INTERNET TOWN HALL

GYAN SWARAJ:
BUILDING A TRULY PUBLIC LIBRARY OF INDIA

Prepared Remarks of Carl Malamud
Shree Guru Gobind Singh Tricentenary University (SGT University)
Gurugram, Haryana, India, July 30, 2019

Thank you Professor Radhika for that kind introduction and to Shri Amogh Dev Rai for the invitation to speak here today.

My friends.

Chancellor Shri Ram Bahadur Rai Ji; Chairman of the Trust Shri Manmohan Singh Chawla Ji; Pro Vice-Chancellor Tuteja Ji; the Director of Academics, Professor Mittal Ji,

Ladies and Gentlemen. Good morning.

It is a great pleasure to be here at SGT University today. I am particularly pleased to be here at a university that was created with a holy cause in mind, a holy cause of propagating the message of Shri Guru Gobind Singh Ji, the great philosopher and social reformer, a message which says "the spread of learning is the best service to mankind."

Learning is indeed the best service to mankind, it is something that you will begin here at the University, but I hope it is but a temporary stop in your path, the beginning of your travel on the road of lifelong learning.

Knowledge is of course essential as you train for your chosen profession and search for economic opportunity. Lifelong learning is essential if you are to teach your children how to go down their own road, for if you do not read, if you do not learn, you cannot expect your children to follow their own path.

But learning is more than the key to earning a good living or being a good example to your family. It is the key to something much bigger, and that is the topic I wish to address today.

Learning is the key to solving pressing problems such as global warming, poverty, and disease. Learning is the key to achieving the goal of making India a $5 trillion economy. Learning is the key to creating a better and more just democracy.

What I wish to talk to you about today is gyan swaraj, why open knowledge—knowledge available to all Indian citizens, knowledge that is freely accessible to all—is the core underpinning of any truly great society.

☸

Before I get to that topic though—about knowledge here in India—I'd like to tell you two stories about my own experiences in the United States. I want to tell you those stories so you can understand my perspective, but also because I believe these goals are universal, that they go beyond India to the rest of the world—and because I would like to explain why I have come to believe that if there will be a revolution in access to knowledge, it will have to start here in India.

For the last 40 years, I have had the privilege—and the honor—to work alongside tens of thousands of my peers on what has become the global Internet, a network of networks that has grown to billions of computers, that connects billions of people. Of course, we must not forget that the Internet still only reaches half the world, and our job must not be considered complete until access to the Internet becomes truly universal.

During the 1980s, when I started working on computer networks, there were many different brands and kinds of computer networks. The idea that any one network might connect the whole world was considered a wild dream, hubris, a dream we all shared perhaps, but one we had to admit was unlikely to become a reality.

The Internet in those days was a network running on a particular set of network protocols. It connected a few hundred computers, then a few thousand computers. There was a file called "hosts.txt" which listed all the computers on the Internet in it. If you wanted to use a modern new service called "e-mail," you'd pull up this file to decide which computer to send your mail to.

But the Internet protocols were not the only game in town. There were commercial networks from companies such as the Digital Equipment Corporation, which they called DECnet. IBM had their own networking solution for connecting big corporate mainframes together, something called the System Network Architecture.

The protocols we use today were developed by a rag-tag group of engineers from around the world called the Internet Engineering Task Force, the IETF. We met three times a year, but most of our work was conducted by email. By rag-tag group, I mean we didn't even exist: there was no formal corporation or association or society, we just worked together.

There was another set of protocols that was much more formal, and it was called the Open Systems Interconnection model, or OSI. It was developed by the oh so very formal International Standards Organization and it was backed by all the great and the good. IBM supported it, the U.S. Department of Defense supported it, all the telephone companies loved it, the European Commission poured buckets of money into it.

OSI was based on the idea of a "smart network." What that means is that the telephone company would build the network on top of their existing lines and provide users with a number of value-added services.

The only surviving example is the Short Messaging Service or SMS, what you now know of as texting. In the early days of SMS, the telephone company offered the service for 25 cents to send a text, and then the receiver was charged another 25 cents to receive it.

In the OSI model, the smart network would allocate a channel for you. You could pay extra for a high-priority channel. You could pay extra for a secure channel. There was a menu of services, and everything was served a la carte, you paid for everything. There was no all-you-can eat buffet.

The menu had the services that the telephone company decided you could have, you couldn't go back into the kitchen and ask the chef to whip you up something different. You certainly couldn't have your own kitchen, or even a hot plate.

Besides being a so-called smart network (which meant a really complicated network because all those services led to incredibly complicated specifications), OSI had one more attribute. The standards documents that defined the network were published by the International Standards Organization and were amazingly expensive.

I spent thousands of dollars buying these standards to learn how the network worked. The standards were copyright restricted and you couldn't share them with anybody else.

By contrast, the Internet I worked on, the Internet Engineering Task Force protocols, were based on a dumb network. The telephone company was our enemy. Our protocols were based on the end-to-end principle. The only job we gave the telephone network was to take bits in one end and throw them out the other side.

To us the telephone system was a dumb pipe. We didn't even care if the network lost the bits (which it often did), our protocols would keep on sending them until a copy made it to the other side.

The end-to-end principle meant that all the smarts were on the computers on the edge of the network. This meant that anybody could create a new service, like Tim Berners-Lee did when he created the World Wide Web, like the engineers who created our modern email did, or streaming protocols for audio and video, or chat rooms, or file sharing, or whatever bright idea any of you here today might have for a way to change the world.

The other thing that made our network different was that our standards were totally open and our process was open. Anybody could read them for free and anybody could copy them. Anybody could propose a new standard.

We had a strong practical ethos: our motto was "rough consensus and working code." If you wanted your idea to be a standard, you had to convince other folks that what you had in mind worked not just in theory, but in practice.

By contrast, the OSI model was all about theory not practice, and as my colleague Marshal T. Rose often reminded us, the distance between theory and practice is far greater in practice than in theory.

What I learned from the Internet was that open works, and what I learned is that there is always some stranger out there that is smarter than you are. There were many, many occasions in which a standard wasn't working properly and some person nobody had heard of would speak up and say "I know a better way."

Innovation always springs from unexpected wells, you can't micromanage that process from the top. The Internet came into being because it was open. Only open can scale. Only open works.

Let me tell you a second story before I turn my attention to India. In the early 1990s, I was running Internet Talk Radio, the first radio station on the Internet. I ran it as a non-profit organization because I was a big fan of our own public radio and TV networks and because I believe in public work, but also because I wanted the flexibility to do things that weren't radio.

One of the things I did was put the U.S. Securities and Exchange Commission database on the Internet for free access. The system was known as EDGAR, the Electronic Data Gathering and Reporting system. In the U.S., all public corporations are required to file periodic reports.

This is what makes our financial markets open and transparent. Companies have to file quarterly and annual reports. Before a company does an Initial Public Offering, an IPO, they have to file extensive disclosures. Mutual funds have to file reports with the SEC. There are many other such reports that make our markets work properly.

In 1993, those reports were very expensive. If you wanted to read IBM's annual report, because maybe you were a journalist reporting on IBM or a student who might want to work for them, you spent $20 or $30 for each report.

The SEC's model was that ordinary people couldn't consume the basic reports. They weren't smart enough. The reports weren't usable. So, they had set up a system of "value-added." They spent $35 million to put together a scheme where a company called Mead Data Central was the data wholesaler, they "added value" to the raw feed.

They then sold feeds for hundreds of thousands per year to data retailers, who "added value" to the documents and sold them to people. This was a $300 million/year market.

Congressman Edward Markey—who chaired the congressional committee with oversight of the financial industry—asked me one day why EDGAR wasn't available on the Internet for free, and I told him I didn't know, but I'd look into it.

I found no conceivable reason why this public information shouldn't be available to the public. The SEC's argument was that no ordinary person would ever want to read these specialized documents except for a few well-heeled Wall Street fat cats, so why should we subsidize their document needs when everybody could make a nice fat profit selling them EDGAR?

I didn't buy that reasoning. So, I decided to do something about it.

I went to the National Science Foundation and got a grant to buy EDGAR from the SEC and put it on the Internet. Think about this a second. I got a grant from the American government to buy the data from the American government to give it away to the American people.

When my grant was announced in the New York Times, all hell broke loose. The vendors screamed bloody murder. A powerful congressional chairman, Mr. Dingell, threatened investigations as to why the National Science Foundation was competing with the private sector.

Luckily, Vice President Al Gore called the New York Times and called this "a big win for the American people," and that bought me a couple of months, so I scrambled and got that service up and running on the Internet.

My SEC service was wildly popular. We reached millions of new users. Senior citizen investment clubs, journalists, students, academic researchers, day traders, corporate employees all suddenly started reading these reports.

After running the service for a year and a half, I shoved it back down the SEC's throat by saying I would terminate the service in 60 days, that it was the SEC's job to run this system, and 17,000 people wrote to the SEC agreeing with me. When the SEC complained they couldn't possibly get this up and running in 60 days, I gave them our source code, loaned them computers, configured their Internet line, and they were on the net.

Two things happened. First, there was a dramatic change at the SEC. When I was running the service, the computer staff over there were very unhappy. They went around telling people our system was unsafe, it was undermining their value-added retail chain, the Internet would lead to viruses getting introduced, and so on and so on. They were not fans.

But, after they took over the service, they went to their bosses and they got to buy fancy new computers, and they found themselves running the busiest web server in the federal government and they got really happy really quick. It was geek heaven for the IT staff.

The other thing that happened was the retail EDGAR industry totally changed their tune. One of the vendors that was making big money selling SEC documents and had protested the loudest came up to me and said "you know, Carl, I wasn't expecting this, but our revenues went up instead of down!"

What had happened is we vastly expanded the number of people reading these documents, and those that were serious about SEC documents went from my free service to the professional offerings, because they had better search, and a better collection of documents, and integrated into their workflow.

Having the core data be open doesn't conflict with private industry's ability to make a tidy profit reselling government data. They just have to exercise their smarts to do something better with the data instead of creating an artificial monopoly.

The lesson from both these stories is that open works. If you want to build a vibrant economy, you need a core open infrastructure that anybody can use. If you want a city with shopping malls and corporate headquarters and factories, you also need public parks and roads and transportation services.

☸

Let me turn my attention to India now. I'd like to give you a few examples of some of the work I'm doing with my colleagues throughout India.

The first example is technical knowledge. I had travelled to India many times in the 1980s and 1990s, but it was always as a visitor. I finished editing my book about databases on a houseboat in Srinigar. When His Holiness the Dalai Lama wrote the foreword to my book about the Internet World's Fair I helped create, I travelled to Dharamshala to present him with a copy.

What brought me back to India was when I started posting technical laws on the Internet. By technical laws, I mean public safety codes, like building codes, fire codes, plumbing codes. These codes have the force of law, and are essential to the public safety, but all over the world they have been sold for high prices and access and use has been restricted.

My work began in the U.S., posting the building code of California, my home state. These codes cost over $1,000 to purchase, and they are an essential part of California law. Even our municipal building inspectors have to buy the codes. A kid that wants to study for a plumbing or electrician's license has to buy the codes.

I bought those codes, scanned them, and posted them on the Internet for free access. We didn't just buy and scan them though—we retyped many of the codes into HTML so they worked on modern browsers and were accessible to the blind. We redrew the diagrams into the modern SVG format, so you could cut and paste the diagrams into your own documents. We exposed the standards to search engines like Google so you could find them.

This work was not without controversy. In the U.S. and in most of the world, these technical codes are made by private not-for-profit organizations who make model codes, and then advocate for governmental authorities to turn them into law.

The National Fire Protection Association makes the model National Electric Code, then has successfully lobbied all 50 states and the federal government to make that code binding law. Despite that, the National Fire Protection Association claims they have the exclusive right to sell this law, and to decide who may or may not read it, and on what terms.

They were not pleased with my work, and the NFPA has joined with 5 other standards development organizations to sue my organization in U.S. federal court. We initially lost the case on copyright grounds, this was then overturned by a unanimous opinion in our favor from the U.S. Court of Appeals, and the case is still pending.

A related case, over our posting of the official laws of the State of Georgia, has just reached the U.S. Supreme Court. Again, we lost at the District Court level, we won a decided victory in the U.S. Court of Appeals, and now our highest court is going to decide the question of edicts of government.

Who may read and speak the laws in a democracy in order to inform their fellow citizens about their rights and obligations under the law? That is the question the court has agreed to hear, and our position is that in a democracy the law is owned by the people.

The government functions as our trustees, they are there at our bidding, and nobody can restrict our rights to read and speak the laws by which we have chosen to govern ourselves. We must all know the edicts of our government for ignorance of the law is no excuse.

☸

In India, technical laws are created by the Bureau of Indian Standards, a governmental body. We purchased all 19,000 Indian Standards and posted them on the Internet for free access.

Over 700 of them have been transformed to HTML. They are all available in bulk, so you can download all the standards by a particular committee or download everything.

The service is wildly popular with students across India, with local and state officials enforced with ensuring public safety, with architects and builders, with factory workers and consumer organizations, with engineers and building owners, with farmers and headmasters, and with ordinary citizens interested in the safety of the world around them. Millions of people have accessed these standards.

These Indian Standards are no casual publication of the government, they are the best codified technical knowledge about public safety in this country.

This is no incidental process. Standards are rules and regulations created by the government. Two Union ministers, 2 members of Parliament, 5 state ministers form the Indian Standards Council and oversee a process with thousands of engineers, government officials, and professors who volunteer their time to create these standards. The standards are then issued for public comment, then approved as an Indian Standard, and a notice is posted in the Official Gazette.

This is essential information. The National Building Code of India specifies proper exits in case of fire for schools, hotels, homes, and other structures, yet it is sold for 14,000 rupees per copy and comes with strict admonishments that you may not copy the building code or even small parts of it.

Indian Standards go far beyond the National Building Code. Many products may not be sold in India without a certification stamp from the Bureau of Indian Standards. In fact, the vast majority of the BIS revenue is from certification, sales of standards is a small fraction of that.

There are many other standards. Frequently in India, we read about workers cleaning sewers dying. But Indian Standard 11972, the "Code of Practice for Safety Precautions To Be Taken When Entering a Sewerage System," explains what every single worker entering these dangerous locations must know. Yet, the document is not widely available.

There are fairly frequent explosions and accidents in chemical laboratories, but it is exceedingly difficult to find Indian Standard 4209, "Code of Safety in Chemical Laboratories."

We know of the great damage caused by typhoons, but have you read Indian Standard 15948, "Guidelines for Improving the Cyclonic Resistance of Low Rise Houses and Other Buildings/Structures"? If you live in Odisha, this would be quite relevant!

After I posted those standards, I sent a letter to the Bureau informing them of what I had done, and offered to give them the HTML we had created and work with them to make standards more accessible.

I suspect I could have simply run the service and they wouldn't have even noticed, but it is an important principle of satyagraha—and have no doubt on this point, posting Indian Standards is an act of satyagraha—it is an important principle that one does not sneak around, one must be forthright.

Before Gandhiji marched to the sea to make salt, he sent a letter to the Viceroy, which famously began "Dear Friend." He told the Viceroy of his actions, and invited him to do the needful, but he did not. It was only then that Gandhiji marched to the sea.

Gandhiji was quoting Justice Ranade when he told us that we petition for two reasons. First it is to warn our rulers, but it is also to inform ourselves of our condition. We must educate ourselves before we may act.

After the Bureau objected to my actions, we petitioned the Hon'ble Ministry of Consumer Affairs, Food and Public Distribution and submitted affidavits from distinguished Indian professors of Engineering. We submitted lists of all the government regulations that used the standards. We documented before and after examples of how we transformed the standards to make them more readable for all consumers, and especially more accessible to the visually impaired.

The Ministry rejected our petition, so we filed a Public Interest Litigation writ before the Hon'ble High Court of Delhi. I am joined by two Indian co-petitioners, Sri Srinivas Kodali of Hyderabad, who is a noted activist on topics such as Aadhaar, and Dr. Sushant Sinha, the creator of Indian Kanoon, the prize-winning site that provides all court opinions for free to the public.

We are represented before the court by Sri Jawahar Raja, who successfully argued the Delhi University copyshop case and Sri Salman Khurshid, the former Minister of Law and Justice. The case is pending, and we appeared yesterday before the Hon'ble Court.

As I said earlier, selling the standards is not a hugely profitable or large business for the Bureau. The big money is in certification. But it is even more than that, it is about more than the money.

When people die in horrific fires, or suffer poisoning from improper application of pesticides, the costs to society in money and anguish is immeasurable. The purpose of public safety codes is to serve a larger purpose, to make our society safe, to make our economy function properly, to educate our engineers and farmers and construction workers, to enable our governments to regulate with knowledge.

Selling a standard is being paisa wise and rupee foolish. They must serve a higher purpose.

☸

Let me turn to a second example of work in India, the posting of books on the Internet for all to read.

There was something called the Digital Library of India, a project that was done under the auspices of the Government of India. They scanned 5.5 lakh books from libraries in 10 different scan centers. These are books in 50 languages and are a very unique collection of materials.

A few years ago, I noticed this collection. It was on a pretty awful server. It kept going down. It was really hard to use. Nobody answered the email. So I started making a copy of the files. I managed to make a copy of about 450,000 books.

I put those books on the Internet Archive. The Internet Archive is a non-profit organization, just like my own NGO, Public Resource. The Internet Archive is strictly non-commercial, they have over 5 million books on-line for free access, and another 10 million documents like newspapers.

There are lots of other resources there, including 5 million video files and 7 million audio files. They also run the Wayback machine, an archive of the entire World Wide Web since 1996.

I use the Internet Archive as my cloud for a number of reasons. It's strictly non-commercial, they never charge for content. It's an open architecture, so not only is there a great search engine and a user interface for end users, for developers like myself there are command-line tools that allow me to manage very large collections. I have several crore objects on the Internet Archive.

Most important though is that I can put data up for bulk access by others. If I upload 8,000 books in Tamil, another user can issue a simple one-line command and download all 8,000 of those books. It might take a while, because that's a lot of data, but bulk access is part of the design, not an afterthought. So, I view the Internet Archive as an initial loading dock.

Well, I uploaded these books from the Digital Library of India to the Internet Archive, and then the government server went down. It no longer exists. So, ironically enough, I have the only copy of the Digital Library of India. I have offered to make disk drives of this collection and bring them to India with all the transformed data and give a copy back to the government.

Now, this collection is, to be quite frank, not so good. The scanning was not done well. Pages are missing. The metadata is bad, titles are wrong. There are books in 50 languages, but the titles were all entered in Roman script, not native scripts.

Most importantly, the government was very sloppy on copyright. There were lots of books there they shouldn't have scanned. I was able to get 4.5 lakh books off their servers. I pulled the metadata up into spreadsheets and knocked out 50,000 books that were clearly in copyright. Things like Oxford University Press from the 1970s. I don't know what they were thinking when they scanned those books and simply put them on line.

That left me with 4 lakh books in my collection. There were still a few books in copyright, but that's always going to be the case with a large public digital library. The key to handling those issues is to respond to incoming queries. If somebody writes to the Internet Archive and says "hey, you have my book!" they get a quick answer and if the book is in fact in copyright, it is immediately removed from public view.

For the collection of books I maintain, we get very few of these copyright takedowns, and we've had over 5 crore views on the collection and it is well exposed on Google and other search engines.

We've supplemented that initial collection by harvesting other web collections on the net. For example, we mirrored 23,000 books from the West Bengal Public Library, and 8,000 books from the Tamil Virtual Academy.

For some of the languages, such as Telugu and Kannada, volunteers have been going through the collection and retyping all the titles and creators into native scripts.

There are other resources we've harvested from the net. With my friend Dr. Sushant Sinha, we've created a collection of 4.5 lakh files of Official Gazettes of India. This is the Union government, but also 17 states. And, unlike the sites run by the Government of India or the states, our archive is searchable. You can easily search on metadata, such as show me all the gazettes from Rajasthan or Kerala from the last week.

For languages—like English—where the Internet Archive does optical character recognition, you can also search inside the texts. For example, if you search the collection for the phrase "Vice Chancellor" you will find all 7,144 issues of Gazettes from 2009 through 2019 that have announcements pertaining to Vice Chancellors.

Dr. Sinha has developed a new program which we're in final testing on that allows us to start adding optical character recognition for most Indian languages. It's a pretty cool program, we pull the files out of the Internet Archive, bounce them off the OCR on Google Vision, then shove them back into the Internet Archive and recreate all the files they would have had if they had done the OCR themselves. It's a hack, but it is working great.

☸

The Digital Library of India, Official Gazettes, West Bengal Public Library, and much more are all harvested from the Internet. But, as I said, many of the scans are pretty awful. And, there are lots of materials that are not available on-line yet that really should be there.

There's a group of us that work in technology who are passionate about making materials available online that have banded together in an informal group we call the Servants of Knowledge. We have a high-end scanner installed at the Indian Academy of Sciences in Bengaluru, which has been highly supportive of our efforts.

Using that facility, as well as scanning facilities in the U.S., we've put a few thousand books on-line. The Bengaluru facility has done over 2 lakh pages. There are lots of books in Kannada and Malayalam, books in Sanskrit.

There is a collection I curate called Hind Swaraj that has the complete works of Gandhiji, Pandit Nehru, Dr. Ambedkar, and lots of other things, like the correspondence of Sardar Patel, books by Radhakrishnan, Rajaji, Tilak, Netaji, Gurudev, a huge number of other works by and about Gandhi, and much more.

The scanner we use there is called a Table Top Scribe, it is the same device that the Internet Archive uses to scan 1 million books per year. It has two high-end cameras. You put the book in a V-shaped holder, and you press down a pedal. That raises the book up to the glass, the cameras take pictures of each of the pages, you let the pedal down, flip the page, press the pedal again.

Once you get good at these Table Top Scribe scanners, you can do 800 pages an hour of very high quality. This operation of the pedal strikes me very much like a spinning wheel, and our motto for the Servants of Knowledge is "scanning is the new spinning."

You may be familiar with Gandhiji's passion for bread labor, the idea that you should do manual labor every day. He took inspiration from Tolstoy, and they both took inspiration from the Sermon on the Mount which proclaimed that we should all "earn our bread in the sweat of our brow."

When you think of bread labor, you may think of the spinning wheel, and those iconic pictures of the Mahatma at his charkha, spinning thread, making khadi. But what you may not know was that bread labor was originally something else.

When Gandhiji decided to form his first ashram at Phoenix, the first thing they did as they moved out of Durban was to dismantle the printing press. They put the press on 4 wagons, each pulled by a team of 16 oxen, and they hauled it out to the Phoenix Ashram.

There was nothing there at the time, just wilderness, no houses. But before they built shelter for themselves, they built a home for their printing press. At the Phoenix Ashram, everybody did typesetting, every day. Typesetting was their bread labor.

They used that printing press to pull in articles from all over the world and published them in newsletters like Indian Opinion. Today, you'd call Indian Opinion a blog. Gandhi was a prolific writer and publisher, satyagraha was his most famous tool, but it was the diffusion of knowledge to educate his peers that also an essential a tool.

We now have three of those high-end scanners in India, and my deal with the Internet Archive is if we can make all three of those sing, scanning lots of stuff, they'll us give us three more. The idea is a decentralized operation. We hope others will learn from our model and begin scanning whatever materials that are their passion, their history, their culture.

Scanning is the new spinning.

☸

I have one last example to give you of work in India, after which I will take a step up and conclude with some thoughts as to why this all matters. I've described how we take print materials (and I should mention we also take movies and audio) and scan and harvest them and put them on the Internet Archive.

That's a railroad train of content, a whole set of steps starting with scans, but including metadata and OCR and search engine optimization and then finally hosting with a search engine and user interface so you can access that content. But, that's not the end of the journey.

The content needs to flow back into India and be made available on many different platforms and in many different styles of use. When I come to India, I bring a bunch of little 2-terabyte drives that have a gold label on the front with a picture of Gandhi.

I call those Gandhi drives and they have copies of the Hind Swaraj collection, government books I've scanned, and much more. I give them to vice chancellors and headmasters of schools when I give speeches, and ask them to put the data on their local servers. I'd love for the government to take one of these drives and put their own content on their own servers.

That's an ad hoc effort, but over the last year, we've built something more substantial, working with my colleagues at Jawaharlal Nehru University and IIT Delhi. We started at JNU and built what we call the JNU Data Depot. It has 576 terabytes of disk on two large servers. Another 250 terabytes of disk is spinning at IIT Delhi.

There are two things we want to do with the data depots. One goal is to take the public resources we have on the Internet Archive and put them in India on these depots, and connect them to the National Knowledge Network, so any university in India can quickly and easily download data in bulk. We're just getting started on that capability, but it has great promise.

There is something else we are doing on the JNU Data Depot which is quite significant. There was an article just last week in Nature, the leading science journal in the world, about how India is leading the world in providing a significant new research facility based on our efforts.

One of the servers has been separated from the Internet. It is what we call air-gapped. It is carefully secured in the machine room, and we've locked it down with very limited access. On that server are 73 million journal articles.

Now, many of these articles are in copyright. We can't give you copies of those articles. There are pirate sites like Sci-Hub out there that let people get copyrighted articles, and those sites are very controversial. But, journals cost so much money, most universities can't possibly afford to subscribe to the scientific information that their students and researchers need to pursue their education and the progress of science.

In India, the top universities spend over 140 crore on access to journals and I've heard that when you add in all the universities and labs, India spends 1,400 crore on access to journals.

That's a huge amount of money and despite that, access is incomplete, so people in India and all the rest of the world use sources like Sci-Hub and Unpaywall and Research Gate. India is the second largest user in the world of Sci-Hub, but the U.S. and the U.K. are in the top ten because even at rich schools like Harvard, you can't get access to the literature of science you need to pursue your education.

We're not trying to solve the public access problem. At the JNU Data Depot we're doing something different. We're allowing scientists to do text and data mining on the collection of articles.

This is called non-consumptive use. They are not consuming the articles by reading them and they are certainly not distributing the articles. They are mining them for facts. There is no copyright on facts. Non-consumptive text and data mining does not violate copyright.

For example, we have a researcher that is looking at the secret language of plants. Plants communicate with each other and with other species using chemicals. She had previously mined a small collection of openly available articles on PubMed and built a database of the name of the plant, the chemical, the effect, and the geographic location. It has been very popular with those that research botany, and using our database she hopes to get significantly better results.

Another example is some research coming out of MIT. They are trying to predict which areas of science are going to be important in the future, a tool that would be invaluable to guide policy makers funding research, corporate researchers and venture capitalists, and even young scientists trying to understand where to aim their careers. Again, by using full text on all articles instead of just searching abstracts or metadata, they will get much better results.

☸

The JNU Data Depot is a bit controversial. You may be asking yourself, but what about copyright, how come you didn't ask permission?

Copyright is a bundle of limited rights for a limited period of time. It is not an absolute right. There are many things you can do with articles, even if they are in copyright.

If you are blind, I can give you any copyrighted materials without exception. There is an international treaty that has codified that right across the globe. Copyright is not absolute.

In India, if you are a student and I give you an article "in the course of your instruction," that is an exception to copyright as was so eloquently explained by Justice Endlaw of the Hon'ble High Court of Delhi in the famous Delhi University copyshop case. Copyright is not absolute.

It is well accepted that a library can buy a book and lend out that book. At the Internet Archive, they do controlled digital lending: they purchase a copy of a book that is under copyright and scan it. The physical book is locked away in the warehouse.

They then lend out the digital copy to one user at a time, using digital rights management software so you can't make additional copies. When the user is done reading the book, they check it back in, and the Internet Archive loans it to the next reader.

There is also a concept called fair dealing and there are exceptions to copyright when the materials have been transformed, used as raw material to create new knowledge. In the U.S., Europe, Japan, and other countries there are specific text and data mining exceptions to copyright.

This is because if you prohibit text and data mining, you are holding back the progress of science, and policy makers all over the world have agreed that would be against the interests of society. Copyright is not absolute.

With text and data mining, you are transforming the original text into other things. For example, you might create a digital signature of every word in every article and use that to detect plagiarism. You are not publishing the words, you are publishing a one-way mathematical transformation of the words. Again, not a copyright violation.

Copyright is a limited sets of rights. There can be no ban on the sharing of knowledge. The Nitisatakam teaches us that "knowledge is such a treasure which cannot be stolen." We must take that teaching to heart.

☸

It is my contention that private publishers, particular in the field of science, have over-reached. They would put it to you that they have absolute rights, and they maintain it is up to them to tell you what you may and may not do with ideas.

This is wrong, this is the colonisation of knowledge. A company such as Reed Elsevier acts as a self-styled East India Company of Science when they tell scientists they may not stand on the shoulders of giants without paying an exhorbitant toll, that they not promote the increase and diffusion of knowledge through text and data mining.

Because knowledge has been colonised in today's modern world, scientists have become the new Indigo farmers. They take out grants to grow their raw materials and they ship them off overseas, ironically enough in many cases to the United Kingdom. Then, they have to buy back high-priced finished goods.

What is Sci-Hub but an unlicensed salt factory on the edge of the ocean of knowledge? Perhaps Sci-Hub is illegal under current laws, but it is clear that Sci-Hub is of great use to the people of India and the students of the world. Salt and knowledge are both essential to life.

One cannot tell people they may not have knowledge, one cannot tell people they may not have salt. If that is what our world looks like, then our world must change.

☸

Now you may say that universal access to human knowledge is perhaps of importance to a few college students, but what about the real problems facing our world?

What about the crisis of global warming and our neglect of that crisis? Rising temperatures and typhoons. What about the crisis of pollution? India has half of the fifty most polluted cities in the world. Respiratory illnesses are fast overtaking malnutrition as a leading cause of death.

What about poverty? India has a surplus of food yet 200 mllion people are dying of starvation. What about the crisis of disease, those $1000 pills that cost pennies to make and could cure fatal illnesses, if only we could afford them?

And, what about the growing economic disparity in India and in the world? All over the world, the rich are getting richer.

I put it to you that access to knowledge is the first step. It is the first step towards a $5 trillion economy, because that goal depends on innovation and education and that means we must empower all youth and all bright minds, not just those born into rich families. You can't predict innovation, it always comes from the must unlikely places.

Access to knowledge is also crucial to building the political will to solve the problems our governments neglect. Democracy is based on an informed citizenry, it fails if we do not inform ourselves and each other. If we are not tacking global warming in the capitals of the world, that will only change if the people of the world stand up and say enough, we must do better, we must fix this problem.

☸

Do you know what made America a scientific, educational, and cultural powerhouse at the end of the 19th century? We had many things wrong in our country in the 1800s. Slavery. Racism. Corruption in government.

But, we did two things right. First, there was a commitment towards universal education. That commitment was sometimes not carried out in practice, witness the dismal state of the schools for people of color. But the U.S. created the country with the highest rate of literacy, the largest reading public in the world.

The second thing was a flood of cheap books all over the country, and a policy of promotion of the dissemination of knowledge through cheap postal rates that encouraged the proliferation of newspapers.

We did many things wrong, but one thing we did right was the dissemination of knowledge. It made America what it is today.

You can see the same thing here in India. The Bengali renaissance helped lay the groundwork for the swadeshi movement, for the call for swaraj. Gandhi was a prolific publisher, read Young India or Harijan and you will be amazed at how much he wrote.

In Tamil Nadu, the passion for the Tamil language stoked the fires of liberation, led by scholars like Rajaji, in Mumbai look at the profound learning of Lokmanya Tilak and the role he played in disseminating information and ideas.

The liberation of India began with the propagation of knowledge.

Changes comes from knowledge. That change can be revolution, or it can be evolution, but when knowledge gets colonised and limited and rationed, you don't get evolution you get stagnation. John F. Kennedy once said "those who make peaceful revolution impossible will make violent revolution inevitable."

I put it to you that if there is to be a revolution in access to knowledge, a peaceful revolution, it must come from India. Gyan swaraj means we must reject the colonisation of knowledge, just like India rejected colonisation and rule by the raj and showed the way to freedom for the rest of the world.

There can be no $5 trillion economy if India must beg permission to access knowledge in order to educate her youth, to promote new businesses, to solve pressing problems in society, to promote science and invention and culture and history and language.

A first step towards gyan swaraj is the creation of a true public library of India. There are vast treasures buried in the libraries of India, there are vast treasures inside the Government of India that lay fallow and unused.

It would not be unreasonable to set a national goal of creating a truly public library of India, to digitize 3 million books a year for a decade and create an open access repository of 30 million books and other resources. The sums expended would be minor in the national scope of things and the benefits would be immeasurable.

Such an effort would have to be decentralized, one cannot depend on one government agency to create something so mammoth, it must be created by the people. But, the government can do it's part. There is so much government information that could be made available much more broadly to help kick-start the effort.

Why are technical standards locked up? Why are the archives of All India Radio not available to all? What about the vast and amazing cultural resources in the Indira Gandhi National Centre for the Arts? The scientific resources of CSIR and the Indian Council for Agricultural Research and other research powerhouses? The comprehensive language resources in the Central Institute for Indian Languages?

☸

I want to close on a personal note, about you and about me. When you hear the words gyan swaraj, you may think this is a national goal, something they do in Delhi, something that politicians make happen.

This is a vending machine style of government. You deposit your taxes in the slot and the government delivers you a product.

But, you own your government. You are the raj in a democracy. Gyan swaraj must start with you.

When you hear the phrase "be the change you wish to see," you may think that is advice one gives to others, about how they should apply themselves. But Gandhi was always very clear that this was advice one has to give to oneself. Swaraj is about self-rule for yourself. Gyan swaraj is about your own personal commitment to learning and to knowledge.

Swaraj is also about something else. It is about public work. The liberation of India came about because many successful people gave it all up for a broader cause. Successful lawyers like Rajaji and Sardar Patel and Motilal Nehru and Gandhiji quit their jobs and devoted themselves to public work, to uplifting the people of India.

Why am I working in India on the cause of access to knowledge? It is because there is a tradition of public work here. I look at people like Aruna Roy, who gave up her plush job in the IAS and moved to a village in Rajasthan and spent the next 25 years laboring for the people, culminating in the Right to Information Act, the most powerful such law in the world when it was enacted.

India has a crying need for knowledge, but so does the whole world. I believe India has a tradition of public work, of tackling difficult problems, that makes this peaceful revolution possible here.

But, this will only happen if we embrace goals of gyan swaraj, if we take up the tradition of public work that created the modern India, and lead the way not only for India, but for the world.

Thank you very much. Jai Gyan. Jai Hind.