DATA DHARA

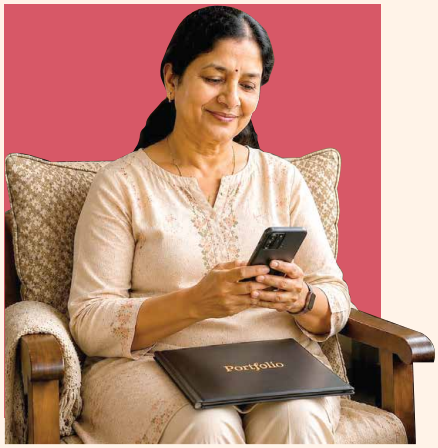
EkStep people+ai

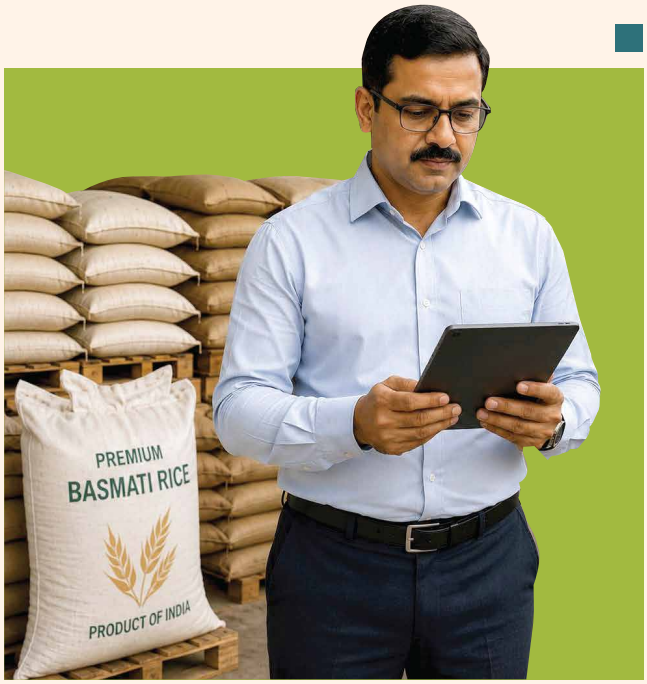


# DATA DHARA

## DATA HARNESSED FOR AI-READY ADVANCEMENTS

A consultative paper exploring India's data power in the AI decade (Executive Summary)



# Why Open Data Matters

India generates one of the world's richest collections of public sector data across sectors such as labour, agriculture, health, education, and finance. Yet much of it remains fragmented, delayed, inconsistently structured, insufficiently granular and difficult to connect across systems, limiting its value not only for its citizens, researchers and businesses but also for the government itself. Since this is ultimately the people's data, held in trust by public institutions, its true value lies not just in being collected, but in being usable to improve governance, policymaking, research, innovative use cases and  better serve the people it represents.

# India's Public Data Goldmine

India's public data ecosystem spans-

***Statistical Datasets*** ***Administrative Datasets*** ***Regulatory Datasets*** ***Geospatial Datasets***

These datasets provide a continuously evolving picture of the country's economy, enterprises, households, and public services. The challenge is not the absence of data, but its usability.

# Two Kinds of Data, Two Readiness Paths

Not all public sector data should follow the same AI-readiness pathway.

CAN BE SOVEREIGN

**Non-personal and aggregate data** can be standardised, catalogued, and made available for AI applications.

**Our work focuses on this category of data**

CAN BE SOVEREIGN

**Personal and household data** requires a different approach for accessibility and AI readiness. AI readiness needs to be designed for consent, secure access and appropriate ownership.

Sovereignty - where data resides and who controls it - is a feature that can apply to either data type above. It is not a third kind of data.

This distinction enables wider data use while safeguarding privacy, individual agency, and institutional trust.

## Five structural gaps constrain the value of public data:

Freshness

Granularity

Interoperability

Accessibility

Format

This results in data that is often delayed, insufficiently detailed, siloed, unavailable, or difficult for developers and AI systems to use.

Unlocking public data is less about creating new datasets and more about making existing ones “ready”- timely, interoperable, and accessible to both humans and machines in a trusted manner. Our months-long primary and secondary research suggests that the first step is data harmonisation: making information across departments comparable, consistent, and interoperable to improve quality, reduce duplication, strengthen coordination, and enable reuse across multiple purposes.

# Assessing and Building AI Readiness

AI readiness begins with individual datasets, as organisational readiness is ultimately built dataset by dataset. Four questions determine whether a dataset is AI-ready:

*Can it be found?*

*Can it be read?*

*Can it interoperate?*

*Can it be trusted?*

Together, these provide a simple framework for assessing whether data is discoverable, machine-readable, connected through common standards, and supported by sufficient metadata and provenance.

# DATA DHARA - 7 Stage Data Lifecycle

**Identified**
Datasets get identified and listed.

**Described**
Datasets get clear descriptions with metadata.

**Interoperable**
Different systems recognise the same thing.

**Catalogued**
Data is discoverable and queryable.

**Governed**
Access controlled, consent enforced.

**Accessible**
AI can pull the exact piece it needs.

**AI-Ready**
Safe for AI application and automation.

***From Raw Datasets to AI-Ready, Trusted Data Assets***

*Rather than prescribing a rigid process, this lifecycle provides a practical roadmap for assessing a dataset's current state and identifying the next steps needed to support AI, evidence-based policymaking, and digital public services.*

# From AI Readiness to Public Value

## *Case Study and Emerging Applications*

In Delhi, a pilot exercise with the Government of NCT of Delhi (Directorate of Economics and Statistics) demonstrated how publicly available **Vital Statistics** can be transformed into AI-ready datasets. PDF reports were converted into structured datasets with standard metadata, unique identifiers, and a searchable catalogue, enabling AI systems to answer questions on vital statistics more accurately and reliably.

In Uttar Pradesh, **InvestUP's Nivesh Suvidha** provides investor facilitation and connects data across 26 departments and 75 districts.

In MahaVISTAAR (Maharashtra), comprehensive crop advisory linking weather, soil, crop, pest, and market data can generate real-time, location-specific advisory services for farmers.

# The Way Forward

## *AI Sandboxes and the DATA DHARA Toolkit*

There is a need for **State AI Sandboxes** that can provide secure environments where government, industry, startups, and researchers develop applications using AI-ready public datasets.

Complementing this, the proposed **DATA DHARA Toolkit** offers departments a practical pathway to improve dataset readiness through inventory creation, metadata standards, harmonisation, cataloguing, and API enablement.

# Conclusion

Making public data AI-ready is a foundational investment in better governance, innovation, and public service delivery. Done incrementally, it can unlock the full value of India's public data ecosystem in support of **Viksit Bharat@2047.**

Scan the QR code to access the full page

DATA DHARA

info@ekstep.org