# Gyan Swaraj:
# Access to Knowledge is a Prerequisite for Progress

Prepared Remarks of Carl Malamud
National Institute of Plant Genome Research (NIPGR)
New Delhi, January 16, 2020

## *1. Thank you for the kind invitation*

- Dr. Gitanjali Yadav
- Dr. Aashish Ranjan

## *2. Summary of my talk today*

- Text and Data Mining
- Access to Knowledge More Generally
- Close with Why This Matters

## *3. Text and Data Mining - Old School*

Christopher Booker worked for 34 years and read hundreds of books. He wrote a book called. "The Seven Basic Plots." The plots plots include: Overcoming the Monster, Rags to Riches, Rebirth, The Quest. Reading lots of materials and then coming up with new results is a time-tested tradition. We always stand on the shoulders of giants.

## *4. Text and Data Mining in Science*

This technique is used in science as well. In a 1901 paper, Karl Pearson, collated data on typhoid inoculation, applied statistical analysis to collate the data from 7 sources. He concluded that typhoid inoculations were not effective. Despite that, the British army kept on using the typhoid vaccines. People don't always listen to scientists.

In a 1940 book on extra-sensory perception, Dr. Joseph Rhine looked at 145 reports on ESP, and attempted to categorize the state of the art. He concluded that ESP explained the results he was seeing, that the results were not guess work. It turns out he was wrong and the real answer was that all the experiments were flawed. Not all scientists should be listened to.

The technique of "meta-analysis" is widely used, particularly in bioinformatics, medicine, and other fields. The technique ystematically assess the results of previous research to derive conclusions about that body of research. It is often used to assess the results of clinical trials.

## *5. Today, computers are used to greatly increase the scope and speed of TDM.*

While Christopher Booker read a few hundred books, a study called "Transformation of Gender" used the Google Books database to read 100,000 novels from 1703-2009.

They came up with a non-intuitive result: the proportion of female characters in novels actually declined. This is perhaps because the proportion of fiction works authored by women decreased over the same period.

## *6. TDM is particularly important for the sciences.*

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 is an important research for those working in the field.

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, building on some of the pioneering work of Acharya JC Bose. 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. She did this by painstakingly reading a large number of journal articles.

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.

## *7. Let me now turn to the JNU Data Depot*

For the last two years, I have been working with my colleague Dr. Andrew Lynn of the School of Computational and Integrative Science at JNU. Andrew and his students have created what we call the JNU Data Depot. I am very pleased that Dr. Lynn and our doctoral student, Mr. Nitin Kumar, are here with us today.

The JNU Data Depot contains over 75 million scientific journal articles that have been gathered from a variety of sources, many of which have overlapping sources. Some of those sources are PubMed and the ArXiv (“archive”) repository. The sources also include a large number of copyrighted articles.

The JNU Data Depot is built on the legal principal that allows what is called "non-consumptive" use of copyrighted materials for research purposes. Non-consumptive means that the articles are not being consumed: they are not being read, they are not being copied for others to read. This is not SciHub, this is a research facility for computer programs to do text and data mining.

Our facility is modeled on the HathiTrust Research Facility in the United States which provides researchers with access to all the books scanned in the Google Books. A number of court cases have clearly established the legality of this facility under U.S. law, stating that non-consumptive use is an exception to copyright. An exception to copyright means that you can used the copyrighted work for the specific purpose for which the exception is provided.

This same legal principal has been carefully established in Europe, Japan, and many other jurisdictions. This is because text and data mining is an essential tool not only for modern science, but for modern business. Web search engines are an example of text and data mining.

In India, we established a very distinguished advisory body of legal experts, and carefully constrained our system to meet the parameters of Indian law.

In support of our position, we have put on the record two legal analyses. The first is by Professor Arul George Scaria, a leading Intellectual Property expert at the National Law University in Delhi.

The second analysis was submitted by Dr. Zakir Thomas, a senior member of the civil service, who submitted his analysis in his personal capacity. Dr. Thomas served as the Register of Copyrights for the Government of India. He is also very familiar with the progress of science, having served on assignment to CSIR to manage the Open Source Drug Discovery program.

We are grateful to Professor Scaria, Dr. Thomas, and the other members of our advisory body for helping guide us.

## *8. Characteristics of the System*

The JNU Data Depot consists of over 500 terabytes of disk on several large Network Attached Storage computers. In addition to the facility at JNU, a replica of the system is in place at IIT Delhi under the guidance of Dr. Sanjiva Prasad, the Head of Department for Computer Science and Engineering.

To meet the requirements of Text and Data Mining, this system is carefully secured, and separated from the broader Internet. Remote access is not allowed. All users of the system must agree to strict terms of use which are modeled on the same terms of use at the Hathi Trust Research facility.

The core data gathered from sources are PDF files. We call this level of raw data "tier 0." The text and images are then extracted using common utilities. Some of the PDFs are just scanned images, so OCR tools such as tesseract are used. Text extraction into raw text, XML, and other formats uses utilities such as pdftotext and grobid. Images and other components are also extracted.

The extracted text files are called "tier 1." This is the level on which computer programs may search the corpus. In addition to the text of articles, a number of other databases are added, such as crossref, which contains metadata.

The top level is called "tier 2," and this contains material that is not under copyright. That can include any public domain databases we have added, but it also includes facts that are extracted from the underlying articles. Facts, of course, are not subject to copyright. In addition to the scientific facts in tier 2, we will be adding other repositories, such as open source software and segments of the Public Library of India, which I will be discussing shortly.

## *9. The Science*

A number of science projects are underway to make use of the JNU Data Depot. We have assisted Professor Haeussler on his Genome Browser, and he is now examining our first pass through the data, and found 643,000 results, a significant increase in the size of his database. A second pass will be made through the data to refine the results.

For Dr. Yadav's work on the innovative Essential Oil Database, we are convinced that we can greatly increase the scope of coverage. Dr. Yadav has been joined in this effort by Trans-Disciplinary University of Ayurvedic Medicine, a leading university that has over 100 Ph.D. students that conduct hard science analysis of Ayurvedic plants.

They do some fascinating research. For example, to test the efficacy of a plant that is purported to cure hepatitis, they grow livers using stem cells, inject both the livers with hepatitis, and one with the medicine.

We will be using their database of over 9,000 Ayurvedic plants in addition to lists of plants that Dr. Yadav has accumulated from a variety of sources  to search the full scientific literature. This project is just getting underway, but we are hoping to have it in full operation this year.

A third project which is also just getting underway is being conducted by Dr. Lynn and his laboratory in cooperation with the University of Virginia School of Data Science. They are hoping to couple the scientific literature with the Wikipedia, Wikidata, Scholia, and other public services.

The idea is to mine the scientific literature for facts, that can be used to substantiate and document claims in systems such as the Wikipedia.

There are a number of other projects being considered. Considerable work is being conducted to refine and improve the core extraction. This is a major undertaking, and crucial to our efforts.

One of the projects I am particularly intrigued by was inspired by the work of the Principal Scientific Advisor to the Government of India which was presented recent at the National Frontiers of Science meeting which Dr. Yadav convened on behalf of the Indian Young Academy of Sciences.

At this meeting, a number of grand goals that the Principal Scientific Advisor to the Government has set were presented. One of those was to take the abstracts of all scientific journal articles and translate those into the major Indian languages.

I have been working with my colleague Dr. Sushant Sinha in Bengaluru, and we are very intrigued by the use of neural network based machine language translation techniques that allow one to upload word pairs for specific domains of knowledge, which are then used to improve the functioning of the neural networks that do the translation. Systems such as Google's machine language translation cloud products use these techniques.

We are hoping to be able to assist the Government of India in this commendable activity and are conducting some preliminary investigations into practicality of doing this translation at scale in languages such as Hindi, Tamil, Kannada, and Gujarati.

## *10. Let Me Turn Now to the Bigger Picture*

Text and data mining on the scholarly corpus is one example of making information available in ways that were previously not possible. This is, of course, the great promise of the Internet, allowing people to communicate and learn in ways that are new. The Internet has changed the world, but we have just begun that journey and there are some important challenges in front of us.

I'd like to briefly describe three other efforts that are underway in India that address some of these challenges. Those efforts are the posting of all Indian Standards on the Internet, a second project that has posted all the Official Gazettes of India in a searchable collection, and finally a massive effort to create a Public Library of India.

## *11. Indian Standards*

It is a well-known principle of any democracy that ignorance of the law is no excuse. For this reason, the law has long been exempt from copyright. In a democracy, the law belongs to the people, not to the politicians and the babus, and anybody is free to read and speak the law to inform their fellow citizens of their rights and of their obligations.

In our modern world, some of the most important laws are technical public safety standards. In India, these are promulgated by the Bureau of Indian Standards, a government agency.

These standards cover some of the most important aspects of our modern life. The National Building Code contains a chapter on life-safety, such as the requirements for proper exits in hospitals, hotels, schools, and office buildings in case of fire.

There are many other such standards which cover the safety of toys, the safety of bicycles, transportation of hazardous materials, the safe application of pesticides in agriculture, the safety of textile manufacturing machines in factories.

There are many codes of safe practices as well. One that is particularly compelling is IS 11972, "Code of practice for safety precautions to be taken when entering a sewerage system." Every year, hundreds of people die working in sewers. This code of practice lists specific precautions every worker should know before entering.

There is much more. Over 19,000 standards have been created. A great many of them are directly incorporated in legislation and regulations. The standards are overseen by a Council consisting of two Union ministers, 5 state ministers, 2 members of parliament, and a raft of senior civil servants.

The standards are created by 650 committees of government workers, members of industry, distinguished professors, and others who all volunteer their time. People are not paid to develop the standards, they do so as public service.

Each standard is issued for public comment in draft, and when approved by the committee and then the Standards Council, becomes an official Indian Standard and is noticed in the Official Gazette of India.

Despite the essential role of these rules and regulations, the Bureau of Indian Standards requires each of these documents to be purchased, often for high prices. The National Building Code, for example, costs 14,000 rupees. They assert strong copyright, strictly prohibiting any duplication of the standards without securing permission and extracting additional revenues.

The money they receive is not important to the Bureau. The vast majority of their funds, well over 95%, are received through their mandatory certification program. If you sell many products in India, you may only do so if they have been certified. This applies to hundreds of products from motorcycle helmets to steel and cement to skimmed milk and bottled water.

India is not unique in selling and restricting access to public safety laws. This has been the case all over the world, and my NGO has been leading the fight to change that situation.

In India, we purchased and posted all 19,000 standards for free and unrestricted access. In the case of several hundred key standards, we retyped and set them into modern HTML format, redrew the diagrams into high quality SVG vector graphic files, and even recoded the mathematical formulas into MathML.

The standards have been wildly popular, gaining millions of views. But, our efforts were not well received by the Bureau. After we received a rather strident letter protesting our efforts, we petition the Ministry of Consumer Affairs, submitting affidavits from leading engineering professors in India, evidence of our substantial transformation of the standards to make them more usable, and other pertinent details supporting our efforts.

The petition was rejected and I, along with two of my colleagues here in India, have submitted a Public Internet Litigation writ to the Hon'ble High Court of Delhi requesting affirmation that our efforts are squarely within the boundaries of the law. The matter is pending, and we are hopeful.

In the meantime, we are continuing our efforts to improve the usability of the standards. We are hopeful this year to begin machine language translation of some of these key standards into Indian languages.

## *12. Official Gazettes*

A second project we have undertaken is to mirror all the Official Gazettes we can find into a searchable archive. The Union government and several states do a good job making their gazettes available, but none of them are searchable. You have to specify the date of the issue you want, and in many cases the user interface on the government systems is exceedingly difficult to use.

For the 19 states and the Union government, we have gathered all the gazettes we can find, a collection of over 4.5 lakh PDF files which we house on the Internet Archive. Using the Google Vision system, we have added optical character recognition in all the Indian languages, so the entire collection is becoming searchable.

In some instances, such as Kerala, there is a complete collection of all gazettes going back to the formation of the state and back into the days of the Raj. In other cases, such as Karnataka, the collection only goes back to 2009.

In other cases, such as Uttar Pradesh, we can't find the gazettes at all! Over 200 million people, and the Official Newspaper of the Government of Uttar Pradesh is nowhere to be found. So, there is still much work to be done.

## *13. Public Library of India*

The last project I'd like to briefly discuss is our collection of over 4 lakh books on the Internet Archive, the largest open library of books about India on the Internet. The collection is now getting 5 crore views a year. We call it the Public Library of India.

The core of the collection was a copy of a system the Government of India had put together, scans of over 400,000 books from a dozen scan centers called the Digital Library of India. I made a copy of the government system and uploaded it to the Internet Archive, and then the government server crashed or was shut down. It has been offline now for several years, so ironically enough, I appear to have the only copy of the former Digital Library of India.

The scans the government did are not very good. Pages are often missing, titles are mislabeled, they often were very sloppy on copyright. We've cleaned a lot of that up, but the collection is still quite flawed. I've also mirrored a number of other repositories around India, such as the West Bengal Public Library and the Tamil Virtual Academy.

Despite the flaws, this repository is the only copy of some very important works. It includes over 46,000 works in Hindi, 33,000 works in Bengali, 33,000 in Sanskrit, 19,000 in Gujarati, and much more.

In addition to the digital books from the net, we've added thousands of high-quality scans. Working with the Indian Academy of Sciences in Bengaluru, our volunteers—who go by the name the Servants of Knowledge—have been operating a high-end scanner known as a Table Top Scribe. They have scanned close to two thousand books in Kannada and a host of important science books in English, such as parts of the personal collection of JC Bose.

We are also assisting the World Konkani Foundation create a repository of hundreds of Konkani book.  Our India Culture collection includes over 10,000 videos of Indian art, and many classic cultural texts, such as Rajaji's Ramayana, Tilak's analysis of the Baghavad Gita, and Radhakrishnan's Upanishads.

I'm particularly proud of the Hind Swaraj collection, which includes the complete works of Gandhi, Nehru, and Ambedkar. There are over 500 works by and about Gandhiji, as well as 129 audio files of the Mahatma speaking on All India Radio after prayer meetings. The collection also has many of the collected works of the Prime Ministers and Presidents, the full set of the correspondence of Sardar Patel, and much more.

## *14.  As I Prepare to Close …*

Let me say that I am often asked why this is important. People tell me there are many more pressing problems in our world, a collection of books in Sanskrit might only be of interest to a few people, Official Gazettes are just for the babus, who cares about Indian Standards since people will ignore them.

Such a view is wrong. I have learned in over 30 years of making large databases available on the Internet that we are always surprised at how popular these databases are, how many people have a thirst for knowledge you could never have imagined unless you make the information available.

Every time I post a database on the Internet—and the India collections are no exception—millions of people rush to use them. People are smart. You can never doubt that.

I put it to you that access to knowledge is a precondition to democracy and essential to the progress of society. Access to knowledge will not, by itself, provide the will to stop global warming, or to solve the growing economic inequality in our world, or to find the will to cure poverty and disease.

But, without access to knowledge, we will never solve any of those problems. In a democracy, an informed citizenry is the key to progress. In a democracy, we own our government. If our society is governed in an unjust way, or if we live in polluted cities, or if our neighbors are dying of poverty, famine, and disease, it is up to us to change that. No one is coming to save us. It is up to us.

If you believe that the very purpose of a

**REPUBLIC** is to secure to all its citizens

**JUSTICE,** social, economic and political;

**LIBERTY** of thought, expression, belief, faith and worship;

**EQUALITY** of status and of opportunity; and to promote among them all

**FRATERNITY** assuring the dignity of the individual and the unity and integrity of the Nation

If you believe that, then I put it to you this cannot happen if you do not live in a world where universal access to knowledge is a human right. We cannot make our world a better place unless we live in a world

WHERE THE mind is without fear and the head is held high;

Where knowledge is free;

Where words come out from the depth of truth;

Where the clear stream of reason has not lost its way.

Gyan Swaraj is about that "heaven of freedom," it is about a country and a people that are awake. Access to knowledge is a human right.

Thank you very much.

Jai Gyan. Jai Hind.