SMITA SHARMA FOR *NATURE*

Carl Malamud in front of the data store of 73 million articles that he plans to let scientists text mine.

THE PLAN TO MINE THE WORLD'S RESEARCH PAPERS

A data store in India could open up vast swathes of science for easy computerized analysis.

BY PRIYANKA PULLA

Carl Malamud is on a crusade to liberate information locked up behind paywalls — and his campaigns have scored many victories. He has spent decades publishing copyrighted legal documents, from building codes to court records, and then arguing that such texts represent public-domain law that ought to be available to any citizen online. Sometimes, he has won those arguments in court. Now, the 60-year-old American technologist is turning his sights on a new objective: freeing paywalled scientific literature. And he thinks he has a legal way to do it.

Over the past year, Malamud has — without asking publishers — teamed up with Indian researchers to build a gigantic store of text and images extracted from 73 million journal articles dating from 1847 up to the present day. The cache, which is still being created, will be kept on a 576-terabyte storage facility at Jawaharlal Nehru University (JNU) in New Delhi. "This is not every journal article ever written, but it's a lot," Malamud says. It's comparable to the size of the core collection in the Web of Science database, for instance. Malamud and his JNU collaborator, bioinformatician Andrew Lynn, call their facility the JNU data depot.

No one will be allowed to read or download work from the repository, because that would breach publishers' copyright. Instead, Malamud envisages, researchers could crawl over its text and data with computer software, scanning through the world's scientific literature to pull out insights without actually reading the text.

The unprecedented project is generating much excitement because it could, for the first time, open up vast swathes of the paywalled literature for easy computerized analysis. Dozens of research groups already mine papers to build databases of genes and chemicals, map associations between proteins and diseases, and generate useful scientific hypotheses. But publishers control — and often limit — the speed and scope of such projects, which typically confine themselves to abstracts, not full text. Researchers in India, the United States and the United Kingdom are already making plans to use the JNU store instead. Malamud and Lynn have held workshops at Indian government laboratories and universities to explain the idea. "We bring in professors and explain what we are doing. They get all excited and they say, 'Oh gosh, this is wonderful'," says Malamud.

But the depot's legal status isn't yet clear. Malamud, who contacted several intellectual-property (IP) lawyers before starting work on the depot, hopes to avoid a lawsuit. "Our position is that what we are doing is perfectly legal," he says. For the moment, he is proceeding with caution: the JNU data depot is air-gapped, meaning that no one can access it from the Internet.

Users have to physically visit the facility, and only researchers who want to mine for non-commercial

purposes are currently allowed in. Malamud says his team does plan to allow remote access in the future. "The hope is to do this slowly and deliberately. We are not throwing this open right away," he says.

THE POWER OF DATA MINING

The JNU data store could sweep aside barriers that still deter scientists from using software to analyse research, says Max Häussler, a bioinformatics researcher at the University of California, Santa Cruz (UCSC). "Text mining of academic papers is close to impossible right now," he says — even for someone like him who already has institutional access to paywalled articles.

Since 2009, Häussler and his colleagues have been building the online UCSC Genome Browser, which links DNA sequences in the human genome to parts of research papers that mention the same sequences. To do that, the researchers have contacted more than 40 publishers to ask permission to use software to rifle through research to find mentions of DNA. But 15 publishers have not responded or have denied permission. Häussler is unsure whether he can legally mine papers without permission, so he isn't trying. In the past, he has found his access blocked by publishers who have spotted his software crawling over their sites. "I spend 90% of my time just contacting publishers or writing software to download papers," says Häussler.

Some countries have changed their laws to affirm that researchers on non-commercial projects don't need a copyright-holder's permission to mine whatever they can legally access. The United Kingdom passed such a law in 2014, and the European Union voted through a similar provision this year. That doesn't help academics in poor nations who don't have legal access to papers. And even in the United Kingdom, publishers can legally place 'reasonable' restrictions on the process, such as channelling scientists through publisher-specific interfaces and limiting the speed of electronic searching or bulk downloading to protect servers from overload. Such limits are a big problem, says John McNaught, deputy director of the National Centre for Text Mining at the University of Manchester, UK. "A limit of, say, one article every five seconds, which sounds fast for a human, is painfully slow for a machine. It would take a year to download around six million articles, and five years to download all published articles concerning just biomedicine," he says.

Wealthy pharmaceutical firms often pay extra to negotiate special text-mining access because their work has a commercial purpose, says McNaught. In some cases, publishers allow these firms to download papers in bulk, thus avoiding rate limits, according to a researcher at a pharmaceutical firm who did not want to be identified because they were not authorized to talk to the media. University academics, however, frequently restrict themselves to mining article abstracts from databases such as PubMed. That provides some information, but full texts are much more useful. In 2018, a team led by computational biologist Søren Brunak at the Technical University of Denmark in Lyngby showed that full-text searches throw up many more gene–disease links than do searches of abstracts (D. Westergaard *et al. PLoS Comput. Biol.* **14**, e1005962; 2018).

Scientists must also overcome technical barriers when mining articles. It is hard to extract text from the various layouts that publishers use — something that the JNU team is struggling with right now. Tools to convert PDFs to plain text don't always distinguish clearly between paragraphs, footnotes and images, for instance. Once the JNU team has done it, however, others will be saved the effort. The team is close to completing the first round of extraction from the corpus of 73 million papers, Malamud says — although they will need to check for errors, so he expects the database won't be ready until the end of the year.

> **"OUR POSITION IS THAT WHAT WE ARE DOING IS PERFECTLY LEGAL."**

A WORLD OF POSSIBILITIES

Early enthusiasts are already gearing up to use the JNU depot. One is Gitanjali Yadav, a computational biologist at Delhi's National Institute of Plant Genome Research (NIPGR) and a lecturer at the University of Cambridge, UK. In 2006, Yadav led an effort at NIPGR to build a database of chemicals secreted by plants. Called EssOilDB, this database is today scoured by groups from drug developers to perfumeries looking for leads. Yadav thinks that "Carl's compendium", as she calls it, could give her database a leg-up.

To make EssOilDB, Yadav's team had to trawl PubMed and Google Scholar for relevant papers, extract data from full texts where they could, and manually visit libraries to copy out tables from rare journals for the rest. The depot could fast-forward this work, says Yadav, whose team is currently writing the queries they will use to extract the data.

Srinivasan Ramachandran, a bioinformatics researcher at Delhi's Institute of Genomics and Integrative Biology, is also excited by Malamud's plan. His team runs a database of genes linked to type 2 diabetes; they've been crawling PubMed abstracts to find papers. Now he hopes the depot could widen his mining net.

And at the Massachusetts Institute of Technology (MIT) in Cambridge, a team called the Knowledge Futures Group says it wants to mine the depot to map how academic publishing has evolved over time. The group hopes to forecast emerging areas of research and identify alternatives to conventional metrics for measuring research impact, says team member James Weis, a doctoral student at MIT Media Lab.

A CAREER UNLOCKING COPYRIGHT

Malamud only recently had the idea of extending his activism to academic publishing. The founder of a non-profit corporation called Public Resource, based in Sebastopol, California, Malamud has focused on buying up government-owned legal works and publishing them. These include, for instance, the state of Georgia's annotated legal code, European toy-safety standards and more than 19,000 Indian standards for everything from buildings and pesticides to surgical equipment.

Because these documents are often a source of revenue for government agencies, some of them have sued Malamud, who has argued back that documents which have the force of the law cannot be locked behind copyright. In the Georgia case, a US appeals court cleared him of infringement charges in 2018, but the state appealed, and the case is with the US Supreme Court. Meanwhile, a German court ruled in 2017 that the publication of toy standards by Public Resource, including a standard on baby dummies (pacifiers), was illegal.

But Malamud has enjoyed victories, too. In 2013, he filed a lawsuit in a US federal court asking the Internal Revenue Service (IRS) to publish the forms it collected from tax-exempt non-profit organizations — data that could help to hold these organizations to account. Here, the court ruled in Malamud's favour, prompting the IRS to release the financial information of thousands of non-profit organizations in a machine-readable format.

In early 2017, aided by the Arcadia Fund, a London-based charity that promotes open access, Malamud turned his attention to research articles. Under US law, works by US federal government employees cannot be copyrighted, and Public Resource says it has found hundreds of thousands of academic articles that are US government works and seem to defy this rule. Malamud has called for such articles to be freed from copyright assertions, but it's not clear whether that would hold up in court. He has posted his preliminary results online, but has put further campaigning on hold, because the project prompted him to take on a wider mission: democratizing access to all scientific literature.

OPPORTUNITY IN INDIA

A trigger for this mission came from a landmark Delhi High Court judgment in 2016. The case revolved around Rameshwari Photocopy Services, a shop on the campus of the University of Delhi. For years, the business had been preparing course packs for students by photocopying pages from expensive

textbooks. With prices ranging between 500 and 19,000 rupees (US$7–277), these textbooks were out of reach for many students.

In 2012, Oxford University Press, Cambridge University Press and Taylor and Francis filed a lawsuit against the university, demanding that it buy a license to reproduce a portion of each text. But the Delhi High Court dismissed the suit. In its judgment, the court cited section 52 of India's 1957 Copyright Act, which allows the reproduction of copyrighted works for education. Another provision in the same section allows reproduction for research purposes.

Malamud has a long association with India: he first travelled there as a tourist in the 1980s, and he wrote one of his first books, on database design, on a houseboat in Srinagar. And around the same time that he heard about the Rameshwari judgment, he had come into possession (he won't say how) of eight hard drives containing millions of journal articles from Sci-Hub, the pirate website that distributes paywalled papers for anyone to read. Sci-Hub itself has lost two lawsuits against publishers in US courts over its copyright infringements, but despite those judgments, some of its domains are still working today.

Malamud began to wonder whether he could legally use the Sci-Hub drives to benefit Indian students. In a 2018 book about his work called *Code Swaraj*, co-authored with Indian tech entrepreneur Sam Pitroda, Malamud writes that he imagined showing up on Indian campuses in the equivalent of an American taco truck, ready to serve the articles up to those who wanted them.

Ultimately, he zeroed in on the idea of the JNU text-mining depot instead. (Malamud has also helped to set up another mining facility with 250 terabytes of data at the Indian Institute of Technology Delhi, which isn't in use yet.) But he is cagey about where the depot's articles come from. Asked directly whether some of the text-mining depot's articles come from Sci-Hub, he said he wouldn't comment, and named only sources that provide free-to-download versions of papers (such as PubMed Central and the 'Unpaywall' tool). But he does say that he does not have contracts with publishers to access the journals in the depot.

IS IT LEGAL?

Malamud says that where he got the articles from shouldn't matter anyway. The data mining, he says, is non-consumptive: a technical term meaning that researchers don't read or display large portions of the works they are analysing. "You cannot punch in a DOI [article identifier] and pull out the article," he says. Malamud argues that it is legally permissible to do such mining on copyrighted content in countries such as the United States. In 2015, for instance, a US court cleared Google Books of copyright infringement charges after it did something similar to the JNU depot: scanning thousands of copyrighted books without buying the rights to do so, and displaying snippets from these books as part of its search service, but not allowing them to be downloaded or read in their entirety by a human.

SAJJAD HUSSAIN/AFP/GETTY

Rameshwari Photocopy Services in New Delhi was taken to court for copying parts of textbooks, and won.

The Google Books case was a test of non-consumptive data mining, says Joseph Gratz, an IP lawyer at the law firm Durie Tangri in San Francisco, California, who represented Google in the case and has previously represented Public Resource. Even though Google was displaying snippets, the court ruled that the text was too limited to amount to infringement. Google was scanning authorized copies of books (from libraries in many cases), even though it did not ask permission. Copyright holders might argue that if Sci-Hub or other unauthorized sources supplied the JNU depot, the situation would be different from the Google Books case, Gratz says. But a case involving unauthorized sources has never been argued in American courts, making it hard to predict the outcome. "There are good reasons why the source shouldn't matter, but there may be arguments that it should," says Gratz.

The question of the facility's legality in the United States might not even be relevant, because international researchers would be getting results from a depot that sits in India, even if they are accessing it remotely. So Indian law is likely to apply to the question of whether it is legal to create the corpus, says Michael W. Carroll, a professor at the American University's Washington College of Law in Washington DC.

Here, India's copyright laws might help Malamud — another reason why the facility is in New Delhi. The research exemption in section 52 means that the JNU data depot's actions would be considered fair under Indian law, argues Arul George Scaria, an assistant professor at Delhi's National Law University. Not everyone agrees with this interpretation, however. Section 52 allows researchers to photocopy a journal article for personal use, but doesn't necessarily allow the blanket reproduction of journals as the JNU depot has done, says T. Prashant Reddy, a legal researcher at the Vidhi Centre for Legal Policy in New Delhi. That entire articles aren't shared with users does help, but the mass reproduction of text used to create the database puts the facility in "a legal grey zone", Reddy says.

RISKY BUSINESS

When *Nature* contacted 15 publishers about the JNU data depot, the six who responded said that this was the first time they had heard of the project, and that they couldn't comment on its legality without further information. But all six — Elsevier, BMJ, the American Chemical Society, Springer Nature, the American Association for the Advancement of Sciences and the US National Academy of Sciences — stated that researchers looking to mine their papers needed their authorization. (Springer Nature publishes this journal; *Nature*'s news team is editorially independent of its publisher.)

Malamud acknowledges that there is some risk in what he is doing. But he argues that it is "morally crucial" to do it, especially in India. Indian universities and government labs spend heavily on journal subscriptions, he says, and still don't have all the publications they need. Data released by Sci-Hub indicate that Indians are among the world's biggest users of their website, suggesting that university licences don't go far enough. Although open-access movements in Europe and the United States are valuable, India needs to lead the way in liberating access to scientific knowledge, Malamud says. "I don't think we can wait for Europe and the United States to solve that problem because the need is so pressing here." ■

Priyanka Pulla *is a freelance journalist based in Bengaluru, India.*

CLARIFICATION

The News Feature 'The plan to mine the world's research papers' (*Nature* **571**, 316–318; 2019) used the term 'fair use' inappropriately — the term isn't relevant under Indian law.