AI4India.org

CP RG CENTER OF POLICY RESEARCH AND GOVERNANCE

# DATADAAN

Donate for a Digital India

## CONSULTATIVE PAPER

# Introduction

Today, India is making rapid strides towards emerging as a global leader in Artificial Intelligence (AI) and Machine Learning (ML) innovation. However, researchers and innovators alone cannot create a flourishing AI/ML ecosystem and generate data-based solutions. A holistic, supportive, and harmonious ecosystem is essential. A key challenge for this 'innovation ecosystem' is the need for a wider access to high-quality local data.

Therefore, India should launch a nationwide programme, '#DataDaan', for voluntary data sharing for the training of AI models and innovation. This programme would enable start-ups and researchers building AI-based solutions to access relevant, usable, and contextually appropriate data for training models and analysis. Concurrently, the formal framework will ensure necessary safeguards for stakeholders such as private individuals, government bodies, and companies, who may participate in this process by voluntarily sharing data.

At the outset, for voluntary data sharing, the definition of data must be expanded beyond its common understanding. Apart from quantitative datasets, it may be textual (like news articles), audio (like customer helpline conversations), or video (like clips of online lecture recordings). Moreover, such data can be sourced from either a private individual (like medical records of deceased individuals obtained from the individuals' families, and books/documents written in local dialects/scripts), a private organisation (like footage of CCTV cameras installed at an office, purchase transactions of a local kirana shop, and employee attendance records), or public/government institutions (real-time traffic violations, data on citizen-state transactions, tax collection data), or any other lawful entity.

# 1. Potential Benefits

Voluntary data sharing has various benefits:

## 1

### Eliminating Algorithm Bias

Numbers (or data) may give unreliable answers if the underlying data is either not representative of the society or measured incorrectly. In such cases, AI models do not always remain objective indicators of on-ground realities, and can even end up perpetuating existing societal biases and stereotypes.[1] The resultant picture is problematic as in areas like law and order, biased algorithm-based solutions can have adverse effects.

In the Indian context, a lack of sufficiently diverse datasets representative of the country's deep internal cultural richness hurts attempts at algorithmic fairness.

To tackle such biases, India must prioritise a re-imagining of algorithmic fairness that is rooted in contextual data and models. The proliferation of diverse datasets commensurately representative of India's cultural diversity is critical for replacing models developed in the West. Besides, the resultant indigenous models can empower all communities equally and enable fair digital solutions in India.

Thus, granular, accurate, and real-time data becomes critical.

## 2

### Building Sovereign AI

With AI's profound implications for economic growth and security, governments globally are increasingly pushing to muster their AI infrastructure, capabilities, and industry. This is "Sovereign AI". It carries several benefits, including spurring domestic economies, building credibility in the booming global digital economy, and assisting in the emerging battlefields of cybersecurity and tech-supported national security.[2]

The use of domestically generated data is crucial for sovereign AI. Thus, through *'#DataDaan'* and similar efforts, India can enable an indigenous, novel, and innovative path to building India's Sovereign AI.

# 1. Potential Benefits

## 3

### Driving Innovation-Based Economy

Historically, the US has been recognised as a scientific powerhouse. However, accessing DNA data for research has long been complicated and circuitous, requiring multiple requests to private labs. This dramatically changed with the launch of the government-led Human Genome Project in 1990, an effort to map the entire sequence of human DNA. Importantly, the Project made its data open-source within 24 hours of its discoveries, unlike targeting and patenting their discoveries as was the norm back then. The results were clear. As per one estimate, in 2010 alone, a $3.8 billion public investment in the Human Genome Project generated over $796 billion in benefits and about 3,10,000 new jobs. Similar data-sharing arrangements substantially sped up the development of the coronavirus vaccines.[3]

Further, the United Kingdom's Transport for London (TfL) is a public agency that has shown how seemingly mundane data with the government can lead to exciting effects. Being the custodian of a monumental dataset of public transportation in London, TfL made a significant amount of it publicly available free of charge. It provided, among other things, accessibility to real-time and customised data, low latency, simplified operations, and support of common web and data formats.[4] As a result, currently, at least 600 apps - being used by over 40% of London residents - are being powered through TfL's dataset.[5] Moreover, data from TfL has contributed to several positive social and environmental shifts.[6]

India has also taken a few similar steps. For example, under 'BHASHINI' (part of the National Language Translation Mission), the government acts as an "orchestrator to bring (data) contributions" from several stakeholders into an open-source repository that will be validated and standardised by a Unified Language Contribution API. To aid start-ups, BHASHINI provides cloud credits and other computing infrastructure to reduce the costs of computational resources for start-ups. Furthermore, BHASHINI has proactively worked with the private sector to build different use cases and test proof of concepts for its AI-centric language translation and database platform. A similar example is Anuvadini, which is working towards strengthening digital technology in education and instruction. Besides, efforts like BHASHINI also provide its data to researchers.[7] In a remarkable step, the Indian Institute of Science's AI and Robotics

# 1. Potential Benefits

Technology Park announced in July 2024, under BHASHINI, that it will make over 16,000 hours of speech data from 80 districts open-source.[8] This is particularly noteworthy as access to speech data in India is highly limited. Other efforts include Anuvadini to promote 'bhasha daan'. Under the aegis of the All India Council for Technical Education (AICTE), Anuvadini employs AI to assist in covering the language gaps in delivering quality and standardised educational content across the country.[9]

A great example of the use of AI in solving key pain points of many citizens is the introduction of the AI-driven module 'Ideal Train Profile' in the Indian Railways. The mainstay of travel across varying distances in the country, the Indian Railways manages several billion passenger journeys annually. Despite such massive popularity, the persistence of the frustration of the waiting list has caused many difficulties for its potential passengers. Therefore, the Railways implemented an AI model that was fed with information on about 200 long-distance trains, including the premier Rajdhanis, to shrink the size of the waiting list by five to six percent. The AI model was able to offer insights into how passengers booked tickets, which origin-destination pairs were in demand and when, and finally, which seats remained vacant and for which part of the journey. The full-fledged model could significantly help the Railways allocate their resources more efficiently not only to assist more potential passengers in finding a confirmed seat but also to reduce its financial losses. Besides, the platform could enable a better real-time understanding of train reservations and make the train a competitive alternative for other modes like flights for short distances.[10]

Furthermore, Google's Maps Content Partners programme is a noteworthy example of private sector collaboration with governments. Herein, Google collaborates with municipal, regional, and national governments worldwide to integrate geospatial data into user-friendly platforms.[11] It is worth thinking about the potential benefits that can accrue when similar large-scale AI/ML programmes are made for other geographies.

These are just a few examples of the incredible possibilities of sharing and processing data. Sharing data and making it available for all users in an open-source manner can greatly facilitate innovation, encourage the growth of novel products, and induce healthy competition among firms and researchers in the economy. Similar data-sharing arrangements substantially sped up the development of the coronavirus vaccines.[12]

# 1. Potential Benefits

Going a step ahead of these globally successful models, India can work towards creating an enabling interface to facilitate a hassle-free experience for any stakeholder who wants to voluntarily provide their data for innovation or any firm that wants to access it. Finally, by helping private entities identify underserved or overlooked market segments, local data can support the development of targeted products to engage with these communities effectively. Additionally, this can benefit the companies by expanding their customer base. This will empower problem-solving that is unique to the Indian context.

### Reducing Costs for a Digital Economy

The early movers in AI have secured a monopoly on critically useful yet extremely expensive datasets. Some early movers have also adopted unscrupulous methods for data collection. Gradually, this has created economic barriers to AI innovation, as only a handful of players have access to adequate data.[13]

It must be recalled that data has always posed a high entry barrier for AI development and new start-ups. Costs around data collection, cleaning, labelling, and management are prohibitive, discouraging several start-ups and companies.[14]

However, as a by-product of #DataDaan-induced intensive data sharing, data would be accumulated at multiple points online at near-zero cost. This would have strong repercussions on data collection costs in several cases.

# 2. Tenets of #DataDaan

Creating a fair, inclusive, and enabling policy framework is the cornerstone for the success of *'#DataDaan'*. In such a framework, while the government may play the role of a facilitator/regulator, the prime focus remains on direct voluntary data donation to private firms and researchers.

For efficiency and effectiveness, *#DataDaan* may develop an interface to encourage data sharing that must adhere to the following tenets:

## 1

### Tenet 1: Trust

Since trust will be integral for the wider and sustained usage of the interface, the features of consent as defined by the Digital Personal Data Protection (DPDP) Act, 2023 may be adopted.[15]

Further, apprehensions about sharing data for consumption by private firms are commonplace. Such apprehensions often arise due to trust deficits from fears of data breaches and misuse. Thus, **data protection norms of the data sharing exercise should enlist the prevention and management of such data breaches.**

Furthermore, the purpose of data sharing may be identified and agreed upon to prevent regulatory confusion and data misuse. This practice would also ensure that data is being shared only for pre-defined legitimate purposes and 'fair use' is being promoted. For instance, in the UK, organisations wanting to pay to access datasets of the National Health Service (NHS) must provide a justification that they have a legal basis to use the data and will do so safely, securely, and appropriately.[16]

Besides, the interests of stakeholders like researchers must also be protected around intellectual property as well as hassle-free usage of data. Thus, their trust can be secured through the usage of relevant licences like Creative Commons 4.0 to support the pursuit of novel research questions fearlessly. The interface must institutionalise the principle that interface membership/data sharing/service provision is not equivalent to the transfer of intellectual property rights over the data accessed.

These measures could build trust across the spectrum — from civil society to corporations — to further the goals of *#DataDaan.*

# 2. Tenets of #DataDaan

2

### Tenet 2: Privacy

After the establishment of trust, privacy must be consistently sustained. Subsequently, any such voluntary sharing of data must incorporate 'Privacy by design'. For example, Personally Identifiable Information (PII) of people and entities should be protected across the data value chain by all stakeholders involved in the ecosystem as a matter of core principles.

At the same time, the interface must not be in a position to identify a specific individual unless the individual provides additional information and permission to enable identification. 'Anonymity by design' must be adopted.

These actions could also address concerns about unregulated data collection methods used to build existing AI products.

3

### Tenet 3: Democratisation of Access to Data

Thereafter, the interface would have to be democratic and meaningful for the intended benefits to be realised. This can be achieved in several ways.

First, attempts must be made to ensure that the data shared is standardised and usable. For instance, the U.S. Energy Information Administration provides free and open data in machine-readable formats to support the private, non-profit, and public sectors in finding novel ways to innovate and create value-added services. This dataset is used globally for energy-related statistics.[17] The Indian Government's National Data and Analytics Platform (NDAP) is a heartening step in a similar direction,[18] but still, a lot more is left to be achieved. It must be explored if data cleaning and conversion can be features of the platform. In many cases, it will prove infeasible to expect data shared to be cleaned and standardised. Various solutions, including AI-based tools, can be explored for improving data usability. Besides, voluntary communities can be formed to facilitate the sharing of data in a usable manner.

# 2. Tenets of #DataDaan

Moreover, active collaboration can be promoted to create more accurate, usable, and even new datasets from existing datasets. This would also significantly enhance data integration, thereby making data more accessible. For instance, TfL engages with navigation and mapping services like Google Maps, and Citymapper to provide end-users with real-time travel information and route-planning options.

Second, data ownership concerns must also be accounted for. In addition to the copyright issues, unauthorised onward transfer should be prevented.

Third, the interface can be an open-source network that any private individual or legal entity may join for free. While there may be no further special conditions to join, joining would translate into agreeing to comply with the defined rules governing the framework, especially the liability to be audited for data processing/usage.

# 3. Best Practices

### Data Marketplace Using DEPA Principles

Using Data Empowerment & Protection Architecture (DEPA) principles, a data marketplace can be offered to give individuals control over how their data is used. Specifically, its foundational principle of combining public digital infrastructure with private market-led innovation is of utmost importance. This eliminates the high barrier of entry by skipping the need to set up expensive and exclusive bilateral data-sharing mechanisms. Ultimately, it creates a marketplace characterised by data portability with the critical role of 'Consent Managers' for serving different market segments. Consent Managers also help in addressing privacy concerns.[19]

### Beckn Protocol

Another particularly exciting example of the government's role as a facilitator in a data-based economy lies in the Beckn Protocol. The Beckn Protocol drives India's Open Network Digital Commerce (ONDC). The Protocol enables a decentralised digital commerce that can support all stakeholders to engage with each other in an open and interoperable fashion.[20] This is an example of how a common data-sharing protocol can facilitate local players to be discovered across industries in e-commerce. The Beckn Protocol can also be replicated for other cases in India to drive economic growth.

# 4. Defining An Interface

*#DataDaan* will require an interface to operationalise the voluntary sharing of data. Such an interface can broadly take three forms, each with its advantages and disadvantages. Nonetheless, all such platforms could include a mandatory assessment of the potential impact on privacy for each case of data sharing, and an auditable trail every time the PII is viewed or used through technologies like blockchain.

1. Government Model - a platform created and managed by the government can enhance the relationship between two parties. However, its presence could be counter-productive, as it could create suspicions. Moreover, it might struggle to maintain critical services like rigorous potential impact assessment.
2. Private Model - a platform developed and managed independently of the government can enable direct donation of data. However, concerns like losing control over data may persist as well as entities like the government and rivals in the same industry may hesitate to participate.
3. DPI-based Model - a 'private' platform wherein the government also participates as a stakeholder might help more in building trust. Measures like 'Confidential Clean Rooms' (CCRs), already introduced in India through the Data Empowerment & Protection Architecture (DEPA) (also called the 'Consent Layer of India Stack'), can also be incorporated.[21] Nonetheless, defining the contours appropriately to cater to all stakeholder interests will be challenging.

However, the precise choice of the interface may be decided by the regulator on criteria it deems appropriate. Such criteria should include technological capability and compatibility along with the consistent ease of fulfilling requests.

# 5. Conclusion

Today, India is actively realising and striving to create conducive arrangements for data sharing. The DPDP Act, 2023 stands as a testament to India's will for change by promoting the creative usage of data for problem-solving while addressing data protection concerns.

#DataDaan is envisaged as a mass movement that requires proactiveness from all stakeholders, especially the common citizen. They must be informed, educated, and assured about how donating their data would help them in their daily lives. Whether it is an affordable AI model that assists in ensuring universal access to quality healthcare in rural India that performs the role of a medical compounder in a small town in India or one that uses real-time traffic data to reduce congestion, the articulation of the benefits must be clear and relatable. Contemporaneously, they must be made comfortable with the associated legalities, especially the fact of not having any legal liability for the use or misuse of the data donated.

*#DataDaan* supports these efforts, accelerating the development of India's conducive, responsive, and accountable regulatory framework. *#DataDaan* is the Indian path of serving the rising aspirations of a country that is once again finding its place globally. Above all, it could be a decisive facilitator for India's burgeoning AI start-up space. *#DataDaan* reflects India's commitment to a better future for all Indians — and by all Indians.

# References


1.
Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining algorithmic fairness in India and beyond. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Retrieved from https://arxiv.org/pdf/2101.09995

2.
World Economic Forum. (2024). Sovereign AI: What it is and ways states are building it. Retrieved from https://www.weforum.org/agenda/2024/04/sovereign-ai-what-is-ways-states-building/,

NVIDIA Blog. (2024). What is sovereign AI? Retrieved from https://blogs.nvidia.com/blog/what-is-sovereign-ai/

3.
Singer, N. (2021). Balancing Privacy, open data, and the public good. The New York Times. Retrieved from https://www.nytimes.com/2021/02/19/business/privacy-open-data-public.html

4.
Transport for London. (n.d.). Unified API. Retrieved from https://tfl.gov.uk/info-for/open-data-users/unified-api#on-this-page-0

5.
Transport for London. (n.d.). About TfL. Retrieved from https://tfl.gov.uk/corporate/about-tfl/

Deloitte. (2017). The economic value of TfL's open data and digital partnerships. Transport for London. Retrieved from https://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf

6.
Deloitte. (2017). The economic value of TfL's open data and digital partnerships. Transport for London. Retrieved from https://content.tfl.gov.uk/deloitte-report-tfl-open-data.pdf

7.
Ministry of Electronics and Information Technology. (n.d.). Bhashini Whitepaper ver 6.0. Retrieved from https://www.meity.gov.in/writereaddata/files/Bhashini%20Whitepaper%20ver%206.0.pdf

Financial Express. (2024). Digital transformation: Govt's AI translation platform Bhashini to tap paid service model. Retrieved from https://www.financialexpress.com/business/digital-transformation-govts-ai-translation-platform-bhashini-to-tap-paid-service-model-3457155/

Moneycontrol. (2024). In talks with fintechs, retail firms to integrate Bhashini platform: Amitabh Nag. Retrieved from https://www.moneycontrol.com/news/technology/in-talks-with-fintechs-retail-firms-to-integrate-bhashini-platform-amitabh-nag-12113511.html

Press Information Bureau. (2022). Digital India: Bhashini - Enabling Multilingual Internet and AI for Bharat. Retrieved from https://static.pib.gov.in/WriteReadData/specificdocs/documents/2022/aug/doc202282696201.pdf

8.
The Economic Times. (2024). Under Bhashini, IISc to open source 16000 hours of speech data. Retrieved from https://economictimes.indiatimes.com/tech/technology/under-bhashini-iisc-to-open-source-16000-hours-of-speech-data/articleshow/111639325.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

9.
All India Council for Technical Education. (n.d.). About Us. Retrieved from https://anuvadini.aicte-india.org/AboutUs

# References


10.
Indian Express. (2023). Indian Railways' AI module brings hope of shorter waiting lists for tickets. Retrieved from https://indianexpress.com/article/india/indian-railways-ai-module-brings-hope-of-shorter-waiting-lists-for-tickets-8398288/

11.
Google Maps. (n.d.). Content partners. Retrieved from https://contentpartners.maps.google.com/

12.
Singer, N. (2021). Balancing Privacy, open data, and the public good. The New York Times. Retrieved from https://www.nytimes.com/2021/02/19/business/privacy-open-data-public.html

13.
Heater, B. (2024). AI training data has a price tag that only big tech can afford. TechCrunch. Retrieved from https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford/?guccounter=1

14.
CHI Software. (2022). AI app development: Cost calculations and factors to consider. Medium. Retrieved from https://chisoftware.medium.com/ai-app-development-cost-calculations-and-factors-to-consider-8244acb17eef

15.
Ministry of Electronics and Information Technology. (2023). Digital Personal Data Protection Act 2023. Retrieved from https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf (please see Chapter II)

16.
Healthwatch Richmond. (2021). NHS sharing your data: What you need to know. Retrieved from https://www.healthwatchrichmond.co.uk/news/2021-08-31/nhs-sharing-your-data-what-you-need-know

17.
U.S. Energy Information Administration. Open data. Retrieved from https://www.eia.gov/opendata/v1/

Patterson, M., Singh, P., & Cho, H. (2022). The current state of the industrial energy assessment and its impacts on the manufacturing industry. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352484722010885?via%3Dihub

18.
National Data and Analytics Platform. (n.d.). Retrieved from https://ndap.niti.gov.in/

19.
NITI Aayog. (2023). Data empowerment and protection architecture: A secure consent-based framework. Retrieved from https://www.niti.gov.in/sites/default/files/2023-03/Data-Empowerment-and-Protection-Architecture-A-Secure-Consent-Based.pdf

20.
Open Network for Digital Commerce. (2024). Redefining digital commerce through Open network framework. Retrieved from https://ondc.org/blog/redefining-digital-commerce-through-open-network-framework/

21.
NITI Aayog. (2020). DEPA executive summary. Retrieved from https://www.niti.gov.in/sites/default/files/2020-09/DEPA-Executive%20-Summary-revised.pdf

**Acknowledgements**
The authors would like to thank all stakeholders who shared their inputs to strengthen the document. All errors are of the authors.



**AI4India.Org**
www.AI4India.org
ai4india@ai4india.org

**Center of Policy Research and Governance**
www.cprgindia.org
office@cprgindia.org