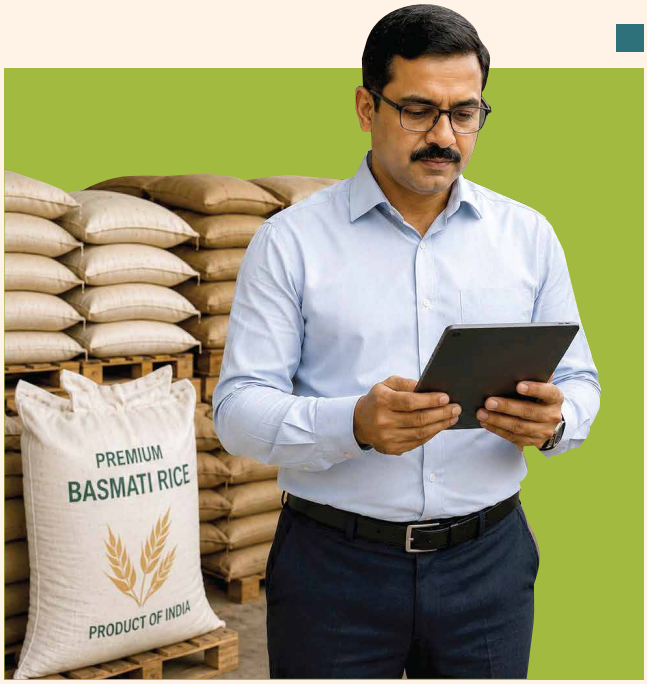
DATA DHARA

EkStep people+ai

DATA DHARA

DATA HARNESSED FOR AI-READY ADVANCEMENTS

A consultative paper exploring India's data power in the AI decade



Copyright & Open Access License

This document has been crafted by People+ai, an initiative of the EkStep Foundation. Individuals and organisations listed under "Inputs Received" provided feedback and consultation during its drafting. Their inclusion in that list reflects their input to the process and does not constitute authorship, co-ownership, or licensing of any content under the terms below.

Suggested attribution and credit: "DATA DHARA," published through People+ai, an initiative of EkStep Foundation.

Table of Contents

Copyright & Open Access License	02
Acknowledgements and Inputs Invited	04
Why Open Data Matters	05
India's Public Data Goldmine	08
Need of Data Harmonisation	12
Assessing and Building AI-Readiness	14
Case Study: Implementation in the state of Delhi	16
Applications Observed Across Sectors	18
What This Means Going Forward	19
Conclusion and Way Forward	21
Open Questions for Ecosystem Consultation	22
References	25

Acknowledgements and Inputs Invited

This consultative paper has been crafted with inputs received from the following individuals and organisations, and further input is sought from a wider ecosystem to explore all dimensions of India's AI readiness with respect to public data and also to address emerging open questions set out in this document:

Mr. P.R. Meshram, Director General (Data Governance), Ministry of Statistics and Programme Implementation (MoSPI)

Dr. Praveen Shukla, Additional Director General, MoSPI

Mr. Rohit Bhardwaj, Additional Director General, MoSPI

Mr. Santosh Vaidya, Principal Secretary (Home), Government of NCT of Delhi

Mr. Premananda Prusty, Director, Directorate of Economics & Statistics

Mr. Rakesh Dubbudu, Factly

Mr. Vinod CV, Mr. Abhishek Purohit, and Dr. Jayaprakash, NanoBI

Mr. Rayulu Villa, Sanketika

The team at Artha Global

Mr. Akhilesh Tilotia and Mr. Karthik Ranganathan, Thurro

Mr. Jagadish Babu, OpenAgriNet

Ms. Keyzom Ngodup Massally, AI Hub for Sustainable Development, Italy

Mr. Arjun Venkatraman and Mr. Suhel Bidani, Gates Foundation

Mr. Randeep Toor, Ms. Amrita Kamat, and Mr. Prem Ramaswami, Data Commons, Google

Mr. Muneender Jillella and Mr. Aditya Chhabra, Kenpath

Mr. Raahil Rai, Anthropic

Dr. Srinath Srinivasa, IIIT-Bangalore

This is a working draft, and we welcome comments from across the ecosystem, including those beyond the institutions listed here, to help shape the version that follows.

Why Open Data Matters

India gathers an extensive range of public data through state governments, sectoral agencies, the **Reserve Bank of India, the Directorate of Economic Surveys, the Ministry of Statistics and Programme Implementation (MoSPI), and other bodies.** This includes granular data on labour markets, household consumption, agricultural production, industrial activity, enterprise demographics, trade flows, health and nutrition outcomes, education access, financial inclusion and more.

Yet this data is often inaccessible, to the departments that hold it, the private sector which builds on it, and the citizens whom it's meant to serve.

It often arrives late, **in incompatible formats, without interoperability, locked away within government departments, making it unavailable to inform policy, AI or accountability**. Public data when held in trust by government departments can only deliver on its social value when it serves the people it was collected from.

This paper argues that making key public datasets AI-ready, granular, fresh and interoperable is one of India's highest-leverage investments. The case for this proposition holds on three fronts.

Government departments could **plan, target, and audit programmes** with connected data instead of contradicting surveys.

Citizens could **get faster, more accurate answers** from the systems meant to serve them.

The private and not-for-profit sector who are already building real use cases with public data could **finally access data they can work with.**

What's missing isn't consumption appetite, it's the supply of ‘usable data’, which can meet this cross-sectoral demand.

This shift is compounded by the **growing usage of AI and automated analytical tools used by governments, citizens, and industry.**

Unlike a human analyst, an AI system **can't infer missing context or work around inconsistent definitions and undocumented changes in structure.** It needs explicit metadata, a stable data model, and a traceable provenance.

Today, consumers of public data have to grapple with;

Contradictory statistics from different departments on the same subject

Beneficiaries missing from one register who are duplicated in another

A systemic inability to view a household across health

Education and welfare systems

Government departments have to bear the time-cost of repeating surveys that better-connected administrative data could have readily answered.

Since the base data already exists, the costs of such fragmentation could be lessened or negated if public data was made AI-ready. However, turning **"existing data" into "usable data" is real institutional work, not a formality.** And this paper seeks to address this transformation by laying out a seven-stage sequence with built-in safeguards for individual agency, rights, and institutional autonomy.

This paper aims to observe rather than present a solution for faultlines across public data systems. Some dilemmas which are presented here remain unresolved, and this paper is better served by documenting them rather than writing around them.

Government data is usually an accurate record of what an institution knows, not necessarily of what is true on the ground.
For example; before 'Jan Aadhaar', Rajasthan's previous 'scheme-wise welfare rolls' accurately reflected their own records. Yet there was no single system which could verify the factor of 'household duplication' across records.

This paper reflects our insights from the field which are still evolving. Our understanding of dilemmas like this one, will keep changing as the work continues. It draws on that fieldwork, but we see this as a journey, where we keep learning along with our peers.

Our initial focus has been on data the government routinely publishes under its open data policy. Many open questions remain, including **India's own evolving policies, debates, and implementation around open data policies and data governance frameworks.**

None of this work starts from a blank slate. Several frameworks are already active and bear directly on the construct of what "AI-ready" should mean. **The National Data Sharing and Accessibility Policy (2012)** has already established an open-license, machine-readable standard for non-personal government data through the Open Government Data Platform. **The Digital Personal Data Protection Act, 2023 and its 2025 Rules** broadly govern personal data and set out a specific standard for how government departments may process it to deliver subsidies, benefits, and services.

This is precisely the terrain of **DATA DHARA's** welfare and vital-statistics examples. **The Ministry of Electronics and Information Technology's National Data Governance Framework Policy** was released for public consultation in May 2022 and, as of this date (still unfinalised) proposes a dedicated national office to standardise metadata, quality, and access rules for the same non-personal, catalogued, machine-readable data that this paper describes.

The Data Empowerment and Protection Architecture, proven in the financial sector through the **Account Aggregator model** and now extending into the health sector through the **Ayushman Bharat Digital Mission**, is the working precedent for sharing an individual's personal data through consent rather than open catalogues. However composing a household-level view from the consent data of multiple individual consented data remains an open design question. **The India AI Governance Guidelines**, released in November 2025, reflect the same national direction; expanding data access to fuel AI adoption, a point that this paper also raises in its seventh readiness stage, 'agent-accessible data', extending to its logical conclusion at the dataset level.

DATA DHARA's role is to potentially give departments and states **a practical, dataset-level sequence for getting ready to participate in whichever of these frameworks ultimately applies to the data in front of them.**

India's Public Data Goldmine

India's public sector data ecosystem is **large by any global comparison- and significantly underleveraged**. Across statistical agencies, administrative systems, regulatory bodies, and line ministries, the Indian state generates a continuous, **multi-layered picture of its economy and population at a scale and granularity** few countries can match.

This ecosystem spans several distinct but complementary layers:

Statistical data

MoSPI's survey apparatus covers labour markets (PLFS), household consumption (HCES), industrial activity (ASI), and enterprise demographics (Economic Census). These form the country's macro-measurement backbone.

Administrative data

Generated as a byproduct of governance at transaction level: GSTN on firm revenues and supply chains, EPFO and ESIC on formal employment and wages, PM-JAY on health claims, MGNREGA and PDS on rural livelihoods, VAHAN on vehicle registrations, and land records systems on property ownership and transfers.

Regulatory data

The Reserve Bank of India on credit flows and banking penetration, SEBI on capital markets, TRAI on telecom usage, and IRDAI on insurance coverage - each holding sectoral intelligence unavailable elsewhere.

Geospatial and sectoral data

ISRO's earth observation archives, the Agriculture Ministry's mandi prices and crop production estimates, UDISE+ on every school in the country, and the health ministry's HMIS on facility-level utilisation.

Viewed together, these layers form a **continuously updated picture of how India works**; from firm to household, from district to sector, from transaction to outcome.

The core problem is not that this data does not exist. It is that it is either **restricted, stale, or exists in diverse formats; PDFs, password-protected portals, annual print releases, and decade-old Excel files** that make it functionally useless for AI applications, real-time product development, and enterprise decision-making.

The Freshness Problem

By the time data reaches its recipient or user to whom it matters, it reflects a country that no longer exists.

> The Periodic Labour Force Survey (“PLFS”) Annual Report for FY 2023-24 was released in mid-2025.

> The Economic Census of 2013 remained the canonical reference for enterprise demographics until 2024.

> The Household Consumption Expenditure Survey (“HCES”) that should have succeeded the 2011-12 round was finally published in 2024, covering 2022-23,an eleven-year gap.

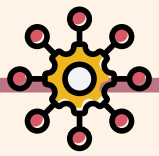
Freshness is not merely a statistical nicety. For a staffing company sizing a talent pool in Tiruppur, a six-month lag is the difference between accurate workforce planning and costly misallocation. For a cross-border payments platform assessing SME export potential in Surat, a two-year lag is the difference between timely credit and missed revenue.

The Granularity Problem

Much of the published data aggregates to **the state level or, at best, to urban/rural splits within states.** India's economic geography does not work at state level. Coimbatore, Chennai, and Dindigul are in the same state but represent entirely different labour markets, industrial ecosystems, and consumption profiles.

The data architecture does not reflect this reality, which means companies either operate with blunt instruments or invest in expensive primary research to compensate.

Birth and Death registrations trends are published at the state level.



The Interoperability Problem

Even where data exists, it is **fragmented across systems** that already assign a business its own identifier ; a GST number, an MCA registration number, an Udyam registration number but does not connect the three.
A business such as 'Sunrise Traders' is recorded under three different numbers across three different databases, with no reliable way to confirm that all three refer to the same entity.

Building composite signals from this data currently requires laborious manual matching, an opportunity cost quietly borne by every company, bank, or agency that attempts it.

The fix is not a new identifier but a reliable crosswalk between existing identifiers to registered business entities, not individuals.

The Accessibility Problem

Many of India's most valuable **government datasets remain inaccessible** even though granular aggregates exist within official systems: firm-level GST summaries, DGCI&S trade records, MCA corporate databases, RBI credit data, and Railways freight movement data.

Some restrictions are legitimate: sovereign or security-sensitive data has a reasonable case for protection.

Released in anonymised aggregated formats, they could power major use cases in MSME finance, supply-chain intelligence, logistics, and trade analytics - markets with strong willingness to pay and no supply to meet them.

The Format Problem

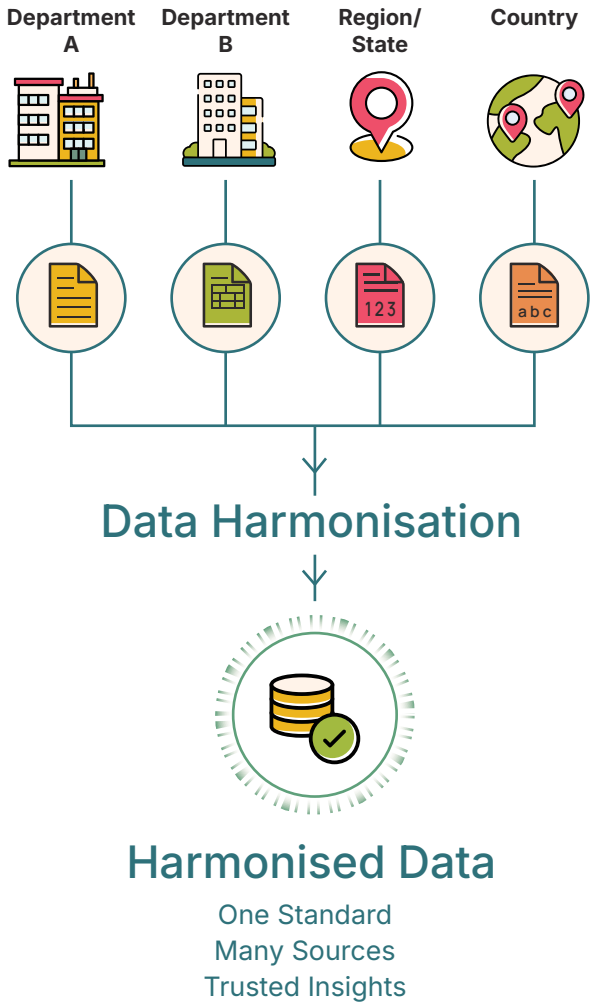
Even where data is nominally public, it is rarely usable. **Harmonized System (HS) codes used in every export process mapping are stored in PDFs**, policies released are not versioned properly, survey results are released as PDF reports - often image-rendered, with tables that cannot be extracted programmatically.

Where Excel files exist, they carry inconsistent schemas, merged cells, and footnotes embedded in data rows that break on import. For an AI pipeline, a PDF is not data, it is a picture of data.

Every company that attempts to use these releases must invest in bespoke parsing logic that breaks with each new round. The cost of that re-engineering falls on every team that tries and stops every team that cannot afford to.

Need of Data Harmonisation

Data harmonisation is the process of ensuring that data collected from different sources, departments, regions, or countries are **comparable, consistent, coherent, and compatible with international standards.**



Harmonisation matters because it **improves data credibility, facilitates policy coordination, and reduces duplication and inconsistencies across institutions.**

Two Kinds of Data, Two Readiness Paths

Not all public sector data carries the same stakes, and the AI-readiness sequence in the section below should not be applied uniformly to all of it. A district-level 'mandi' price and a household's welfare enrollment record are both 'public sector data,' but making them AI-ready means different things.

Non-personal and aggregate data

Mandi prices, weather, macro survey statistics (PLFS, ASI), firm-level identifiers under GST/MCA can move through the full seven-stage sequence toward open, machine-readable, catalogued, and agent-accessible form.

This is where most of the use cases in 'Section 3' already sit, and readiness here is primarily a technical and organisational task.

Personal and household data

Vital statistics, health records, welfare and scheme enrollment, education records tied to an individual carry a different requirement. Readiness for this category cannot mean open catalogue or bulk API exposure. It means that the data remains with its original custodian and moves only through a consented, purpose-specific, revocable exchange.

This is the model India has already built and proven through the Data Empowerment and Protection Architecture (DEPA) and the Account Aggregator framework in the finance sector, which is now being extended to health.

The practical implication: **The seven-stage sequence** in the next 'Section' describes readiness for the first category. For the second, 'readiness' should be defined as consent-artefact compatibility. The question is whether this dataset can be exposed through a DEPA-style consent flow, with a data principal or their authorised representative in the loop, for each specific use, rather than discoverability and open API access. Section 5 sets out these open questions in more detail.

Why Harmonisation Matters

It Ensures

Consistency
Stable Data Models

Context
Explicit Metadata

Comparability
Machine Readable Provenace

Comparability
With International Standards

It Improvements

Advanced Analytics Across Departments

Data Credibility

Facilitates Policy Coordination

Reduces Duplication

Inconsistencies across Ministries

Assessing and Building AI-Readiness

Defining AI Readiness:
AI Readiness can be interpreted in many ways. Here is a sample process where inputs from the ecosystem are sought. At a high level, the progression looks like this:

AI Readiness Pathway

1. **Administrative Data**
   Data Collected through government operations & services
2. **Harmonisation**
   Standardized formats, definitions, codes and classifications
3. **Interoperability**
   Enable linkage across systems and datasets using common standards
4. **AI Readiness**
   Ensure data is complete, consistent, reliable & fit for AI use
5. **Data Reuse**
   Make Data accessible for multiple use cases and Stakeholders
6. **Advanced Analytics**
   Derive insights and support evidence-based policy and decision

A pathway from administrative data to advanced analytics.

AI-readiness can be assessed at two levels: **the individual dataset, and the department or organisation that is a custodian of it.** Most of the practical work happens at the dataset level, because that is where a team can act directly. Knowing where each dataset stands also gives a fair picture of where the organisation as a whole stands - readiness at the dataset level adds up to readiness at the organisation level.

Four questions determine whether a dataset is ready for AI use:

Can it be found? A dataset is discoverable when it appears in a catalogue with a clear description, rather than requiring someone to already know it exists and ask around for it.

Can a machine read it? A dataset is machine-readable when it is available in a structured format a program can process directly, not only as a PDF, scanned form, or static report.

Can it be interoperable? A dataset is interoperable when it uses identifiers, codes, and definitions consistent with other datasets, so records can be matched and combined across sources.

Can it be trusted? A dataset is reliable when its origin, quality, and version history are documented well enough for someone to judge whether it is fit for their purpose.

These four questions can be applied plainly to public sector data, where the starting point is usually messier: several formats, inconsistent codes, no shared identifiers, and little documentation to begin with.

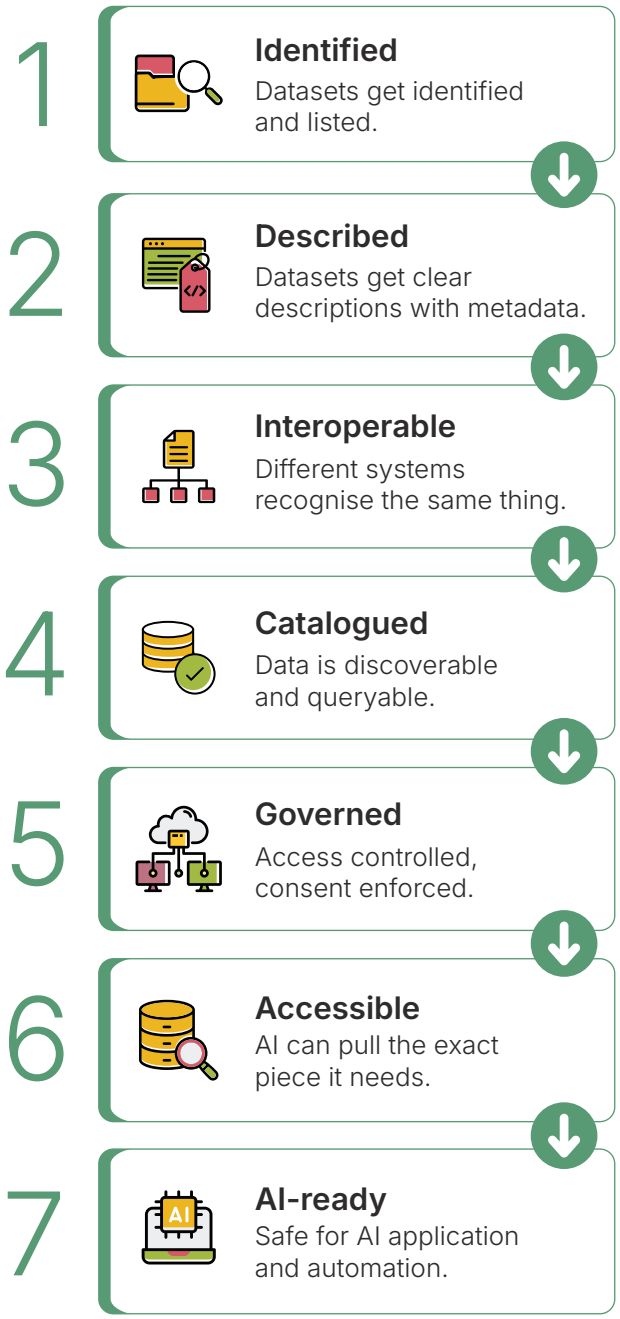
A Seven-Stage Data Lifecycle:

Regardless of sectors or states, turning any dataset into an AI ready asset, entails progression through a consistent set of stages.

A seven stage life-cycle is set out here because it appears to hold across different contexts, not because it is the only way to organise this work.

Seven-Stage Data Lifecycle

From raw datasets to AI- Ready, trusted Data Assets



Few datasets begin without any prior work. Most departments already maintain some metadata, or already publish certain datasets. What this life-cycle is useful for is diagnosis: it tells a team not only how ready a dataset is overall, but exactly where to focus next.

Case Study: Implementation in the state of Delhi

Before

A readiness effort for 2024 Birth and Death Registration dataset was undertaken with the Government of NCT of Delhi (Directorate of Economics and Statistics). The data was already present as tables in documents published by the department in the form of pdf reports on their website. The data was extracted from these publicly available pdfs to build public datasets. Example: Aggregate data was present as: Date of 'number of births' in the 'Municipal corporation of Delhi', by women, 18-25 years. There was however, no PII and names, parentage, addresses, were not mentioned. The department published this data for policy and decision making, but routinely found that LLMs gave wrong answers to questions like, ' How many girl children were born in the month of July 2025 in Delhi?'.

The LLM may or many not have been able to access the answer which was present in a pdf on the website.

After

The above birth and death datasets were categorized into adequate data product groups. **Unique identifiers** to each dataset was assigned, NMDS 2.0 (National Metadata Structure) compliant meta data was captured, data was harmonized by capturing classification codes, and downloadable dataset and meta-data files in machine-readable structured format was made available in a 'Data Catalogue' with a search enabled browser.

Outcome achieved could now be demonstrated by LLMs providing more specific and reliable answers to questions on vital statistics.

The same datasets are now reliable, discoverable, interoperable and machine-readable.

DES Workflow

Directorate of Economics and Statistics, Govt of NCT of Delhi
Workflow for AI readiness of Delhi Vital Statistics datasets.

1 Idenitified

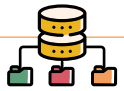
5 Data Product Groups organized, Unique Dataset Ids assigned

Data Product Groups	# of Datasets
Still Birth Data	2
Infant & Mother Mortality Data	10
Medical Cause of Death Data	2
Death Rates Data	12
Live Birth Rates Data	60

2 Described

- NMDS 2.0 compliant Meta-data captured.
- Classification groups derived and codes captured.

3 Interoperable



- Harmonised data values
- Tabular data structure in Staging with metadata

4 Catalogued

Search enabled Catalogue browser available. Datasets & meta-data are now discoverable

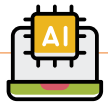
5 Governed

Machine-readable format for Dataset and Meta-data staged in an access-controlled secure environment.

6 Accessible

API end points available. Datasets and associated meta-data are accessible programmatically

7 AI-ready



MCP server available. Auto-discovery and secure access by AI agents and LLMs enabled via MCP.

Applications Observed Across Sectors

Investment, Uttar Pradesh:InvestUP [2,3]

Investor-facing information; policy documents, notifications, and procedural guidelines are spread across sources, which makes it difficult for investors to obtain clear, timely answers. Structured and connected, this information could power conversational tools, multilingual guidance, and calculators that give investors direct answers, rather than documents to search through.

Agriculture, Maharashtra: MahaVISTAAR [4]

The information farmers need already exists across weather, soil, crop, pest, and market systems, but each source gives only a partial picture on its own. Linked through shared fields such as district, crop, and season, this data could generate real-time, practical advice on sowing, nutrients, disease risk, and market timing instead of static bulletins.

Trade and fintech: Exporter Verification

For a cross-border payments company, information on small exporters is spread across company, tax, trade, and payments systems, making it difficult to verify businesses beyond the largest, well-known clusters. Better-connected public sector data on business identity, trade activity, and tax records would reduce onboarding friction and let fintechs serve smaller exporters with more confidence.

[2] Press Release Invest UP | [3] Nivesh Suvidha | [4] The Hindu article on Mahavistar

What This Means Going Forward

Data readiness unlocks diverse use cases

The examples in this report span different sectors, states, and levels of government, but the underlying pattern repeats itself. Once a dataset can be found, read by machines, connected by a shared identifier, and trusted enough to act on, new applications become possible that no one has had to build individually, from scratch, for each use case.

Opportunity for Data & AI Sandboxes

An 'AI Sandbox' is a governed environment where harmonised public sector datasets are made accessible to approved external builders, industry, startups, researchers, and other government departments — to develop and test AI use cases against real public sector data. It draws on a curated selection of AI-ready datasets from a state's or department's catalogue, available in structured formats through secure API access, and is intended to turn the readiness work already underway into visible, testable applications rather than a static catalogue.

Access is governed rather than open. Organisations wishing to build on this data apply for entry, and approved builders receive sandbox credentials and defined usage limits, scoped to the tier of data access described.

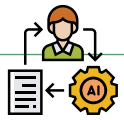
Priority use case areas include welfare and scheme targeting, health and vital statistics, urban planning, climate and air quality, and employment and MSME formalisation. For urban planning, climate, and MSME/employment data — largely non-personal or at a firm-level, the 'Sandbox' can offer direct, credentialed access to structured datasets. For welfare scheme targeting, and health and vital statistics, 'Sandbox' access is limited to aggregate or de-identified statistical outputs, unless a specific consented data flow, of the kind described in this document, has been established for the use case in question. This distinction is stated explicitly in the 'Sandbox' onboarding documentation, so that builders know from the outset which tier of data they are working with and what obligations are attached to each.

DATA DHARA Toolkit

We request inputs from the ecosystem on the 'DATA DHARA Toolkit' as an operational package built towards achieving AI readiness of datasets. It can potentially guide a data custodian through the steps from building an inventory, capturing metadata against National Metadata Structure 2.0 (NMDS 2.0) and Metadata and Data Standards (MDDS), harmonising classifications, publishing a catalogue, to exposing the dataset through API and MCP endpoints. Measuring readiness and improving it are part of one workflow, so that a department always knows both; where a dataset stands in terms of AI readiness, and what to do next. This can become an open-source collection of utilities, libraries and the configuration kit.

Enable an Ecosystem of Builders

Making data AI-ready creates supply, and the 'Sandbox' creates access. What neither does on its own is create demand that requires programs that actively bring builders in, rather than waiting for them to arrive. A state or department can run four such programs.



With industry and startups:
'Challenge programs' and innovation calls inviting companies to compete on specific governance problems using AI-ready data.

With academia, partnerships with local universities and research institutions:
Supporting policy research, use-case evaluation, and published findings.

With other government departments:
Cross-departmental working groups identify shared data problems and build linked-data use cases across health, welfare, and urban systems.

With civil society and media:
Open access to non-sensitive datasets supporting independent journalism, watchdog work, and public-interest analysis.

Conclusion and Way Forward

India faces not a shortage of data, but a structural gap between the data it holds and the insights that data could generate. The question is no longer where the data is; it is how to make it usable, connectable, and trustworthy.

The way forward is **incremental & additive:** dataset by dataset, department by department, State by State.

Each dataset improved today generates value for every application built on it tomorrow, and each standard adopted in one domain reduces friction in the next. Executed with discipline, this shared effort has the potential to make India's data architecture a stronger foundation for precise programme targeting, efficient public resource management, and outcome-driven governance: in service of Viksit Bharat@2047.

Open Questions for Ecosystem Consultation

The sections above set out a case and a sequence. Several dimensions of that work are still open, and this paper is better served by naming them directly than by resolving them internally. This section lists those dimensions as specific questions, organised by theme, and identifies the parts of the ecosystem best placed to weigh in on each.

Responses can take any form - a written note, a marked-up section of this paper, or a conversation. Where a question maps to work already underway elsewhere (DEPA, the DPDP Act and Rules, sectoral consent-manager frameworks), we would rather point to and align with that work than duplicate it, and we are relying on this consultation to tell us where those connections should be made.

Personal Data and Consent

Section 1 distinguishes non-personal from personal and household data, and proposes that the latter be shared only through a consent-mediated flow. We don't yet have a settled position on where the line sits or how such a flow should be operationalised for public sector data.

Should household-level data across health, education, and welfare systems be pursued through the same catalogue- and - API model as non-personal data, or should it require a consent - mediated flow of the kind DEPA and the Account Aggregator framework already provide?

Where should the line sit between the two categories - is it about the sector (health vs. agriculture), the identifiability of the record (aggregate vs. individual), or something else?

For datasets already published as open catalogues (e.g. the Delhi birth and death registers), what level of granularity is appropriate for the ecosystem to find benefit if they want to use it?

What would it take for a state or department's personal-data assets to be made AI-ready in a way that is compatible with, rather than parallel to, the consent architecture already live in finance and emerging in health?

Regulatory Alignment

The Digital Personal Data Protection Act, 2023 and its 2025 Rules are now in force in a phased manner, with government processing standards specified separately from private-sector obligations. This paper is written to sit inside that framework, and we want the ecosystem's help in identifying potential gaps we must consider and fill.

Which parts of the readiness sequence (metadata, structure transformation, dissemination, API access) **engage DPDP obligations**, and at what point in that sequence should a compliance checkpoint sit?

For datasets that fall under the **Act's government-processing provisions,** what does 'AI-ready' need to additionally guarantee — Data Principal intimation, purpose documentation, retention limits?

Should the **DATA DHARA Toolkit include a DPDPA compliance checklist as a standard component**, and if so, who is best placed to define it?

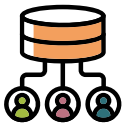
AI Sandbox: Access Tiers and Governance

Section 4B proposes different access tiers for personal versus non-personal data in the 'Sandbox'. This is a starting position, not a settled design, and we want the ecosystem's view before any sandbox is piloted.

For welfare, health, and vital statistics use cases, should builder access be **limited to aggregate or de-identified outputs by default**, with individual-level access requiring a separate, purpose-specific authorisation?

Who should **set and audit** those access tiers of the 'Sandbox' — the originating department, a central sandbox authority, or a third-party auditor?

What obligations should approved builders carry once they have used sandbox data — **publication of methodology, restrictions on re-identification attempts, data deletion timelines**?



Institutional Coordination

This paper proposes new institutional artefacts — a Sandbox model, a **DATA DHARA** Toolkit — in a space where DEPA, the Account Aggregator framework, and ongoing MeitY/IndiaAI initiatives already operate. We want to be additive to that landscape, not duplicative of it.

Who should own the 'AI Sandbox' within the state? A new entity, or state the existing entity.

What governance body, if any, should have visibility, so a dataset doesn't end up readable through one path but restricted through the other?

Where does **DATA DHARA's** readiness sequence and proposed Toolkit sit relative to the National Data Governance Framework Policy and the India Data Management Office (IDMO) once that policy is finalised — should DATA DHARA's state- and department-level work feed into IDMO's national India Datasets programme, operate as a parallel track for personal-adjacent data IDMO doesn't cover, or something else?

Public Trust and Redress

None of the mechanisms above resolve what happens when something goes wrong — a dataset is misused, a household is misidentified, an inference is drawn that a citizen disputes. This paper does not yet have an answer to that, and we think it needs inputs here.

What redress mechanism should exist for a citizen who believes their data was used inappropriately?

How should trust in this system be measured or reported over time — audit logs, periodic public reporting, something else?

How to Respond

We are not looking for consensus on all of this before **DATA DHARA** moves forward. Some of these questions will have different answers for different sectors and states, and that is a fine outcome. What we need is enough input on each dimension to take a version of this paper forward that names its own boundaries clearly, rather than leaving them implicit. Comments, written responses, or a short conversation with the People+ai team are all useful; whichever is easiest is the right one.

References

1 Ministry of Statistics and Programme Implementation (MoSPI), Data Informatics & Innovation Division (DIID). Data Harmonisation Handbook. [No public URL available at time of writing; cited as an institutional source, consistent with its acknowledgement above.]

2 Data Informatics & Innovation Division (DIID), Ministry of Statistics and Programme Implementation, Government of India. https://www.mospi.gov.in/data-informatics-innovation-division-diid

3 NITI Aayog. Data Empowerment and Protection Architecture (DEPA) — Executive Summary (Draft for Discussion), August 2020. https://www.niti.gov.in/sites/default/files/2020-09/DEPA-Executive-Summary.pdf

4 Account Aggregator (AA) Framework, Department of Financial Services, Ministry of Finance, Government of India. https://financialservices.gov.in/beta/en/account-aggregator-framework — operational framework maintained by Sahamati, the RBI-recognised Self-Regulatory Organisation for the AA ecosystem: https://sahamati.org.in/

5 The Digital Personal Data Protection Act, 2023, Ministry of Electronics and Information Technology, Government of India. https://www.meity.gov.in/content/digital-personal-data-protection-act-2023 · Digital Personal Data Protection Rules, 2025. https://www.meity.gov.in/documents/act-and-policies/digital-personal-data-protection-rules-2025-gDOxUjMtQWa

EkStep people+ai