The Elephant in the Room
Prepared Remarks of Carl Malamud

National Frontiers of Science
Indian Young Academy of Sciences (INYAS)
November 6, 2019, Jaipur

Thank you Dr. Yadav and the Indian Young National Academy of Science for your kind invitation to be with you.

Honored guests. My friends. Good evening.

There is an elephant in the room with us, and I believe we cannot continue to pretend this is not so. We must confront the issues it has raised, and that is the topic I wish to address.

Knowledge has become colonized. This trend is true for knowledge in all fields, but it is particularly true for science. Your work has all-too-often become the private property of a few large corporate concerns, immensely rich for-profit trading companies that have claimed exclusive control over the corpus of scientific literature.

In these modern times when the Internet makes universal access to human knowledge possible, it is ironic that the scholarly literature has become less and less accessible.

Prices for journals have soared while the costs of production have plummeted. Physical copies have been replaced with digital files that are locked down with digital rights management, terms of use, obscure user interfaces, and aggressive and unfriendly gatekeepers who continually remind scientists like you to keep off the grass.

In this world we have come to inhabit, scientists have become the new Indigo farmers. You ship your preprints and other raw materials off to London where modern-day East India Companies like Reid Elsevier force you to buy back high-priced finish goods.

These merchant trading companies of knowledge—both for-profit and sadly even many purportedly not-for-profit scholarly societies—these companies are the elephant in the room. They have become a knowledge Raj.

As an author, you are told you may not share your own work without permission, and such permission often comes with a tax and severe limitations. You are told you may not make any use of the scholarly corpus without first applying for a license, and such licenses are often arbitrarily denied.

You may not even make copies of the so-called "version of record" of your own articles for your family or your students, at least according to the fear, uncertainty, and doubt the gatekeepers spread among you.

I have two problems with the idea that an East India Company of knowledge can tell you what—and how—you may read in the course of your pursuit of scientific knowledge.

The big problem of course is that the students and scholars of the world do not have access to the scholarly corpus in order to further their studies. I am pretty sure if I asked you how many of your students used Sci-Hub to do their research, almost every hand would go up.

That's a big problem. A huge problem. A moral travesty.

Education is a fundamental right in our society, the way that people from any walk of life who have the capability to learn may earn a better living, teach themselves a craft or an art, become a professional, or—like all of you here—to practice science to further the increase and diffusion of knowledge.

But, that is not the problem I wish to speak to you about. We are gathered here to discuss the frontiers of science. One of those frontiers is what you may know of as "big data" or "machine learning" or "text and data mining." We use computers to examine the work of those that came before us so that we may stand on the shoulders of giants and reach for new heights.

As scientists, text and data mining offers some amazing opportunities. To conduct this research, however, you must have access to the scholarly corpus. Sci-Hub has approximately 75 million journal articles, Crossref lists approximately 100 million objects with Digital Object Identifiers. Your research labs and universities have access to only a fraction of this body of knowledge.

This second problem—the use of text and data mining on the scholarly corpus for the purpose of scientific research—is a pressing issue for our times. If a scientist believes she can perhaps better cure cancer if she is able to search the text of previous research, then it would be immoral for private parties to tell that scientist she may not proceed upon this inquiry.

But, that is what happens today. Text and data mining is purportedly prohibited. By prohibited, I mean that there are people who make a lot of noise about why this can't possibly be allowed without unjustified monetization and licenses—without clearing the details of your inquiry with them before you may proceed.

☸

Let me give you the example of Max. Dr. Maximilian Haeussler is a researcher at the University of Santa Cruz Genomics Institute. He is using text and data mining to search for references to chromosomal locations in scientific articles, then makes those available in a genome browser. This genocoding software is 200 lines of Python code that searches texts for different ways to refer to a chromosomal location, such as gene symbols, SNP mutation identifiers, or cytogenetic band names.

Max put together a letter requesting permission to crawl all articles on a publisher's site that were published after 1980 (which is the advent of routine reporting of DNA sequences). The code only pulls out 200 character snippets around the match, it is clearly non-consumptive, by which I mean people are not reading or disseminating the article, they are using computers to extract a very small portion.

He sent the letter to 43 publishers. All 43 specifically prohibit crawling their site in the terms of use. For 28, he got some form of partial permission, but in many cases that permission was empty—no site license was forthcoming and technical measures have prevented crawling the site. Fifteen of the answers were an outright no or they simply ignored him.

He has been unable to complete this important work. He has been blocked because gatekeepers don't approve of his research.

A second example of text and data mining is your own Professor Gitanjali Yadav. She is doing a fascinating research project that is examining the silent language of plants. Plants communicate with each other and with other species using chemicals. Each plant has a chemical fingerprint, a unique bouquet of scents and emissions.

The text and data mining she is conducting consists of searching journal articles looking for the names of plant species and their parts, then extracting the names of any volatile compounds associated with those plants as well as details such as where they were reported and the date.

This work began 10 years ago when she searched open sources such as PubMed and created the Essential Oil Database. The database presently contains 1.2 lakh essential oil records with data from 92 plant taxonomic families. But, that is based on only a small set of articles and Dr. Yadav is convinced that she will be able to greatly increase this database with a search of the full scholarly corpus.

A third example was recently featured in Nature, this one in the area of materials science. The discovery of new materials is a mixture of craft and science. It is often a trial-and-error process, and is a very inefficient, almost artisanal process.

Using 3.3 million scientific abstracts, the researchers created a 500,000 word vocabulary, then looked at co-occurrences of words—such as "iron" or "steel"—and other terms, such as chemical compositions—using unsupervised machine learning. These word vectors were then associated with various materials, which were then clustered around major categories of uses, such as superconductors, battery materials, photovoltaics, and organic compounds.

This example shows the potential of data mining, but what if they had more than just abstracts to work with? Would the results be better?

A study in PLOS Computational Biology did text mining of protein-protein, disease-gene, and protein subcellular associations to examine that question.

This study compared the results from performing extraction on 15 million abstracts with the results from the same procedure on 15 million full text articles. As one would intuit, the results were far superior with full text. The reason is simple. Abstracts are highly summarized and the full text has much more detail.

Text and data mining is not just for the hard sciences. Legal informatics has used text and data mining to examine similarity in court opinions to see, for example, how U.S. District Court and Court of Appeals decisions influenced the U.S. Supreme Court.

Text and data mining is a key component of modern search engines, it is used for machine-assisted translation, it was even used recently to to determine what makes people happy!

The Economist reported on this recent study, which examined over 8 million books and millions of newspaper articles for terms with a psychological valence of happiness. Researchers found that as wealth increased people became happier, but that was incidental. Significantly more important for happiness was the health of the population and the absence of war.

Text and data mining can also be used for plagiarism detection. You may be familiar with Dr. Elisabeth Bik, the "lab fairy," who has been highlighting unattributed or manipulated reuse of imagery in articles. Image recognition on the full text of articles could substantially assist in this enterprise, as could techniques to mathematically model the words of all journal articles in order to do plagiarism detection on texts.

This clearly is an important frontier of science, but it is an unexplored frontier.

☸

We have set out to change that.

For the past 18 months, I have been collaborating with my colleague Dr. Andrew Lynn of the School of Computational and Integrative Sciences at Jawaharlal Nehru University. Many of you are familiar with Professor Lynn's distinguished career in applying informatics to the biological sciences.

We have built a system we call the JNU Data Depot. A replica of this system is now spinning at IIT Delhi under the direction of Dr. Sanjiva Prasad, a distinguished computer scientist.

The JNU Data Depot consists of 2 large systems, each with 24 disk drives, and a cluster of smaller towers. The computers are cut off from the broader Internet, what we call an air-gapped system. We are spinning a bit over 500 terabytes of data.

On that system are the texts of over 125 million journal articles culled from a variety of sources. There is overlap in many of the sources, but we believe we have approximately 75 million unique articles.

The text has been extracted from underlying PDF files using common utilities such as pdftotext and grobid, which builds on pdftotext to add XML-based structure to the extract. Images and other components are also extracted for analysis. A number of corollary data sets are also on the system, such as the Crossref database of citations and a number of important biological databases such as the Elixir data sets.

The system contains copyrighted data, so it is carefully secured from the rest of the Internet. You need to bring your computer to campus, apply to Dr. Lynn for permission, and must agree to the JNU Data Depot terms of use. Our terms of use are a direct copy from the Hathi Trust text and data mining facility in the U.S. and strictly prohibit any redistribution of the underlying texts.

The Hathi Trust system contains all the books scanned by Google Books. Their terms of use—and ours as well—establish that the system may only be used for non-consumptive text and data mining. You are not reading or even looking at the full text of the PDF files, you are using computers to mine the text for facts.

We are limiting the system to non-commercial uses only, and have placed a number of other limitations on the system. The system was described recently in an article in Nature magazine.

India is on the frontier of this revolution, one of the first places in the world where researchers may do this type of research on a collection that approaches the full scientific corpus. We are in the early days of the system, and are still bringing it up to speed, but the system exists and is in use by a number of researchers.

☸

You may rightly ask, is this legal?

We approached the JNU Data Depot in an exceedingly deliberate and careful manner. In particular, we discussed the issues extensively with a number of legal scholars, and have put on the record a legal analysis by Professor Arul Georgia Scaria, a leading intellectual property expert at National Law University, Delhi, and by Dr. Zakir Thomas, a senior member of the civil service who was formerly the Registrar of Copyright for the Government of India, who contributed his analysis in his personal capacity.

We also consulted with a number of other legal scholars, including Professor Feroz Ali at IIT Madras, Professor N.S. Gopalakrishnan of the Cochi Institute of Science and Technology—widely considered the dean of the intellectual property community in India—and Professor Lawrence Liang of the Ambedkar University School of Law.

I also discussed the matter extensively with Professor Shamnad Basheer, who I am sad to say recently passed away at an all-too-young age, after a brilliant career of public work and public advocacy. He is sorely missed.

I am not a lawyer, but I am familiar with many of the issues because of my work making knowledge available in the U.S., Europe, and India. I am always very careful before I make a database available, and believe it is crucial that one have a strong legal grounding for any action that is taken. I strongly believe that is the case here, and the legal experts agree.

In the U.S., the question of text and data mining was extensively litigated when Google started scanning all the books of the world, and did so without first asking for permission. Google was not distributing the books, they scanned them for the purpose of text and data mining—for showing snippets to users, for driving search engine results.

Hathi Trust is the consortium of universities that provided Google with the books to scan, and received a digital copy back in return. They made the books available to their members, which are the major universities of the United States, to show their users previews and snippets, and in many cases the full text.

Hathi Trust went even further, and created a text and data mining facility where researchers may—using the same model we are using at the JNU Data Depot—mine the full text of all the books that Google scanned.

The courts have ruled repeatedly—and definitively—that the uses of both Google and of Hathi Trust are legal.

But, is this legal in India you may ask? And, are journal articles the same as books? The answer to both questions is yes.

India is a common law country. There are statutes, but ultimately any new use must be judged by the courts if somebody were to object. But we believe we stand on very solid ground and we do not expect to be challenged because of the careful and deliberate way we have gone about this, the wide-spread participation and support for this endeavor from major universities and government research labs throughout India, and the obvious and compelling need for this facility.

A basic concept in copyright law is that there is no copyright in facts and ideas, it only protects the expression of those ideas as fixed into a tangible medium. One cannot use end user license agreements to prohibit the use of facts and ideas to override the underlying purposes and rights under copyright law and the Constitution. That is morally wrong and legally wrong.

It is also important to understand that copyright law is about the rights of users as much as about the rights of creators. That was underscored in the opinion of the Hon'ble Justice Endlaw in the famous Delhi University Copyshop case, where he said that the rights of copyright users are not be read narrowly or strictly.

Under Indian law, there are a number of exceptions to copyright. What that means is that even if a work is under copyright, certain uses are allowed. The Delhi University Copyshop case was about one of those exceptions. A professor may assign a course pack of materials that are under copyright because one of the exceptions is when materials are furnished by a teacher to a student in the course of instruction.

Another exception to copyright is for private or personal uses, specifically including research uses. That is what is happening in the JNU Data Depot.

The exceptions to copyright are part of an analysis known as fair dealing. What is "fair" is a matter of degree, it depends on the circumstances and a number of factors, and it is a matter for the courts if a controversy should arise. We hope and believe our work will not reach the courts because it is clearly fair dealing, it is clearly vital to the future of scientific inquiry, and it is in line with the goals and objectives that have been laid out by the Government of India to continue to keep India at the leading edge of the frontiers of science.

In the JNU Data Depot, only small snippets are being extracted. The purpose of our dealing is strictly non-commercial. It is being done as non-consumptive use. Articles are not being distributed, they are not being made available to others, our use is not in any way harming the market for these works.

One of the most important factors in evaluating fair dealing is to examine the alternatives that are available. In the case of text and data mining, there is no possibility of licensing the entire scientific corpus. Yet, this corpus is what is known as an essential utility—one cannot conduct science properly without using this corpus, just as one could not operate a hospital without electricity. Even in the physical world, if a property may not be accessed without trespassing another's property, the law permits easement rights.

Some might say that keeping a copy of these materials without the permission of publishers is an infringement. But, the whole purpose of copyright exceptions is that one may keep copyrighted material if those materials are within the scope of the exceptions. The JNU Data Depot is clearly within the scope.

We strongly believe—and eminent legal experts agree with us—that this is legal.

☸

Before I conclude, I would be remiss if I did not discuss who let the elephant into this room, into our laboratories, our ivory towers. It was us. We are responsible. We are the ones that left the door open.

Since the beginnings of journals, back to the Philosophical Transactions of the Royal Society and before, the dissemination of scholarly results has been a collective endeavor.

When we present a lecture at a conference, this is not an experience for which we demand a fee. When we peer review a colleague's paper, we do this as a service to our profession. When we submit a paper to a journal, we do so for fame and glory, or to promote the increase and diffusion of knowledge, not for bags of gold.

Until recently, the editing of journals was always a matter of prestige and learning. Even today, it is almost always done for the advancement of science, not for the pursuit of pecuniary riches.

For the last 50 years, journal prices have skyrocketed to absurd heights even as costs have plummeted. The academic publishing "industry" has engaged in predatory practices, the oligarchy has built the walls ever-higher.

This is our fault.

When we complain, the elephants tell us there is no free lunch. They say the process is by its nature exceedingly expensive and complicated, they say publishing is difficult, they say we must not sacrifice quality by permitting amateurs into their walled garden.

This is nonsense.

A.J. Liebling once said "freedom of the press belongs to those who own one" and in today's world of the Internet and computers, we all own a printing press, we all have ready and easy access to the means of distribution.

Publishers are now pretending to embrace openness, but they are doing so with smokescreens. Article publication charges are ridiculously high, lock-down periods are unnecessary and in many cases illegal, exclusive publishing platforms with abusive terms of use are a travesty.

This is not open. Publishers are attempting to fool us with their jadoowallah routines.

But, I am pleased to report that the elephants are on the run.

Plan S is a great start. The funders of science have said this must all change. They have said enough is enough. Artificial metrics are a lazy alternative to judging individual articles on their merits. Open publishing must be a requirement, not an alternative. Plan S makes it clear across the globe that change must come.

I was delighted to see the strong support of Plan S from Dr. VijayRaghavan and the Government of India. He joins leaders from around the world in supporting your efforts as scientists and scholars.

But, we must go further.

We must not depend on scholarly merchant houses for the dissemination of knowledge. As scholars, we must take back control of science. If a high-quality open journal does not exist, make one. If you publish your results, you must do so openly. There is no excuse to do otherwise.

I have great hope for the future of open publishing. The wind is at our backs. People like Dr. VijayRaghavan are helping row this boat. It will happen.

But, we must still stand on the shoulders of giants if we wish to reach new heights. This is why text and data mining on the existing corpus is so important. This is why we have created the JNU Data Depot.

You are no doubt familiar with Mahatma Gandhi's seven social sins. Science without humanity is of course one of those sins. One of the great things about science in India, for the most part at least—and particularly so with those of you gathered here today—is that science is practiced with humanity. So many of you are working to solve pressing problems that afflict India and the world. It is inspiring.

But, another sin is commerce without morality, and I put it to you that that prohibiting text and data mining—telling a scientist who believes she can better find a cure for a disease or understand the roots of poverty—that is the very embodiment of that sin. Gandhi Ji would have been aghast at this unholy proposition.

Today, knowledge has been colonized. You are being told you cannot make knowledge without paying a tax and securing a license. How is this different than being told you may not make salt without paying a tax and securing a permit?

Knowledge and salt are both vital to the functioning of society, both are essential to human life. Taxing knowledge and prohibiting its consumption, taxing salt and prohibiting its consumption, these are both examples of commerce without morality.

You are no doubt also familiar with Gandhi Ji's use of the concept of swadeshi, one he of course adopted from the works of many others throughout India. Most people think of this in terms of Bapu and his spinning wheel, making kadhi as a way of fighting the colonization of India.

This was bread labor, the idea from the Sermon on the Mount that you should "earn thy bread by the sweat of thy brow." Bread labor and swadeshi, public work and satyagraha, these are the tools the Mahatma knew would lead to swaraj, to the liberation of India.

What you may not know was that bread labor for Gandhi Ji was originally not the spinning wheel, it was the printing press. At the Phoenix Ashram, their bread labor was that everybody must typeset every day. It was the increase and diffusion of knowledge that would educate the people.

Before satyagraha must come awareness of their condition among the people, and awareness comes from knowledge. Only with knowledge could the shackles of the Raj be removed.

Today, it is knowledge that has been colonized and I believe we cannot ignore this elephant in the room. We cannot sit back and allow others to say that it is wrong to pursue knowledge, that it is wrong to try and alleviate poverty, that it is wrong to make technologies that are appropriate to our modern lives and the problems our society face, that will help enlighten humanity.

Gyan swaraj is a great challenge of our times. It is a true frontier of science. You are all public workers, you have devoted your careers to the pursuit of knowledge, to the betterment of our society, to educating our youth.

If there is a way to carry out that mission better, we must not shirk from that task. If we all stand up and say text and data mining is an important frontier, then it will be so. But, it will only be so if we do this together.

Jain Hind. Jai Gyan.

Thank you very much.