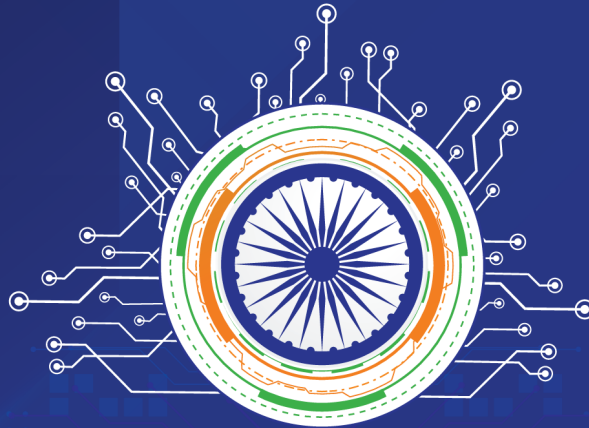**NITI Aayog**

NITI Frontier Tech Hub

FUTURE FRONT
QUARTERLY FRONTIER TECH INSIGHTS

India's Data Imperative
The Pivot Towards Quality

Issue-3 I June 2025

India's Data Imperative:
The Pivot Towards Quality

How high-quality data can unlock better outcomes, build public trust, and power the next decade of digital governance

Knowledge Partner:

Gramener
A Straive Company

Foreword

India's Data Imperative: The Pivot Towards Quality

I am delighted to present this edition of the NITI Aayog Frontier Technology Hub's Quarterly Insight on "India's Data Imperative: The Pivot Towards Quality". This timely report underscores a crucial shift in India's data journey – from achieving unprecedented scale to ensuring uncompromising precision and trustworthiness in our data.

Over the past decade, India has truly distinguished itself as a global leader in digital public infrastructure. The Unified Payments Interface (UPI) moves trillions of rupees monthly, Aadhaar has authenticated billions of identities, and Ayushman Bharat has extended healthcare access to more than 350 million citizens. These are not merely statistics; they are testaments to a foundational transformation, enabling financial inclusion, empowering citizens, and streamlining welfare delivery on a scale previously unimaginable. Our collective efforts have laid robust digital rails, demonstrating to the world how technology can empower a nation of 1.4 billion.

However, as we embark on the next phase of our digital evolution, the focus must decisively pivot towards the quality of the data that fuels these incredible platforms. As this report highlights, data quality is no longer a peripheral technical consideration but a fundamental prerequisite for good governance and sustained public trust. When a single erroneous digit can delay a benefit, or a duplicate record inflates welfare outlays, the true cost of poor data becomes painfully apparent – impacting budgets, distorting policy, and eroding the faith citizens place in our digital systems. The move from "scale to precision" is not just an aspiration; it is a national imperative.

I particularly commend the inclusion of two practical, easy-to-use tools in the report: The **Data-Quality Scorecard** and the **Data-Quality Maturity Framework** that are designed to enable every department to conveniently assess their current data landscape, identify gaps, and chart a clear roadmap for improvement. Whether it is implementing real-time validation at the source, assigning data stewards, or linking grievance redressal to backend corrections, the "*Starter Kit*" outlined in this report offers tangible pathways to achieve quick, verifiable wins.

This report serves as an invaluable guide for all government Ministries

and Departments. It meticulously dissects the challenges across the entire data value chain – from generation and storage to sharing, use, and retirement. I urge the Ministries and Departments across the Government of India and the State Governments to thoroughly engage with this report. The insights and recommendations in the report are aimed at embedding a culture of data excellence – one where quality is not an afterthought but a shared responsibility, deeply ingrained in our institutional roles, incentivized effectively, and supported by robust interoperability.

Our digital future will be defined not just by the platforms we build, but by the unwavering trust we cultivate in them. This trust begins with data that is precise, reliable, and truly ready to serve. Let us champion clean, usable data as the foundation of good governance, ensuring that the dividends of our digital journey reach every citizen with precision and equity.

DR. SAURABH GARG
Secretary, MoSPI

As India matures as a Digital Economy, we must increasingly prioritize a critical imperative: data quality. The shift from scale to precision makes high-quality data not just desirable, but essential. Even a single incorrect digit can halt a pension, deny healthcare, or misdirect subsidies, profoundly eroding citizen trust and costing billions in fiscal leakage.

Data quality must therefore become a frontline governance imperative, integral to building Digital India on a robust foundation of trust. Trust is not merely a desirable outcome—it is the bedrock upon which our digital ambitions stand and scale.

This edition of NITI Frontier Tech Hub's Quarterly Insights, Future Front, developed with our knowledge partner Gramener, emphasizes the urgent need to radically transform our approach to data quality. Strengthening data quality is foundational to realizing our digital ambitions—whether enabling AI-driven governance, ensuring targeted welfare delivery, or driving cross-sector innovation. It requires collective action, clear accountability, and a shared commitment from every department and ministry.

DEBJANI GHOSH
Distinguished Fellow, NITI Aayog;
Chief Architect, NITI Frontier Tech Hub

EXECUTIVE SUMMARY:

India's digital public infrastructure has grown at a breakneck speed. UPI now moves trillions of rupees each month, Aadhaar verifies billions of identities, and Ayushman Bharat touches half a billion citizens. As data volumes explode, it is imperative for data quality to keep pace across both the torrents generated today and the decades of legacy records still in use. Records need to be precise, complete, and up-to-date by reducing or eliminating gaps, duplicates, or timing lags that may exist today. When one wrong digit can freeze a pension or misroute a subsidy, quality is no longer a technical afterthought but a frontline service obligation. India's digital governance will only be as strong as the quality of the data that powers it. As India moves from scale to precision, data quality must become a national priority—on par with infrastructure and platforms.

Why Data Quality Matters:

1. Fiscal leakage, erroneous or duplicate beneficiary records drain budgets, inflating welfare outlays by an estimated 4–7 per cent annually.

2. Policy blind spots, inconsistent, non-standardised datasets distort evidence, leading to mis-targeted schemes or delayed course-corrections.

3. Public trust erosion, citizens confronting mismatched records, rejected claims, or endless rectification queues lose confidence in digital governance.

This edition of the NITI Frontier Tech Hub Quarterly Insight, created in partnership with Gramener, lays out data-quality challenges across the full data value chain, analyses their root causes, and offers two easy-to-use tools to measure data quality and maturity in any department.

1. SETTING THE STAGE: INDIA'S DIGITAL & DATA MOMENT

India's digital public infrastructure now operates at a scale few imagined a decade ago. In April 2025, the Unified Payments Interface (UPI) processed 17.89 billion transactions worth ₹23.9trillion,[1] rivalling the monthly GDP of several mid-sized economies. Aadhaar authenticated 27.07 billion identity requests in FY 2024-25,[2] confirming its place as the common key for banking, welfare and a widening array of private-sector services. On the social-protection front, more than 369 million Ayushman cards are now in circulation,[3] bringing cashless hospital care to roughly half a billion citizens.

But the next test is precision, not reach. What if just 5 % of records are flawed, and those belong to our most vulnerable citizens? The true test of digital infrastructure is how precisely it represents each individual, not just how many people it onboards. As platforms mature, the focus must shift from expansion to fidelity. Digital highways only create public value when every record is accurate, complete, and current. And no stack of audits can rescue accuracy if the people who create or rely on a record have no sight or stake in it. Confidence endures only when the people who depend on a record can also see and correct it.

1.1 Small errors carry huge costs

A single wrong digit during Aadhaar enrolment can block a pension. A duplicated mobile number can stall a hospital claim. A misspelt name in a land record can delay compensation after a natural disaster.

Multiply such slips across several central schemes and thousands of state programmes and the fiscal dent is staggering. Every wasted rupee crowds out an honest claim; every wrong dashboard alert delays a life-saving decision. Bad data costs real lives and real money.

1.2 From scale to fidelity

Over the last decade India laid the rails, identity, payments, registries, that make inclusive innovation possible. The coming decade's services, farm-credit via account aggregators, portable skill credentials, AI-powered health alerts, depend on the integrity of the data flowing over those rails. Where records are rigorous at the moment of capture, new services can scale at marginal cost; where they are inconsistent, each added layer inherits and amplifies a structural deficit.

Recent clean-ups underline the pay-off

Action	Estimated Savings
Deleted 17.1 m ineligible PM-Kisan names[4]	₹90 billion (FY 2024)
Weed out 35 m bogus LPG connections[5]	₹210 billion (in two years)
Drop 16 m fake ration cards[6]	~₹100 billion per year

1.3 What holds us back

Across ministries and states, platforms are built as separate silos that store data in clashing formats and use incompatible identifiers, forcing even routine joins to be stitched together by hand. Many of these systems run on ageing back-ends that lack basic validation rules, audit trails or version control, so a small tweak in one module can topple an entire workflow. Because records shuttle between teams without a named custodian, obvious errors linger unchecked, and field programmes are often driven by enrolment targets that prize speed over accuracy. Over time this breeds a "good-enough" culture, once 80 percent accuracy is accepted as normal, the system quietly stops trying to prevent error and the data quality debt compounds.

1.4 Towards "data confidence"

Building confidence begins at the point of capture: automated checks and standard pick-lists must stop bad entries before they enter the system. Every high-value dataset then needs a clearly identified steward—someone empowered, and accountable, to correct and continually improve the records. Finally, common schemas for people, places and programmes must be adopted so that data can travel seamlessly across agencies without constant translation. Together, these three practices turn today's fragmented information into a reliable asset that analytics, AI and service delivery can use without sacrificing trust.

The pivot from scale to fidelity demands a new scorecard—one that values precision and trust as highly as reach.

The table below shows how priorities must shift, toward iterative architecture, stronger interoperability rails, and shared stewardship roles for AI-ready data.

Scale Era	Fidelity Era
Number of users onboarded	Precision in every record
Platform speed and uptime	Trust in data at every stage
Build core digital public goods, identity, payments, registries	Iterate architecture, strengthen interoperability rails, and embed shared stewardship roles

Without quality, data is just noise.

India has proved it can scale digital public goods. Showing the world we can sustain high-fidelity data will cement India Stack as a governance model, not just a technology export,

and ensure the first dividends reach those who need them most. Getting there will demand as much organisational reform as technology: frontline training, budget lines for data stewardship, and lightweight policy guardrails that make quality a default requirement rather than a discretionary upgrade.

2. SEEING THE WHOLE ELEPHANT: THE PUBLIC SECTOR DATA VALUE CHAIN

A farmer's subsidy application, a child's vaccination record, or a woman's property title all start as a few keystrokes or biometric scans. Those fragments soon underpin service delivery and policy. Between capture and a final decision, data travels an often-hidden chain - stored, transmitted, transformed, analysed, and eventually archived. Quality can slip at any link, with losses that snowball for citizens and the agencies that rely on the record.

These five stages of the data value chain – generation, storage, sharing, use, and retirement – present distinct challenges and play a role in shaping data quality. We'll illustrate each stage with Lakshmi Devi's PM-Kisan record, a single row that travels the chain.

2.1. Generation - Where It Begins

Most data-quality faults are born at the moment of capture, whether a field agent taps details into an enrolment app, a clerk uploads a survey sheet, or a sensor pushes logs to the cloud. In resource-tight settings a single typo in a name, Aadhaar, or bank IFSC can ricochet through every downstream system, blocking payments, skewing analytics, and triggering costly reconciliations. Fixing them later often requires time-consuming manual reconciliations and frustrated citizens.

Example - At the village service centre, Lakshmi Devi's PM-Kisan details were entered with two digits of her bank IFSC transposed. Because the enrolment screen showed only a confirmation tick, neither she nor the officer noticed the slip.

2.2. Storage: The Quiet Fade of Confidence

Once generated, data must be stored in a way that preserves its structure, context, and integrity. This includes defining how data is formatted, tagged, and secured; how it is updated over time; and how different departments manage their own databases. In practice, storage practices are fragmented. Some systems now employ cloud-native, version-controlled stores, while others still depend on legacy databases with no audit trail.

Example - When the day's data synchronises, Lakshmi's record is copied from the enrolment app to the state PM-Kisan server and then to a legacy backup database. Because neither system runs cross-check or version control on incoming fields, the transposed IFSC is stored unchanged in both locations, quietly duplicating the error.

2.3. Sharing: When Silos Collide

Sharing data across ministries, departments, or platforms is where digital governance can truly shine and where friction most often flares. Crucially, inconsistency is not only horizontal (one ministry versus another) but also vertical (different schemes, dashboards, or district offices inside the same programme). A pension portal may store household size at the family level, while a village survey logs each individual; a health-worker app records a child's age in months, yet the state immunisation dashboard expects years. The grain of data, the frequency of refresh, and even the date formats vary wildly, turning simple joins into costly clean-ups that stall service delivery.

Example - At quarter-end, the agriculture portal exports beneficiary data to the treasury gateway for bulk transfers. Lakshmi's row fails the bank's validation check - "invalid IFSC"

—and is returned as an exception. Her payment is held until staff review the file, correct the code, and resubmit the batch.

2.4. Use: Where the Stakes Get Real - and the Rewards too

When government data feeds a service queue, a policy dashboard, or an AI model, any upstream flaw becomes a public-facing fault. Clean, time-stamped records can power tools that flag disease outbreaks, fine-tune subsidy targeting, or route traffic in real time. But if two datasets record farms at different granularities or age a child in months versus years, even the smartest algorithm will misfire.

Architecture matters too: a single data lake risks breaches, while sealed silos hide cross-domain patterns. Newer methods—federated learning, synthetic data, privacy-enhancing computation, let models train where data sits, but only if every record carries consistent schemas, consent flags, and update logs. In short, garbage in, algorithmic garbage out: disciplined capture, harmonised storage, and governed sharing are the non-negotiable tripod beneath any citizen-centric AI.

Example - The programme manager builds a drought-relief model that relies on "successful payment" history to spot cash-constrained clusters. Because Lakshmi's transfer shows as failed, her village appears slightly better off than it is, nudging fertiliser stock toward other blocks.

2.5. Retirement: The Data We Forget

Archival is the stage most systems skip. With no clear "sunset" rules, obsolete entries linger, inflating counts, slowing queries, and sometimes exposing private details. Welfare rolls still list families that have moved away; backup servers hold duplicates long after production tables are cleaned.

Fixing this is less about storage and more about policy: time-bound retention schedules, annual "alive-and-eligible" checks, and safe deletion once legal purpose lapses. Regular pruning keeps databases lean, dashboards truthful, and public trust intact, freeing analysts to focus on trends that matter, not ghosts in the machine.

Example - Six months later a data-quality drive updates Lakshmi's IFSC in the main server, but the duplicate row in an older backup is missed. Unless that orphan entry is archived or reconciled, it will keep resurfacing during audits and quarterly reconciliations, requiring manual checks each time.

Together, these five stages show how one mistyped field can ripple from capture to archive. The next chapter tallies the real-world cost of letting those cracks persist, and why they demand urgent repair.

This journey reveals how cracks can form at any link—and why fixing them demands a closer look at the systems behind the data.

3. THE NATURE OF THE CHALLENGE

India's public data engines hum louder each year, powering cash transfers, hospital authorisations, fertiliser caps, and pandemic dashboards. Yet familiar cracks reappear: errors slip through, gaps don't close, fixes don't hold. These aren't stray glitches, they stem from how data is captured, shared, and rewarded.

This chapter identifies six recurring design flaws, pairs each with a real-world incident, and highlights the risks associated with leaving them unaddressed.

3.1 Systemic, not Isolated

Data quality lives or dies in system design. Every dataset, whether a 1995 ledger or a 2025 API stream, inherits the roles, incentives and validation rules baked into its first mile. When incentives favour quantity over correctness, mistakes recur, no matter how many downstream audits we conduct.

Example - During a 2017 linkage drive, three regions uncovered 4.4 lakh "ghost students" claiming midday meal funds. In one state alone, more than two lakh fictitious names vanished once biometric attendance became mandatory, proof that lasting quality follows when accuracy is rewarded.

Risk: phantom beneficiaries drain budgets and trust every time a new scheme launches.

No stack of audits can substitute for giving the record's true owner - teacher, farmer, patient - live visibility and the right to correct it. The "two screen" strategy used at Aadhaar enrolment centres – exposing the data to the Aadhaar holder to view and correct the record – is an excellent example of implementing this.

3.2 Fragmented by Design

Platforms often start with good intentions but end up maturing in isolation. A ministry launches a portal tailored to its flagship scheme; states then clone and tweak that code to meet local rules; over time, each branch adds its own columns, codes, and shortcuts. Formats drift apart, hectares in one table, acres in another, full names in one file, initials in the next, dates logged as DD/MM/YYYY in one district but MM-DD-YY next door. What should be a simple join across systems turns into days of manual cross-walks and spreadsheet contortions, delaying benefits and muddying analysis.

Example - In 2021, roughly one-third of crop insurance claims in an eastern district stalled because land titles listed parents as owners while crop-loss forms named the farmers themselves. About 1.5 lakh applications sat idle until staff reconciled the two schemas

Risk: fractured views create blind spots and delay help when citizens need it most.

3.3 Legacy Systems, Modern Pressures

Many core registries still run on infrastructure built for a different era. They've been patched over the years, but the core design remains unchanged. Basic features, such as validation logic, audit trails, or version control, are often missing. Dashboards may have been added to improve visibility, but they sit atop brittle backends. As policies evolve and field realities shift, these systems struggle to keep pace. Even minor changes, such as updating a beneficiary status, can require custom workarounds.

Example- In 2024, elderly ration holders in a western district queued for months after fingerprint readers failed to recognise worn prints. Iris scanners arrived late, but thousands missed subsidised grain at the peak of inflation.

Risk: brittle systems turn minor tweaks into outages for households on the margin.

3.4 Everyone's Job, No One's Responsibility

Data moves across departments and systems but rarely comes with a clear owner. One team collects it, another uses it, and a third reports on it. When errors show up, it's not always clear who is responsible for correcting them or whether anyone has the authority to do so. This lack of custodianship means even obvious mistakes can persist for years, not because they're hard to spot but because they fall between the cracks of institutional roles.

Example - In late 2022, 17,000 health insurance cards were stuck because beneficiary names or biometrics failed identity checks. Health staff labelled it an "ID problem," the ID team saw a "health-database issue," and elderly citizens paid cash until a joint task force cleaned the queue.

Risk: obvious errors linger for years, eroding faith and loading staff with avoidable rework.

Industry practice shows that durable quality emerges when systems expose errors to the people who benefit most from clean data, instead of layering extra audits.

3.5 When Speed Trumps Quality

Most programmes are evaluated by how fast they move: how many people were enrolled, how quickly benefits were delivered, and how soon dashboard indicators turn green. These indicators are helpful, but they can create unintended consequences. When speed is rewarded, but accuracy is not, frontline staff may rush through data entry. Mid-level officials may hesitate to reopen past data for fear of triggering audit flags. Over time, this creates a culture where getting it done is more important than getting it right.

Example- During a large LPG subsidy roll-out in 2013, ID bank linkages covered only 60 per cent of households; the remaining 40 per cent were bulk-rejected to keep the launch on schedule. It took two years of painstaking review to unlock the full savings—showing that haste can cost more than delay.

Risk: performance metrics turn into vanity stats, hiding debts that grow costlier over time.

3.6 The Cost of Low Expectation

Once systems learn to live with error, they stop trying to prevent it. A database that's 80% accurate becomes "good enough." A mismatch between two registries is seen as a minor technical hiccup rather than a barrier to service delivery. This mindset shapes everything: how platforms are designed, how staff are trained, and how performance is judged. Slowly, quietly, quality becomes optional.

Example- A large northern state declared itself open-defecation-free in 2019; however, an audit of 590 rural houses found that nearly half lacked toilets. Dashboards still showed 100 per cent coverage, and funds kept moving—evidence that data judged "close enough" can override field reality.

Risk: policy rides on numbers no one fully trusts, undercutting evidence-based governance.

These six flaws reveal a design defect, not a tech shortfall. Quality will improve only when priorities shift from speed to precision, ownership, and aligned incentives. The next chapter calculates the cost of persistent cracks in rupees, including delays and the erosion of public trust. It explains why fixing them now is more cost-effective than patching them indefinitel.

To move forward, we need a shared benchmark of what quality truly means—beyond error rates, into attributes that build trust.

4. WHAT GOOD LOOKS LIKE: ANATOMY OF HIGH-QUALITY DATA - SIX ATTRIBUTES

Picture a GPS that shows every road but scrambles half the street names—you'd still get lost. Data behaves the same way.

Before we can improve data quality, we must define it—clearly and measurably. Without a shared understanding, every clean-up drive veers off course, leaving efforts inconsistent and results short-lived. Six core attributes—accuracy, completeness, timeliness, consistency, validity, and uniqueness—provide a practical framework to assess and improve public datasets.

Six attributes of 'good' data

Attribute	What 'Good' Looks Like
Accuracy	Names, codes, and coordinates match the trusted source on the first try, no corrections needed.
Completeness	Every required field is filled so the record can trigger a payment, call, or report without chasing extra details.
Consistency	The same person or place carries identical values wherever it appears, no double spellings, no age mismatch
Timeliness	Records are refreshed quickly enough to drive the intended action: minutes for payments, days for planning, never "stale."
Validity	Every entry passes format and logic checks—dates in range, PIN codes that exist, IDs that clear checksums.
Uniqueness	One row equals one real-world entity; duplicates are flagged and merged before numbers go to the dashboard.

Defining "good" is only half the battle; But even the clearest definition needs visibility to be useful. A simple Data-Quality Scorecard does precisely that. Designed for programme officers, MIS teams, and data stewards, the scorecard transforms the six attributes into a one-page, colour-banded tracker that shows, quarter by quarter, where a dataset meets the mark, where gaps persist, and who is responsible for each fix. Departments can adapt the template and instantly gain the visibility they need to prioritise corrections and hold teams accountable.

Here's a sample structure that departments can adapt to monitor key dimensions of data quality.

Sample Data Quality Scorecard

Illustrative example for programme registries or sectoral databases

Attribute	Definition	Indicator Example	Target Threshold	Current Performance
Accuracy	Correctness of data compared to real-world values	% of beneficiary bank-account records that pass NPCI's Account Verification Service (IFSC + account-number check)	≥ 98%	96.40%
Completeness	Presence of all required fields and records	% of records with non-null mobile number and bank details	≥ 95%	89.70%
Consistency	Internal coherence across fields and systems	% of matched age entries across health and education databases	≥ 90%	81.20%
Timeliness	Data is up-to-date and reflects current status	Average time lag between scheme approval and registry update (days)	≤ 3 days	6.1 days
Validity	Data conforms to expected format, range, and logic	% of records passing format validation rules	≥ 99%	97.80%
Uniqueness	Each real-world entity appears only once in the dataset	% of duplicate household IDs identified in deduplication audit	≤ 1%	3.50%

Note: The values shown in the "Current Performance" column are illustrative dummy figures, meant only to demonstrate how you can use this scorecard to benchmark and compare your own real metrics. Adjust each cell with your actual data when you run an assessment.

By turning abstract principles into trackable metrics, the scorecard makes quality visible—and therefore actionable.

Used consistently, the scorecard turns abstract ideals into visible gaps—making data quality a problem that teams can see, own, and fix.

And once the scorecard makes the red flags impossible to ignore, the next question is obvious: can the organisation keep data healthy—day in, day out—without endless fire drills?

High-quality systems do more than patch errors; they bake safeguards into every step: structured capture, schema controls at rest, real-time checks in motion, and planned archival. The goal is persistence, not perfection. No stack of audits can rescue accuracy if the people who rely on a record can't see and correct it themselves.

Durable quality follows only when roles are clear, incentives reward fidelity, and platforms prevent decay by design.

That's where the Data-Quality Maturity Framework comes in

- What it is: a four-pillar self-assessment—Governance, Validation, Monitoring, Integration—that rates each pillar from ad-hoc to best practice.

- Who uses it: department heads, nodal officers, and cross-functional data leads.

- Why it matters: an annual (with mid-year pulse) check turns "fix the data" into a timed, budgeted roadmap—fund a steward, automate rule engines, adopt shared schemas—rather than vague good intentions.

With scorecards tracking today's vitals and the maturity framework securing tomorrow's health, agencies gain both a dashboard and a maintenance plan for trustworthy public data.

Data Quality Maturity Framework
A self-assessment guide for public institutions

How to Use This Framework:
Self-assess each row by identifying which description best matches your current state.
Use the result to identify gaps and **prioritise improvements;** e.g., move from "Emerging" to "Established" over the next year.
This tool can be applied at the department, scheme, or even platform level (e.g., PMAY, eShram, DigiLocker).
It can be used to guide capacity-building, budget planning, and platform redesigns.

Dimension	Level 1: Foundational	Level 2: Emerging	Level 3: Established	Level 4: Advanced	Level 5: Institutionalised
1. Data Governance & Ownership	No defined data owners or custodians	Named data focal points exist, but roles are informal or limited in scope	Department-level data stewards identified with clear responsibilities	Cross-department governance with coordination mechanisms in place	Stewardship embedded in institutional roles; linked to performance and budgets
2. Standards & Metadata	Data fields and formats vary across systems	Some common formats used, but documentation is sparse	Data dictionaries and naming conventions adopted within departments	Cross-sector metadata standards adopted; documented and updated regularly	Standards institutionalised across levels of government; aligned with national frameworks
3. Data Capture & Validation Controls	Minimal validation at entry; errors are common and go undetected	Basic field-level checks (e.g., format, mandatory fields) in place	Automated validation and duplication checks used during data entry	Real-time validation against reference data; alerts and escalation built in	Validation built into service design; field staff trained and supported to capture clean data

Dimension	Level 1: Foundational	Level 2: Emerging	Level 3: Established	Level 4: Advanced	Level 5: Institutionalised
4. Quality Monitoring & Reporting	No routine monitoring; quality issues identified only during audits	Ad hoc reviews; some manual corrections made	Periodic checks using scripts, reports or third-party assessments	Dashboards track quality metrics (e.g., error rates, missing fields)	Data quality KPIs tracked and published regularly; linked to planning and delivery reviews
5. Correction & Feedback Loops	Corrections made on request; no formal process	Manual correction workflows exist but are slow or informal	Designated process for correction and grievance redressal	Feedback loops exist from users to data owners; tracking of correction status	Continuous feedback integrated into workflows; audit trails and accountability mechanisms in place
6. Interoperability & Integration	Systems operate in silos; no data exchange across departments	Bilateral data exchanges through manual reports or shared files	APIs exist between key systems; limited real-time data flows	Federated systems with shared identifiers and automated exchange	Seamless, consented data flow across services; common reference data and vocabularies shared at scale
7. Culture & Capacity	Data quality not prioritised; limited awareness or training	Staff recognise the need but lack tools or time to act	Regular training and quality awareness activities underway	Quality practices seen as core to programme performance	Culture of data ownership; quality is part of organisational identity and decision-making

The next question is how far each dataset already stands from these targets

Institutions can use this framework to move from reactive fixes to a planned, progressive approach to data quality, aligned with their own context and capacity. This is not a ranking exercise, but a diagnostic aid. It helps organisations identify gaps, prioritise actions, and create a shared language around what it means to build data systems that are not only functional, but resilient and future-ready.

Used together, the scorecard and maturity framework turn firefighting into foresight—exposing blind spots, ordering fixes, and giving every team a shared language for resilient, future-ready data. Yet diagnosis is only the prologue. The next step is turning insight into quick, verifiable wins. Chapter 5—the Starter Kit— distils proven tactics that any programme can pilot rapidly, translating "what good looks like" into measurable improvements on the ground.

5. STARTER KIT: QUICK WINS AND TACTICAL PATHWAYS

With a shared yard-stick is in hand—the six core attributes, a scorecard to flag red cells, and a maturity lens to show whether fixes will stick—the next question is what to tackle first.

Drawing on proven enterprise playbooks, three field-ready moves stand out—Fix it at the Source, Keep it Clean, and Make it Matter. Each can be rolled out in under a quarter and tracked on the existing scorecard. No sweeping rewrites or new legislation are required—just light tweaks to capture screens, monitoring cycles, and accountability loops that any public programme can pilot and scale.

5.1 Fix it at the source

(Focus on validation, deduplication, and metadata)

A. Real-time validation during data entry

Most errors originate at the point of capture. Preventing them upfront saves time and effort later.

- Enforce field-level checks (e.g., PIN code format, date logic).
- Use dropdowns or autocomplete for standard entries (e.g., district names, job categories).

- Flag missing mandatory fields before submission.

Some digital identity platforms now include real-time validation features, like auto-formatting fields, verifying entries against known datasets, or flagging duplicates, to catch errors before submission.

B. Deduplicate registries using open-source tools

Duplicates inflate counts, misdirect benefits, and complicate audits.

- Use lightweight tools (like Dedupe.io or Python-based matching) to flag probable duplicates.
- Prioritise deduplication before scheme expansion or transfers.

C. Publish and maintain a basic data dictionary

Ambiguity in field definitions creates inconsistency in data entry and analytics.

- List field names, accepted formats, mandatory/optional status, and update logic.
- Share internally and with vendors or state-level partners.

When organisations use standardised field definitions and data templates, they reduce ambiguity at the point of capture, making downstream data cleaner and easier to work with. Simpler the data is to understand, the more useful it becomes.

5.2 Keep it clean

Audits, stewards, and correction workflows

A. Conduct regular, sample-based data audits

Frequent, light-touch reviews can catch minor issues before they compound.

- Sample recent entries (e.g., 5–10%) each month for key fields.
- Check for blanks, invalid values, or mismatches.
- Document recurring patterns for system or training updates.

In many settings, even basic monthly reviews of key fields have helped reduce avoidable errors and improve staff awareness.

B. Assign internal data stewards

Designated points of contact reduce ambiguity and improve responsiveness.

- Identify one staff member per dataset or registry.
- Give them edit rights and visibility on known issues.
- Review data quality performance as part of regular team meetings.

C. Link grievance redressal with backend correction

Citizens often notice errors before systems do.

- Add "Report a mistake" options on portals or SMS updates.
- Route flagged issues to the correction workflows.
- Notify users once the update is complete.

Some grievance redressal systems now link user complaints directly to backend databases, allowing reported issues to trigger updates in service records, not just status tickets.

5.3 Make it matter

Dashboards, recognition, and feedback loops to make data available in a highly accessible manner.

A. Use dashboards to track quality, not just outputs

Service delivery dashboards often monitor coverage, but not data integrity.

- Add indicators like % of blank fields, error rates, and duplication flags.
- Show trends over time by district or block.
- Highlight gaps for targeted support.

B. Recognise and reward data quality improvements

Positive reinforcement builds a culture of care.

- Highlight top-performing blocks in monthly reviews.
- Publicly acknowledge field teams with high-quality submissions.
- Include data quality in annual appraisals or incentive frameworks.

C. Run a 'Data Quality Week' as an internal campaign

Short, focused campaigns help rally attention and clean up long-standing issues.

- Set goals (e.g., fill missing phone numbers, remove dead records).
- Track and share progress across teams.
- Debrief and document what worked for future replication.

These interventions do not require new laws, budgets, or platforms. They need ownership, clarity, and consistent follow-through. They're a way to start today, and build habits that last. When repeated and scaled, small moves like these shift the culture around data. They reduce the cost of correction. They make systems more responsive. And they signal to staff, leadership, and citizens that quality is not optional. It is essential. Over time, these efforts lay the foundation for a shared culture of quality that must be reinforced, coordinated, and made visible across institutions. The next step is to move from individual habits to collective norms, where quality becomes a shared expectation across systems, institutions, and levels of government.

But to turn quick wins into lasting norms, data quality needs deeper roots—anchored in roles, incentives, and systems that don't reset with each programme cycle.

6. BUILDING BLOCKS FOR CHANGE: INSTITUTIONS, INCENTIVES & INTEROPERABILITY

Improving data quality isn't just a matter of better tools or tighter rules; it's about shifting how institutions think, behave, and collaborate. The previous chapters laid out what quality looks like and how to measure it. But making quality last requires something more enduring: ownership, incentives, and infrastructure that support good data as a habit, not a one-off intervention. This chapter offers three practical levers to institutionalise quality from within the system.

6.1 Institutionalising Ownership

Quality suffers when data has no clear owner. In many programmes, data moves across departments and levels, but no single entity is responsible for maintaining its integrity from end to end. The fix lies in assigning stewardship, not just for compliance, but for ongoing care. This means designating custodians at each level (national, state, district) who are

empowered and accountable for data health. These roles don't need to be new hires; they can be embedded within existing structures with a defined mandate, clear escalation paths, and regular review cycles.

In addition, quality must be seen as a shared responsibility. Programme heads, IT teams, and field staff all have roles to play. When Ownership is structured and supported, errors don't just get corrected—they get prevented.

Under the U.S. Foundations for Evidence-Based Policymaking Act (2019), every federal agency was required to designate a Chief Data Officer responsible for data governance and quality. By October 2022, the Government Accountability Office reported that 82 agencies had officially appointed CDOs, embedding stewardship at the heart of their data operations.[7]

6.2 Getting Incentives Right: Rewarding Quality, Not Just Speed

Most public programmes reward what they can count: enrolments completed, forms submitted, and dashboards updated. These indicators are useful, but they can create unintended consequences. When speed is rewarded but accuracy is not, field staff may rush through data entry. When outputs matter more than inputs, verification becomes a burden rather than a standard. Data quality indicators—such as error rates, completion levels, and timeliness—can be tracked and factored into programme reviews, not as punitive audits but as a reflection of delivery strength.

Some states have implemented scorecards to rank departments based on data quality, fostering a constructive sense of competition. Others have tied data improvements to budget flexibility or expedited approvals. Although the approaches may differ, the fundamental principle remains the same: when quality is measured, recognized, and incentivized, it becomes significant.

6.3 Making Systems Work Together: The Case for Interoperability

Even clean data loses value if systems can't use it together. Interoperability—across platforms, departments, and time—is essential to making public data work harder. It depends on shared schemas, reference codes, and APIs that allow secure, controlled access between programmes.

India's toolbox is steadily growing:

- IndEA provides a common vocabulary and reference architecture.
- NDGFP sets rules for sharing, consent, and anonymisation.
- Account Aggregator, DigiLocker, DEPA show how consent-driven flows can reduce duplication.
- Community-led tools—from CDPI to iSPIRT—offer off-the-shelf templates, APIs, and governance playbooks.

Still, most systems remain siloed, relying on PDFs or manual copy-paste. The next step is scale: making these standards the default, not the exception. That means using common data models where possible—for people, households, land parcels—and building new systems on existing rails rather than starting from scratch.

Singapore's Government Data Architecture shows what's possible: structured datasets, clear access rules, shared protocols. India already has the building blocks. The challenge now is connection—between systems, teams, and incentives. That's what makes quality stick.

7. THE ROAD AHEAD: TOWARDS A SHARED CULTURE OF DATA EXCELLENCE

India's digital foundations are in place. The challenge now is ensuring that what flows through them—our data—is accurate, complete, and trusted. Data quality is no longer a backend concern; it is central to public trust, effective service delivery, and the success of India's own AI ecosystem.

But this isn't a problem technology alone can solve. True change requires a cultural shift—where quality is everyone's responsibility, not just the IT team's. From frontline staff to policymakers, every actor must see themselves as a data steward.

This shift must be supported by light-touch, collaborative mechanisms: voluntary scorecards, peer-learning networks, flexible incentives, and playbooks of best practices that evolve with experience. These resources—maintained by a neutral body like NITI Aayog or the proposed India Data Management Office—should blend technical, institutional, and behavioural insights from across government and sectors.

States can lead the way by embedding data quality cells, linking quality to service outcomes, and recognising excellence. Capacity building, leadership development, and hands-on support will be essential to sustain momentum.

Above all, this transformation needs visible commitment from the top. A government that champions clean, usable data sends a powerful signal: that quality is not an afterthought—it is the foundation of good governance.

India's digital future will be shaped not just by how many platforms we build, but by how much trust we build into them. And that begins with data that's ready to serve.

REFERENCES:

1. https://www.financialexpress.com/business/banking-finance-upi-transactions-up-34-on-year-in-april-down-a-tad-from-march-3829424/

2. https://economictimes.indiatimes.com/news/india/aadhaar-authentications-touch-2707-crore-in-2024-25/articleshow/120703093.cms?

3. https://www.linkedin.com/pulse/24th-march-2025-over-369-crore-ayushman-cards-have-m91vc/

4. https://www.business-standard.com/india-news/centre-to-save-rs-18-000-cr-by-removing-fake-accounts-from-welfare-schemes-123101700528_1.html?

5. https://timesofindia.indiatimes.com/business/india-business/govt-sraps-rs-3-5-crore-bogus-lpg-connections-saves-rs-21000-crore-subsidy-in-2-years/articleshow/53326003.cms?

6. https://cag.gov.in/uploads/download_audit_report/2023/Report-No.-11-of-2023_PA-on-PMJAY_English-PDF-A-064d22bab2b83b5.38721048.pdf?

7. https://www.gao.gov/assets/gao-23-105514.pdf