The India Datasets Programme

Putting the Cart Before the Horse

Ashwin Arun
12 November 2026

Data protection frameworks worldwide find their normative groundings in the right to privacy. The Justice Srikrishna Committee Report, which shaped India's Digital Personal Data Protection Act, 2023 (DPDP Act), recognised *Puttaswamy v. Union of India* as laying down the fundamental right to privacy as the primary value underlying data protection legislation. Consequently, the Indian DPDP Act governs only personal data, or data that can identify individuals directly or indirectly.

Like the DPDP Act, most data protection laws worldwide also exempt non-personal data (NPD) from regulation entirely, creating a legal presumption that data incapable of identifying individuals cannot threaten privacy at all. This presumption, however, is not entirely true. "NPD" is still a blanket legal term to describe data of varying degrees of anonymity, and legally 'exempt' NPD can still be used to re-identify individuals, and in aggregate, provide monetisable insights. This has led to regulators worldwide struggling to define the exact limits of NPD, since such an exercise requires delicately balancing legal definitions and technical standards to ensure some semblance of regulation falls upon entities handling NPD (for instance, Singapore, Canada, the United Kingdom, the European Union). It is surprising, then, to observe that despite worldwide efforts to define standards, and check the unrestrained processing of NPD, Indian regulators have so far sidestepped the regulation of NPD, and have powered through to trying to enable maximum monetisation. Through the lens of India's emerging NPD framework, namely the India Datasets Programme and its successor, AIKosh, this article argues that the rush to unlock economic value from anonymised data has outpaced the development of adequate anonymisation standards and privacy safeguards, putting the "cart before the horse".

Why Personal And Non-Personal Data Are Not Binaries

The DPDP Act defines personal data as data about an *'individual who is identifiable by or in relation to such data'*. Any data falling outside this definition is, by exclusion, non-personal under Indian law. This bright line optimistically assumes a clear-cut distinction. In practice, data lies on a spectrum of identifiability.

The Act prescribes no standards for adequate anonymisation, nor does it distinguish de-identified data (where direct identifiers are removed but indirect re-identification remains possible) from truly anonymised data (where re-identification is irreversible). The withdrawn Personal Data Protection Bill, 2019 had defined anonymised data as *'irreversibly transformed'* data, but the DPDP Act abandoned this standard without replacement.

The European Union's (EU) General Data Protection Regulation (GDPR), from which the DPDP Act borrows heavily, takes a more textured approach. Recital 26 excludes anonymous data from data protection, but defines anonymity as requiring that the data subject is *'not or no longer identifiable'* by any means *'reasonably likely to be used'* by any person. The GDPR accordingly treats pseudonymised data as insufficiently anonymous to bypass regulation. The Kris Gopalakrishnan Committee similarly

recommended treating de-identified data as personal, since it retains the potential for indirect re-identification.

Two risks follow from keeping these distinctions binary and unregulated, as the DPDP Act does.

First, datasets thought to be anonymised can be de-anonymised by artificial intelligence (AI). Rocher et al demonstrated that 99.98% of Americans could be re-identified in any dataset using just fifteen demographic attributes. Nyffenegger et al have shown that large language models can re-identify individuals from ostensibly anonymised datasets by cross-referencing disparate data points, and the European Data Protection Board's (EDPB) 2025 report on AI privacy risks confirms that generative models can *'correctly infer personally identifiable information [...] by stitching together information from disparate sources'*.

Second, aggregated NPD, regardless of its re-identification risk, can be used to profile and target specific ethnic, religious, or political groups, a risk scholars call *'group privacy'*, which most legal frameworks leave unregulated. NPD analytics can produce discriminatory outcomes at a collective level even where no individual is identified, circumventing the consent paradigm on which data protection law rests.

The India Datasets Programme And AIKosh: Economic Ambition Bereft Of Privacy Guardrails

NPD's economic potential has not gone unnoticed by the Indian Government. The draft National Data Governance Framework Policy (NDGF Policy), released by the Ministry of Electronics and Information Technology (MeitY) in May 2022, launched the India Datasets Programme, a *'data commons'* model in which a dedicated office under MeitY would facilitate and incentivise data-sharing between public and private entities. The NDGF Policy's treatment of privacy, however, is perfunctory, limited to a single sentence:

> *'4.1 The NDGFP standards and rules will ensure Data security and informational privacy.'*

In March 2025, MeitY launched AIKosh, a centralised platform under the IndiaAI Mission hosting over 300 non-personal datasets and 80 AI models. In March 2026, Right to Information (RTI) filings revealed plans to establish 50 AI curation units across ministries to upload 'non-personal' government datasets to AIKosh, without specifying how datasets would be classified as non-personal or what anonymisation standards would apply.

The Kris Gopalakrishnan Committee's 2020 Report forms the framework's backbone, asserting India's sovereign rights over data generated within its borders. The Report introduces *'data trustees'* (government or non-profit bodies) to manage *'high-value datasets'* for the community. However, the definition of 'community' is vague, the look framework for trustees is largely symbolic, and there are insufficient safeguards against conflicts of interest where government bodies act as both data sharers and trustees. The then-Minister Rajeev Chandrasekhar spoke of sharing data only with 'trusted' entities, without specifying criteria, while reports indicate a shift towards mandatorily requiring big technology companies to share their NPD, putting trade secrets and intellectual property at risk.

These concerns are not hypothetical. In 2019, the Ministry of Road Transport and Highways sold two vehicle-owner databases to private parties without user consent. Mishra and Agarwal characterise India's approach as a *'techno-political'* governance model that has bypassed traditional legislation in favour of executive policy and digital infrastructure, granting overbroad *'sovereign purpose'* exceptions for government data requests that the proposed 'Non-Personal Data Protection Authority' (NPDA) cannot adjudicate.

Technical Standards For Anonymisation As A Spectrum

Before pursuing NPD-sharing at this scale, India must develop authoritative anonymisation standards. Its only governmental attempt, a MeitY report on anonymisation guidelines for e-governance, was withdrawn within five days of release. Any standard must recognise that anonymisation is not binary but exists on a spectrum, with data granularity generally inversely proportional to the degree of anonymisation.

Two technical metrics for quantifying anonymity have found widespread acceptance. *K*-anonymity is an accepted metric for generality – a dataset is k-anonymous if every record is indistinguishable from at least *k*-1 others sharing the same quasi-identifiers. Differential privacy is another metric that measures the statistical noise added to a dataset to prevent the singling-out of any individual's contribution. Singapore's Personal Data Protection Commission (PDPC) has prescribed minimum *k* values for internal and external data sharing, offering a model which India could work off of. At the regulatory level, the United Kingdom Information Commissioner's Office's (ICO) 'motivated intruder' test asks whether a reasonably competent person could re-identify individuals using *'all the means reasonably likely to be used'*, which is a contextual legal standard that offers to ground the line between personal and non-personal data.

The Court of Justice of the European Union (CJEU), in *Breyer v. Bundesrepublik Deutschland*, carried forward the ICO's contextual approach in holding that dynamic IP addresses constitute personal data where a website operator has the "legal means" to identify the user. More recently, in *EDPS v. Single Resolution Board*, it clarified that pseudonymised data may not constitute personal data for a recipient who lacks the *means* needed to reverse the pseudonymisation – again, piggybacking off the ICO's 'motivated intruder' standard. The standard for anonymisation must therefore consider the person in whose hands the dataset lies. Alongside European decisions finding Google Analytics transfers unlawful for inadequate anonymisation, including the Italian Garante's holding that IP addresses remain personal even when truncated, these rulings indicate growing consensus that anonymisation must be assessed contextually, and not merely assumed from the removal of direct identifiers.

These are lessons for India. Naively anonymised datasets are dangerous, stripped only of direct identifiers, can be cross-referenced with publicly available data to re-identify individuals. Without standards addressing indirect re-identification through quasi-identifiers, AIKosh risks sharing datasets that are anonymised in name only.

Putting The Horse Back Before The Cart

The notification of the DPDP Rules in November 2025 advances personal data protection but remains silent on NPD, which stays entirely unregulated even as AIKosh accelerates the sharing of government datasets at scale. Notably, the commons-based model underlying the India Datasets Programme is not bereft of academic backing, with Gangjee endorsing it as a viable alternative to proprietary data rights following the failure of the EU's Data Producer's Right. Data's non-rivalrous nature means sharing can generate value without diminishing it for the original holder. However, this economic logic cannot be severed from the privacy concerns that threaten it, and a common model without robust guardrails risks the very harm it seeks to avoid.

Four reforms are necessary to put the horse back before the cart. Firstly, India must legislate binding anonymisation standards, drawing on *k*-anonymity, differential privacy, and the motivated intruder test, before permitting large-scale NPD sharing. Secondly, data trustees and the proposed NPDA must be genuinely independent of government ministries that are simultaneously the largest data controllers. Thirdly, group privacy must be recognised as a distinct regulatory concern, given that aggregated NPD can produce discriminatory outcomes even absent individual identification. Lastly, mandatory

data-sharing must be balanced against intellectual property protections and accompanied by fair compensation mechanisms, lest it amount to the nationalisation of data.

India's ambition to leverage NPD for AI-driven innovation is legitimate and ambitious, but its current trajectory leaves anonymisation undefined, group privacy unrecognised, and institutional accountability unaddressed. These lacunae strike at the very right to privacy that *Puttaswamy* enshrined as fundamental. A step back is due – anonymisation standards must be laid down first, and only then should the focus shift to the economic gains of the policy.

Cite as (OSCOLA, 5th ed.): Ashwin Arun, 'The India Datasets Programme: Putting the Cart Before the Horse' (*The Innovation Brief*, 12 November 2026) <https://innovationbrief.nludelhi.ac.in/essays/the-india-datasets-programme> accessed 15 November 2026.

Ashwin Arun is a fifth-year student at National Law University Delhi.

The opinions expressed in the Blog are personal to the author/writer. The University does not subscribe to the views expressed in the article / blog and does not take any responsibility for the same.

Published by the Centre for Innovation, Intellectual Property and Competition, National Law University Delhi. Read online at https://innovationbrief.nludelhi.ac.in/essays/the-india-datasets-programme.