORD-19 data set. Unveiled last week, it automatically provides suggestions of papers containing related themes to help refine search results. It also displays connections between papers as browseable maps.

Word of such tools has yet to reach many scientists. A half-dozen contacted by *Science*Insider, for example, said they sounded promising but hadn't heard of them. And some added that they didn't have time to try them.

That reticence highlights yet another challenge: getting researchers to deviate from their usual ways of sifting through the literature. "Even if you have that perfect tool, it's hard to change [scientists'] way of information foraging and searching within a pandemic," West says. "It's like going into an emergency room and giving the doctors a different scalpel and saying: 'This is actually better.' It's going to take some time to get people to change their habits."

In the meantime, many researchers say they are falling back on some time-tested ways of identifying key COVID-19 papers, including reading bulletins from scientific societies and a few leading journals, as well as relying on word of mouth—including tweets—from trusted colleagues.

"You do what you can. … There are more fact sheets and webinars than most people can digest," says Sherry Chou, a neurologist at the University of Pittsburgh Medical Center who organized an international research consortium studying neurologic complications of COVID-19. "When I'm not taking care of patients, I see how much more I can absorb."

But the quantity of new information is daunting, she says. "It's like what you would get in a medical conference that used to happen yearly. Now that's happening daily."

***Correction, 15 May, 10 a.m.:** *This story has been updated to remove an incorrect statement that the CORD-19 data set holds exclusively English-language papers.*