इलेक्ट्रॉनिकी और सूचना प्रौद्योगिकी मंत्रालय
भारत सरकार

Ministry of Electronics & Information Technology
Government of India

REPORT OF COMMITTEE - A ON PLATFORMS AND DATA ON ARTIFICIAL INTELLIGENCE

July 2019

Contents

1. Glossary

Initially we define a few key terms to ensure an easy reading through the report.

a. **Data Platform:** an online tool or services that hosts publicly available data sets from government agencies, universities, public institutions and other organizations, usually under open terms of use.
b. **Open Data**: data that is free to access, use, modify, and share by any entity usually but not necessarily, accompanied by the requirement that derivative works based on the data are attributed to the source of the data.
c. **Open Source:** software whose original source code is made freely available and may be redistributed and modified by any entity.
d. **Framework:** the required infrastructure to support the management, deployment, execution and implementation of an idea, regulation or policy.
e. **Cloud:** the infrastructure that allows users to deploy and run code remotely on machines that are accessed and available over a network, including the internet.
f. **Machine Learning:** algorithms and techniques that allow computers to "learn" from and make predictions on data. It is the science of getting computers to act without being explicitly being programmed.

2. Executive Summary

This report **recommends the development of an enriched National Artificial Intelligence (AI) Resource Platform (NAIRP) of India:** a platform that will bring together all publicly shareable data, information, tools, literature, solutions, best-practices to enable a large number of people to individually and in collaboration take up AI tasks to fuel all aspects from capacity building to building solutions in different domains that will benefit the society, enrich national prosperity and enable international cooperation. The platform will also have scope for sharing and driving standards, policy guidelines, entrepreneurship and developing a creative economy.

It is expected that this platform will be built in a contributory and participatory manner by all stakeholders, initially primarily driven and mainly funded by the Government along with Academic and Research Institutions, Industry and corporate bodies, Entrepreneurs, and

Thought Leaders. The platform will be a key component of the Indian AI ecosystem which will also consist of Knowledge Parks, Mission Programmes and Projects, Capacity Building and Re-Skilling and Policies and Guidelines. This National AI Resource Platform (NAIRP) has the potential to develop into a central repository of various components of the AI Ecosystem making it a critical knowledge integration and dissemination base.

This National AI Platform (NAIRP) will be an Open Data and Knowledge-cum-Innovation Platform that will enable usage by all categories of users for a variety of purposes including but not limited to training, research, projects including educational, competitive, funded and mission projects, start-ups and commercial development for socio-economic good. It will encourage the highest quality talent and innovators from all over the country and world to participate in this programme and help solve national challenges.

This platform will also catalyze the development of a partnership/ collaboration/ contribution/ participation model for knowledge sharing, data sharing, meta-data structure, annotation, API framework, IP creation, innovation, value added AI services, government adoption and human interactions.

The success of the National Digital Library of India (NDLI) Project (https://ndl.iitkgp.ac.in/) will be replicated for developing this AI repository and the potential of the data.gov.in resource can be the starting point to develop this National AI Resource Platform (NAIRP).

It is recommended that the initial development of NAIRP be carried out in project mode to be funded by the Government of India through MeitY. This Committee is willing to take up the responsibility to carry out this project through a National Institute of Importance such as IIT Kharagpur (which has successfully developed the National Digital Library of India) in collaboration with National Informatics Centre (NIC), other academic institutes and industry partners.

To develop a public open platform, a network of partners will be essential. . A unique meta-data and date / knowledge sharing paradigm will be adopted for populating and integrating elements in the platform. A gap analysis of the existing open data policy and how best to ensure recommendations above for quality data platform can be initiated. A set of metrics will be defined for evaluation of the platform for its performance, efficiencies and utilities. Finally it will disseminate the information, knowledge and awareness of AI through various on-line initiatives including those related to skilling, projects, policies and innovation.

It is estimated that the development will require a budget of about Rs 100 crores over a period of 3 to 4 years after a detailed proposal and scoping is made.

3. Background on AI/ML

Artificial intelligence (A.I.) is no longer the realm of science fiction but a practical software tool used to help millions of people every day. This has been driven by recent breakthroughs in the field of machine learning, a branch of artificial intelligence that specifically studies algorithms which learn and improve from training examples. While these developments have been decades in the making, they are only now becoming practical because of the availability of computational power, a growing community of talent across the globe and richer sources of information and data.

It has become imperative for both the government and industry to embrace AI in order to remain competitive, generate new economic growth, drive social progress and improve the health of our environment. Globally, various countries are leveraging data and associated IPs as a strategic asset and for global dominance.

India's capability and readiness in Artificial Intelligence (AI) will be one of its major drivers in the Knowledge Economy. India has been reported to be among the top 3 locations in the globe (after the US and China) for the development and iteration of AI related technologies[1] and has been praised as having over 58%[2] of its technological uses of AI in the implementation stage, i.e beyond pilot and test projects.

a. Key requirements

Elucidating a few key requirements that act as pillars for developing an AI ecosystem

i. Capable Talent

The availability of well trained and contemporary workforce is the key to building a sustainable ecosystem of AI in a nation. The creation of such capable talent is usually

[1] These three countries are winning the global robot race, Rishi Iyengar. Available at: http://money.cnn.com/2017/08/21/technology/future/artificial-intelligence-robots-india-china-us/index.html

[2] India leads the artificial intelligence race thanks to the local offices of US firms, Ananya Bhattacharya. Available at: https://qz.com/1073903/india-leads-the-artificial-intelligence-race-thanks-to-the-local-offices-of-us-firms/

predicated upon the availability of experts, existence of premier institutions to train budding computer science graduates and creating a demand in the industry to hire graduates in the field of AI and machine learning. Re-skilling, i.e. the training of current employees or the external workforce in frontier technologies such as AI, is also another avenue available to jump start the demand and supply cycle in the industry for well trained talent. There are quite a few good global examples on re skilling that we can pick up globally to replicate here in India from countries like US, Singapore and Canada. In India for example, NASSCOM has already developed 'FutureSkills' platform for enabling training of workforce on 8 emerging technologies. Towards it, NASSCOM is also developing the details of the job roles, the associated skillsets expected by the industry and model curriculum for the various job roles in vogue. Measures that can be taken on a policy front to further enable this include incentivising graduates to pursue AI and machine learning via scholarships, ensuring the retention of AI experts by offering competitive remuneration in universities, allowing the percolation of industry experts into academia in the form of guest lectures and funnelling funding for research in AI technology at a Central and State level. A further aspect which while not directly skilling related is the aspect of dispelling the concern that AI will take away jobs. The Policy initiatives should also focus on ensuring that there will be the creation of more high quality jobs due to AI.

ii. Infrastructure

The development and deployment of AI technologies requires vast infrastructural resources, both in terms of raw computing power as well as network connectivity. While few educational institutions and research institutes have dedicated supercomputers, these resources are usually booked for months (if not years) in advance for long term research projects. In order to enable access to such resources to the average student, the only scalable solution is to make cloud based infrastructure available at affordable prices or for free with minimum permissions required to access it. Unlike traditional computer programmes, which can run on a single PC or laptop, AI and ML programmes take a lot more processing power to run and can take days to complete. Therefore, policy should focus on making such resources available to universities, research centers and vetted independent practitioners by either setting up infrastructure dedicated to this task or by collaborating with the private industry to make these

resources available at a low cost with a centralised cost-negotiated model. Thus the cloud ecosystem which is critical for the same needs to be encouraged in the country. The first focus should be to encourage the setting up of data centres in the country by looking at incentives related to lighter regulation for dark fibre, sustained power, incentives for green energy etc. The second area of focus would need to be around areas like cross border data flows, proper security processes to ensure the sanctity of the data etc.

iii. Enabling regulatory environment

The rapid pace of change in the AI and ML space requires an agile regulatory environment where academicians and practitioners can play an active role is asking for and implementing policy changes. This will allow the right resources, knowledge and opportunities to be made available where they are most required and will have the maximum impact. The ideal way to do this is to create a broad framework which allows for decisions to be taken at units like a university or research centre or an organization level with budgets and minimal appropriate guidelines guiding their actions. The benefits of such distributed decision making guided by core guidelines will allow for resources to be directed to where they are needed the most while ensuring accountability. In order to enable regulation with experimentation, it may be worthwhile to consider a light touch regulatory environment with a sand boxed approach which enables innovation and at the same time ensures security and privacy of data and also remaining within certain ethics and legal boundaries which prevent and harm and ensures no discrimination

iv. Industry, Government, Start-up demand

Jobs and employability are the key motivating factors behind the decision of students to pursue a stream in their education. In order to make India a leader in the field of AI, the focus of the industry needs, government projects and start-up initiatives to pivot from a pure services framework to a hybrid service sector and original product framework. While both these streams can exist concurrently, the creation of world leading teams that create original work for their organisations will not only attract talent

to the field but also make the outputs be utilised for its huge potential in the country and across the globe, adding to India's reputation as an innovation hub. This in turn will attract more funding as well as corporations to hire and provide scope for projects, further contributing to the growth of the sector.

v. High quality data sets

The AI of today is heavily reliant on the collection, usage and processing of big data. Bernard Marr[3] notes that data is invaluable in AI devices understanding how humans think and feel, thereby speeding up their learning process. It is cyclical – *the more information there is to process, the more data the system is given, the more it learns and ultimately the more accurate it becomes.*

India enjoys a large and diverse population accompanied by a heterogeneous economic complexion. The volume of data available for AI algorithms to utilise is rapidly increasing thanks to flagship government programmes like Digital India, GSTN, etc. It could be discerned that the country is presently an untapped AI goldmine. All it requires is appropriate regulatory responses and coherent strategies to catalyse market growth and requisite research and development.

India could develop capabilities and enablers through

I. Collecting and increasing availability of data and knowledge in standard discoverable formats across data types (text, images, voice, video, etc. in multiple languages) that are indexed, searchable and retrievable on an open data platform to fuel AI solutions.

II. Making available Open-source Tools for data-retrieval, analysis and building ML solutions.

III. Encouraging various Ministries to experiment with machine learning solutions in different public domains. This could be done by identifying use cases in important

[3] Marr, B. (2017, July 15). Why AI Would Be Nothing Without Big Data. Retrieved November 25, 2017, from https://www.forbes.com/sites/bernardmarr/2017/06/09/why-ai-would-be-nothing-without-big-data/#6f3b02994f6d. https://cis-india.org/internet-governance/blog/artificial-intelligence-in-india-a-compendium)

domains around small tasks, acquiring the data required and developing ML driven solutions for that task. (eg. Ministry of Health building a model to predict disease spread patterns in relation to changing weather conditions. This can help them plug the spread of specific diseases by proactive outreach to the right demographic.)

IV. Government as a Consumer: Globally, all the leading nations has a strong local ecosystem where the solutions created are consumed locally. More importantly, AI has the potential to revolutionize the public sector—and save billions of dollars. Artificial intelligence already helps run government, with cognitive applications doing everything from reducing the time to detect & treat cancers, cutting costs to handling tasks we can't easily do on its own, such as predicting fraudulent transactions and identifying criminal suspects via facial recognition etc.

Increasing access to quality data sets is key to creating a competitive and equitable AI ecosystem, where innovation can flourish. While various committees set up by MeitY are deliberating on various key aspects to supercharge the AI ecosystem in India, this report would largely focus on recommendations on increasing access to high quality data sets through robust Data Platforms.

4. Data Platform for AI

Governments globally have a unifying role to play in enabling safe and ethical innovation through application of Artificial Intelligence and Machine learning. This role includes ensuring availability of talent, deployment of necessary infrastructure and networking systems (either directly or via third party providers), and maintaining the availability of open source machine learning tools/framework and high-quality data sets.

Increasing access to quality data sets is key to creating an enabling ecosystem.. Researchers need large swathes of high-quality data to train Algorithms systems for specific tasks. Governments are well placed to enable scaled aggregation and subsequent availability of these data sets in machine-readable formats.

Thus, a data platform in an Indian context should

1. Drive data acquisition efforts. Not all data platforms are created equal. Their usefulness can be widely classified by choices made across 4 parameters: a. Technical (data is aggregated as link collection, download catalogs, etc.), b. Access (data is made available through webforms or API support, etc.), c. Organizational (moderation and curation) and d. Integration (free-for-all upload, platforms that restrict metadata attribute values for more standardization, etc.).
2. Catalyze collation of knowledge. This includes the creation of knowledge repositories on AI tools, literature on latest research/solutions, trainings and other important resources.
3. Enable crowd-sourcing of AI/ML solutions.

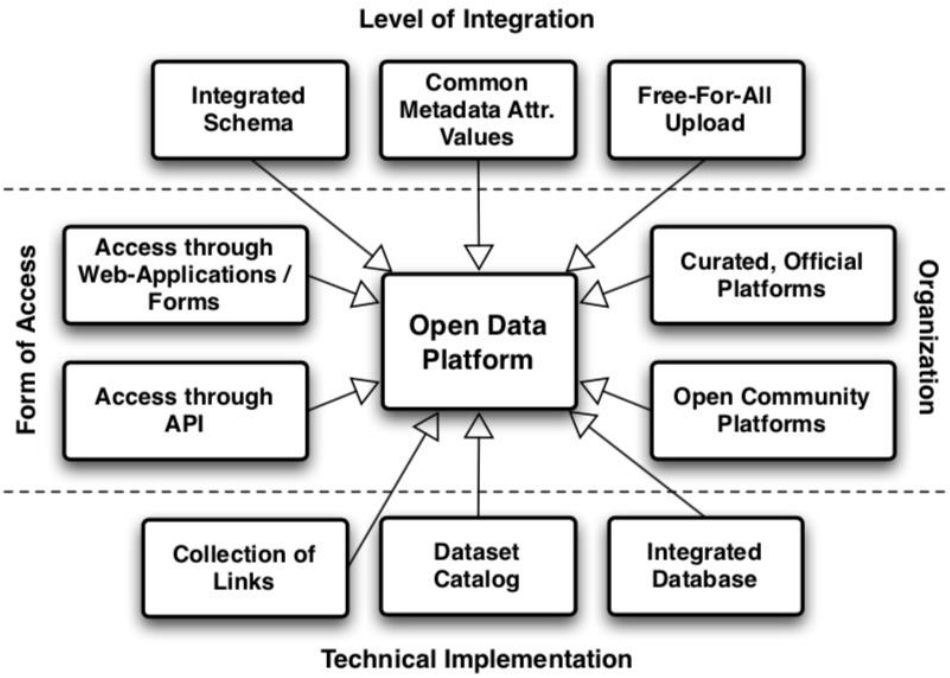

Image taken from: Braunschweig, Katrin, Julian Eberius, Maik Thiele, and Wolfgang Lehner. "The state of open data." Limits of current open data platforms (2012). Available at: https://wwwdb.inf.tu-dresden.de/opendatasurvey/www2012_short.pdf

Creation of a Technical Committee to help make these choices in an informed manner, while evaluating and constantly course correcting the progress (in terms of quantity, reach, effectiveness, data policies, etc.) of a data platform would be critical to the success of such an initiative.

a. Ensuring availability of Quality data

The utilization of published data sets greatly depends upon the nature in which the data is made available. Government agencies and other organizations generate large amounts of data through the various services they offer. Large swathes of this data in its raw form are not very conducive for collection, dissemination and eventual use. Data must be reliable, usable and continuously available. Therefore there is a general need to ensure that the aggregated data sets are cleaned, customized and made available in standardized formats with key considerations in mind.

Key factors that need to be considered to ensure the availability of quality and processing of data sets:

- **Sharing the right data:** This is one of the most important factors to any successful AI strategy. If the right data that is relevant to the work being done in the ecosystem is not released, it would reduce its effectiveness for practitioners and researchers. For example, providing highly localised weather data is pivotal to ensuring models developed to aid farmers can suggest more accurate actions or corrective measures. Hence, it is important to ensure that the right data strategically aligned to national priorities is periodically released and made available.

- **Interoperability**: It would be critical to develop and implement standards for data formats to ensure the data provided are operable in common AI platforms and frameworks used by AI researchers. Many such standards already exist with open licenses, certain governments (such as the ICO in the UK) have been proactive at implementing such standards which can be studied and suitably adapted to the Indian context.

- **Data Management Practices:** It is important to define the standards and processes for data collection, data sanitisation, anonymization / pseudonymization (such as encryption and or removing personally identifiable information, removing classified information, data labelling/annotating, storage, monitoring, reviewing, quality assurance and reporting). Having detailed security guidelines for the use of such data (if relevant) driven by industry standards and best practices is key to building trust in the ecosystem.

- **Data Privacy**: Implement privacy standards and controls like terms of use, choice, consent, transparency and accountability. This point closely ties in with Management of Data presented above. The Srikrishna Committee and its White Paper, which may be accompanied by a draft law, can serve as guiding points for the recommendations made by the relevant subcommittees. This should ideally include clear responsibilities in terms of data ownership, licenses, breach of terms or privacy etc.

- **Unbiased Data:** Data reflects the social, historical and political conditions in which it was created. Artificial intelligence systems 'learn' based on the data they are given. This, along with many other factors, can lead to biased, inaccurate, and unfair outcomes. It is essential to ensure that data being utilized for socio-technical machine learning systems remains free from bias (ethnic, cultural, etc.) to ensure non-discriminatory decision making. This is key to ensuring the principles of Fairness, Accountability and Transparency (FAT) are followed by the entities who utilise the data sets. Defining the criteria for the implementation of the methodologies that will be followed for ensuring such sensitivities are accounted for should be a multi-stakeholder effort governed by a technical sub-committee on applied ethics (with adequate representation from Government academics, industry, civil society and industry). International efforts such as the FAT conference and the Partnership on AI can be used as a sample yardstick for these efforts. It is important to note that a MeitY committee will surface a more detailed report on this key consideration.

- **Data Flows:** Artificial intelligence and Machine learning is an international phenomenon, drawing on researchers, computing infrastructure and data sets from around the world. Global flows of all types encourage growth by promoting innovation and raising productivity levels. Supporting liberal data flow regimes backed up with robust data protection practices would amplify this effect by broadening participation through effective knowledge sharing and creating a more efficient market.

- **Reliability**: AI systems must be designed to operate within clear parameters and undergo rigorous testing to ensure that they respond safely to unanticipated

situations and do not evolve in ways that are inconsistent with original expectations. People should play a critical role in making decisions about how and when AI systems are deployed.

- **Fairness**: In a country as diverse as India with different cultures, languages, religious beliefs, etc., it becomes critical that when AI systems make decisions about, they need to be fair

- **Future scope of expansions for higher data quality**

    - **Data harmonization:** Often data comes in different formats, forms and sizes. It is important to harmonize the data to improve the quality and utility of this data. Data harmonization techniques interpret existing characteristics of data and action taken on data and use that information to transform or suggest subsequent data quality improvements.

    - **Availability of Quality Data Models:** For those users who would like to use the data portal only to infer from the fresh data, it will be time consuming to create first a trained model and then infer. The proposed AI data portal can expand to create and provide reusable high-quality data models for those who would like to use the facility only for inferencing. This model can also be used in places where the raw data is sensitive and cannot be shared in raw format and instead only the trained model for inferencing could be made available.

    - **AI Data Cloud and DR Infrastructure**: It could be expected that by introducing the proposed AI data platform, the amount of data being aggregated would grow rapidly and exponentially along with high dependability on this portal for future AI developments. As a result, to ensure higher reliability, availability and serviceability, the underlying infrastructure would need to be beefed up (either directly or via third party involvement). The National Supercomputing Mission could be such a resource platform in addition to contributions from public and private partners and contributors.

b. Guidance and norms for platforms

The aggregation of quality data sets to enable research and development would have to be a well-coordinated collective movement across government, industry, academia and the technical community.

To ensure there is effective and continuous availability of data, the proposed technical committee should take the following broad principles into consideration:

1. **Equity:** In order to ensure an open platform, uploaders across public and private sector should share the good quality data sets in an open and interoperable format. Uploaders should have the necessary permissions to make the data available publicly and aimed towards identifiable problem solving. This principle will offer the framework which enables sourcing of data sets for AI/ML use. In order to ensure open access to the platform, the AI/ML models should be shared under open source license. This principle will enable its use in an effective and efficient manner.

2. **Openness:** Data should be accessible to anyone in the country from this repository as long as the use is defined (be it social or commercial). This principle will lay down the framework/conditions of downloading data. This will be based on different kinds of terms and conditions of the open license. Based on the license conditions, the downloader may have to attribute or use the data or model for either commercial or non-commercial purposes.

3. **Ethical Grounding:** AI/ML should be used for the purpose of development and to solve real world problems in a privacy preserving manner. The overarching legal framework could be derived from the Sri Krishna committee draft once it is ready.

5. Recommendations

1. Development of an Open National Artificial Intelligence Resource Platform (NAIRP) to become the central hub for knowledge integration and dissemination in Artificial Intelligence and Machine Learning. This can be built drawing ideas and learnings from

available national platforms like AI Singapore, efforts like Kaggle and Indian efforts like the National Digital Library of India and data.gov.in

2. Develop a generalized meta-data standard for NAIRP that will enable integration of a variety of resources including but not limited to data, tools, literature, etc. from multiple resources and owners of these resources.

3. Create mechanisms for data / meta-data harvesting and integration from all contributors and partners to ensure information in NAIRP is updated and owners take responsibility for their own data.

4. Strengthen data.gov.in and use it as a base data source of NAIRP for storing Government and other data from public sources.

5. Encourage unbiased, reliable, safe, open by default, inclusive data sharing. Suitably develop data standards, access, federation, usage, security, privacy and rights issues for data integration and dissemination in NAIRP

6. Use the National Digital Library of India (https://ndl.iitkgp.ac.in/) as the base source for literature and knowledge for NAIRP to enable rapid development.

7. Carry out a Gap Analysis to help all stakeholders to ensure that enriched quality data and information is made available through NAIRP

8. Create a Technical Committee to help Monitor and Evaluate the progress of NAIRP in terms of quantity, reach, effectiveness, data policies and other monitoring aspect.
9. Create a Data Committee to Monitor and Evaluate the progress of data / meta-data / links received from various public and private contributors in NAIRP.
10. Create an Ethical committee which can monitor the ethical aspects of the use of AI its interplay with the laws related to security, Privacy etc as well as having the ability to ask the relevant questions on practices followed by AI systems.

11. Create a User Community of registered users of NAIRP for access of data and resources in addition to public display of general information.

12. Create a NAIRP Club of AI/ML users and experts to rapidly annotate, curate, share and create India's AI/ML Knowledge Directory and Catalogue in a well-designed crowd-sourced manner using hackathons and other mechanisms for exponential knowledge / services creation and dissemination that will make this internationally unique.

13. Enable national and international players to participate in solving national problems using the NAIRP platform.

14. Partner with an appropriate Institution to develop a basic compute infrastructure for AI/ML around NAIRP from the National Supercomputing Mission Programme.

15. Enable development of Knowledge Verticals, Capacity Building, Training Programmes, National Missions, Commercial and Entrepreneurship ecosystem, Policies and Regulatory Framework around NAIRP.

16. Evaluate the need for developing an AI Data Exchange in the future with appropriate policies and guidelines in place.

17. Fund a project for 3 years to develop NAIRP with an initial budget of around Rs 100 crores through MeitY with all stakeholders.

6. Conclusion

As illustrated by this report, there is a pressing need for a public platform that can make data, tools and knowledge pertaining to AI technologies available in an easy to use, reliable and effective manner for researchers and the masses alike. The large push towards e-Governance and ICT technologies in general has enabled a vast amount of data to be captured and generated by the relevant government departments. After following privacy and security best

practices on this data, releasing it into the public domain as open data will attract some of the brightest minds in the field to identify and solve India's problems. The knowledge and best practices available on the platform will also serve a great utility to teachers, students and researchers for guidance in their individual learning processes by being an authoritative source of information. The industry will also benefit from these developments, as the overall quality and quantity of fresh qualified practitioners of AI technologies will increase over time due to such a programme. This will allow for locally sourced talent to be available for stakeholders to utilise according to their needs and the practitioners' interests. Finally, the government and society will also benefit from such a programme as the organic and systematic development of the AI/ML ecosystem will lead to a natural (and possibly exponential) increase in the applications of AI technologies for public benefit. The benefits ensuing to all stakeholders in the ecosystem along with the increase in the public's capacity to understand the AI are necessary first steps to making India a leader in the AI space. We hope that the recommendations presented in this report will aid the government in achieving this goal.