A Vision for National Statistics

Building an Integrated National Data Infrastructure for Pakistan

Zahid Asghar, School of Economics, Quaid-i-Azam University Islamabad

Compiled from the provided Quarto (.qmd) chapters.

Table of Contents

Ethics, Privacy, and the Rights of Data Subjects **65**

Preparing for the Future: AI-Ready Data **72**

A Vision for National Statistics

A Vision for National Statistics Building an Integrated National Data Infrastructure for Pakistan ::: {.subtitle} A Framework for Transforming How Pakistan Produces, Shares, and Uses Statistical Information — 2026

About This Document Pakistan collects an enormous amount of data. NADRA holds biometric records for over 200 million citizens. The Federal Board of Revenue maintains tax files. BISP manages registries of millions of beneficiary households. Provincial health departments operate facility-level reporting systems. Education departments track enrolments and teacher deployments. The Pakistan Bureau of Statistics conducts censuses, labour force surveys, and household economic surveys. Yet this data largely sits in institutional silos — collected by different agencies, for different purposes, using different definitions, in incompatible formats, with no systematic mechanism for combining it. The result is that a country of 241 million people makes policy decisions on the basis of information that is often incomplete, outdated, or fragmented. The data exists. What is missing is the infrastructure to make it usable. This document argues that Pakistan needs a fundamental shift in how it produces and uses statistical information. Not by abandoning surveys, which remain essential, but by building a national data infrastructure that blends survey data with administrative records and other sources to produce statistics that are more timely, more granular, and more useful than any single source can provide alone. This requires repositioning the Pakistan Bureau of Statistics from a data collector to a national coordinator, establishing governance and legal frameworks for data sharing, investing in human capacity, protecting the rights of data subjects, and designing the system for the analytical demands of an AI-driven future. The chapters that follow develop this argument through diagnosis, international evidence, institutional analysis, and a sequenced set of priorities for implementation. --- ## Chapters ::: {.section-group} The Problem

::: {.chapter-card} ### 1. Why Statistics Matter and Where Pakistan Stands Today (01-why-statistics-matter.qmd) Statistics as national infrastructure, the pressures facing Pakistan's survey-based system, and why the current model — siloed agencies, periodic censuses, standalone surveys — can no longer meet the country's needs.

::: {.section-group} The Framework

::: {.chapter-card} ### 2. Understanding Data Infrastructure (02-understanding-data-infrastructure.qmd) What data infrastructure actually means — not just technology, but the combination of data assets, institutional arrangements, skills, governance frameworks, and the communities that sustain them.

3. The Case for Blended Data (03-case-for-blended-data.qmd) Why no single data source is sufficient and what it takes to combine survey, administrative, geospatial, and other data sources to produce richer and more timely statistics.

4. Key Data Holders in Pakistan (04-key-data-holders.qmd) Who holds the data that matters — from PBS, NADRA, FBR and SECP to provincial governments, BISP, private sector enterprises, and why coordination across these holders is the central challenge.

::: {.section-group} Principles & Governance

::: {.chapter-card} ### 5. Core Principles for a Modern Data Infrastructure (05-core-principles.qmd) The foundational principles — using data for the common good, proportionality, transparency, privacy by design, and institutional independence — that should guide Pakistan's infrastructure.

6. Repositioning PBS: From Data Collector to National Coordinator (06-repositioning-pbs.qmd) The case for transforming PBS from an agency that collects its own data into one that coordinates, curates, and governs data flows across the entire national statistical system.

7. Data Quality: Fitness for Use (07-data-quality.qmd) Why quality means more than accuracy — fitness for use, the dimensions of quality for blended data, metadata requirements, and the FAIR principles as a framework for making data usable.

::: {.section-group} Implementation

::: {.chapter-card} ### 8. Organisational Options for the New Infrastructure (08-organisational-options.qmd) International models — from the Netherlands' centralised CBS to Estonia's distributed X-Road to Australia's AIHW — and what Pakistan can learn for designing its own institutional arrangement.

9. Ethics, Privacy, and the Rights of Data Subjects (09-ethics-privacy.qmd) The care.data cautionary tale, re-identification risks, social licence, group privacy, the Five Safes framework, and why Pakistan's absent data protection law is a credibility crisis for the entire infrastructure.

10. Preparing for the Future: AI-Ready Data (10-ai-ready-data.qmd) What AI-ready means in practice, the FAIR principles for machine consumption, the UK National Data Library concept, and why the data infrastructure is the binding constraint — not the algorithm.

11. Moving Forward: Priorities and Conclusion (11-moving-forward.qmd) A sequenced plan — from data inventory and pilot projects to legal reform and institutional design — and why Pakistan's 241 million people deserve a statistical system equal to the complexity of their lives.

--- ::: {.callout-note} ## Guiding Idea "Just as bridges and highways facilitate the transportation necessary for commerce, statistical information informs decisions by governments, business enterprises, and individuals." — National Academies of Sciences, Engineering, and Medicine (2023)

Why Statistics Matter and Where Pakistan Stands Today

Why Statistics Matter and Where Pakistan Stands Today

Every modern government makes decisions on the basis of statistical information, whether it recognises this dependency or not. When a finance minister allocates development funds across provinces, the allocation formula relies on population counts and economic indicators. When a health ministry decides where to build new facilities, it draws on data about disease burden, facility utilisation, and geographic coverage. When a planning commission sets growth targets, it uses national accounts estimates and labour force statistics. When a political opposition challenges the government's claims of progress, the challenge is credible only if there are independent numbers to point to.

> Statistics, in this sense, are not a technical luxury. They are the informational foundation on which governance, accountability, and democratic debate all rest.

When the numbers are absent, outdated, or untrustworthy, the consequences are not abstract. Resources are allocated on the basis of political bargaining rather than evidence. Policies are designed on the basis of assumptions rather than facts. Programme failures go undetected because there are no baseline measurements against which to assess performance. Citizens who suspect the government is misrepresenting conditions have no authentic source to consult. The absence of credible statistics creates a space that is filled by rumour, manipulation, and guesswork. As a result, people who suffer most from this are those who are already least visible to the state.

There are parallels between data infrastructure and physical infrastructure and its not merely a convenient analogy. A road is useful not in itself but because people and goods move on it. A statistical system is useful not because it collects data but because the data is used — by governments designing policies, by businesses planning investments, by researchers studying social and economic dynamics, by journalists holding power to account, and by ordinary citizens trying to understand the conditions of their own lives. The value of data infrastructure, like the value of transport infrastructure, is realised through use. When data sits unused in the filing systems of government agencies, its potential is wasted as surely as a highway built to nowhere.

The United States National Academies of Sciences, Engineering and Medicine made this comparison the organising framework of its landmark 2023 report, *Toward a 21st Century National Data Infrastructure*. The report argued that "just as bridges and highways facilitate the transportation necessary for commerce, statistical information informs decisions by governments, business enterprises, and individuals" (NASEM, 2023). The report went further: it described a paradox now confronting statistical systems worldwide. The traditional method of producing national statistics — large-scale probability surveys conducted by government statistical agencies — is under severe pressure from declining participation rates, rising costs, and growing demand for more timely and more granular information. Yet at the same time, the volume of digital data being generated about the activities of individuals and businesses has never been greater. The challenge is to build infrastructure that can mobilise these new data assets while maintaining the rigour and public trust that legitimate statistics require.

This challenge is universal. But its stakes are particularly high in countries like Pakistan, where the gap between what the statistical system produces and what governance requires is already large and growing wider. In a country where political discourse frequently revolves around competing claims — about the size of the economy, the number of people living in poverty, the quality of public services, the distribution of resources across provinces — the absence of authoritative, timely, and trusted statistics is not merely an inconvenience. It is a structural weakness in the democratic process itself. When citizens and their elected representatives cannot agree on the facts, debate becomes a contest of assertions rather than an exchange of evidence. An efficient and accountable government depend fundamentally on the availability of data that all parties accept as credible.

What Pakistan Needs to Know About Itself

Pakistan is a country of 241 million people, the fifth most populous nation on earth, with a young and rapidly growing population that is urbanising at one of the fastest rates in South Asia. The challenges facing the country are enormous and interconnected: a stunting rate that affects roughly 40 percent of children under five; a net primary enrolment rate that still leaves almost 22 million children out of school; a labour market in which the majority of employment is informal and therefore largely invisible to official statistics; an economy in which tax collection as a share of GDP remains among the lowest in the region; energy systems under strain; agricultural production increasingly vulnerable to climate shocks; and persistent inequalities across provinces, between urban and rural areas, and between men and women.

None of these challenges can be addressed effectively without credible data. To design a social protection programme that reaches the poorest households, you need to know who is poor, where they live, and what characterises their deprivation. To improve primary education, you need to know not just how many children are enrolled but whether they are learning and what happens to them after they leave school. To reform the tax system, you need to understand who is paying, who is not, and why. To respond to a flood or a drought, you need real-time information about which areas are affected, how many people are displaced, and what resources they need. To hold provincial governments accountable for their use of development funds after the 18th Amendment devolved responsibilities to the provinces, you need provincial- and district-level data that is comparable, timely, and independently verifiable.

Pakistan needs, in other words, a statistical system capable of answering the questions that matter most for its future. The question is whether the system it currently has is up to this task.

A System Built for a Different Era

Pakistan's current statistical infrastructure, like other countries in the world, was designed for the conditions of the 20^{th} century. It has served the country with reasonable adequacy under those conditions. The Pakistan Bureau of Statistics (PBS) — the national statistical agency operating under the Statistics Division within the Ministry of Planning, Development and Special Initiatives — has historically relied on two principal instruments: periodic censuses and sample surveys. The population census, constitutionally mandated every ten years, provides the demographic baseline. The Pakistan Social and Living Standards Measurement Survey (PSLM), the Household Integrated Economic Survey (HIES), the Labour Force Survey (LFS), and other periodic exercises provide the social and

economic indicators on which policy has been built.

This model has genuine strengths. Probability sampling, when properly implemented, allows a relatively small number of surveyed households to generate statistically valid estimates for large populations. The surveys conducted by PBS have produced the core indicators — literacy rates, enrolment ratios, poverty estimates, unemployment rates, health access measures — that have shaped Pakistan's development discourse for decades. The 2023 Digital Census, which deployed 121,000 enumerators equipped with tablets linked to geographic information systems, represented a significant technological advance and geo-tagged approximately 40 million structures across the country. The HIES 2024-25, released in early 2026, was the first fully digital household economic survey in PBS history.

But the model is under pressure from multiple directions simultaneously, and the pressures are intensifying.

The Pressures Are Real

First, the economics of survey-based data collection are increasingly unfavourable. Surveys are expensive to design, field, and process. The PSLM district-level survey requires enumerating roughly 195,000 households across thousands of sampling units — a logistically enormous operation requiring trained field staff, transport, supervision, quality control, and data processing infrastructure. As costs rise, the frequency, sample size, or geographic coverage of surveys must be sacrificed. Pakistan cannot simultaneously afford annual surveys at the district level, quarterly labour force surveys at the provincial level, and the specialised surveys (on time use, disability, migration, agriculture) that specific policy questions demand. Something has to give, and what gives is either timeliness, granularity, or both.

Second, survey response rates globally are declining, and Pakistan is not immune. Respondent burden — the time and effort required to complete lengthy questionnaires — is a growing concern, particularly in urban areas where households are more mobile and less willing to participate. The HIES 2024-25 reported a non-response rate of 1.86 percent, which appears low but understates the challenge: 157 primary sampling units out of 2,500 were dropped entirely due to "administrative and other issues." In many countries, the decline in survey participation has been dramatic enough to threaten the validity of the estimates produced. The NASEM report described this as a "severe threat to the quality of statistical information" (NASEM, 2023). The problem is self-reinforcing: as participation declines, the remaining respondents become less representative of the population, which erodes the quality of estimates, which erodes public trust in the numbers, which makes future participation even harder to obtain.

Third, and perhaps most important, the demand for statistical information has changed qualitatively. Policymakers no longer want national-level estimates published twelve to eighteen months after data collection. They want provincial and district-level data. They want it quarterly or monthly, not annually. They want to be able to track the effects of specific interventions — a cash transfer programme, a school construction initiative, a health campaign — in near real-time. They want to understand the intersections between different dimensions of wellbeing: how health affects educational outcomes, how education affects employment, how employment affects poverty. None of these demands can be met by a system that relies exclusively on periodic, standalone surveys, each designed for a single

purpose and published on its own schedule.

Fourth, the current system operates in silos. PBS conducts its surveys. NADRA maintains its biometric databases. BISP manages its beneficiary registry. The Federal Board of Revenue holds tax records. Provincial health departments operate DHIS2. Provincial education departments maintain NEMIS. Each agency collects data for its own administrative purposes, using its own definitions, coding systems, and formats. There is no systematic mechanism for combining these data sources to produce richer statistical products. When a researcher or policymaker needs to understand the relationship between, say, household income, health utilisation, and educational attainment, they find that the data exists in multiple agencies but cannot be linked, because no common identifiers, shared standards, or institutional arrangements for data sharing are in place.

The poet Coleridge's observation — "water, water, everywhere, nor any drop to drink" — captures the situation precisely. Pakistan is not short of data. It is short of infrastructure that makes data usable.

The Census as a Case Study

The population census illustrates both the strengths and the limitations of the current system. The Constitution mandates a census every ten years, but Pakistan has managed only seven since independence — in 1951, 1961, 1972, 1981, 1998, 2017, and 2023. The gap between 1998 and 2017 — nineteen years — meant that for nearly two decades, the country's most basic demographic information was outdated, with profound consequences for political representation, resource allocation, and development planning.

The 2023 Digital Census was a significant achievement, deploying advanced technology at enormous scale. But the census also illustrated systemic challenges. Enumeration, originally planned for one month, was extended five times as coverage in major cities — Karachi, Lahore, Faisalabad — proved difficult. The results, like those of 2017 before them, became politically contested, with the Government of Sindh and several political parties challenging the count for Karachi. These controversies are not unique to Pakistan — census results are politically sensitive everywhere because they determine political representation and resource allocation. But they are exacerbated by the long intervals between censuses. When a country goes nineteen years without a census, as Pakistan did between 1998 and 2017, the stakes attached to each individual count become so high that the exercise itself become a source of political conflict rather than a resolution of it.

The census also consumed enormous institutional energy and resources that, in a more diversified statistical system, might have been complemented by continuously updated population estimates derived from administrative records — birth registrations, NADRA identity records, school enrolments — rather than depending entirely on a single decennial exercise. The Nordic countries, for example, have largely replaced traditional field-based censuses with register-based censuses that use linked administrative data to produce population statistics at a fraction of the cost and with far greater frequency. Pakistan is not in a position to replicate this approach immediately, but the direction of travel is clear: the future of population statistics lies in the continuous integration of multiple data sources, not in periodic mobilisations of six-figure enumerator armies.

This is the fundamental limitation: the current system places too much weight on large, periodic, standalone data collection exercises and too little on the continuous, systematic use of the data that government agencies already collect in the course of their daily operations. The data is there. What is

missing is the infrastructure — technical, institutional, and legal — to connect it.

The Case for Transformation

The argument of this document, developed across the chapters that follow, is that Pakistan needs a fundamental shift in how it produces and uses statistical information. This shift involves not abandoning surveys — which remain essential for measuring dimensions of life that administrative records cannot capture — but supplementing them with a data infrastructure that can blend survey data with administrative records, geospatial data, and other sources to produce statistics that are more timely, more granular, and more useful than any single source can provide alone.

This requires action on multiple fronts. It requires repositioning PBS from a data collector into a national data coordinator and steward. It requires establishing common standards for data quality, documentation, and interoperability. It requires building the institutional and legal frameworks for governed data sharing across agencies. It requires investing in human capacity — in data science, record linkage, statistical disclosure control, and data governance. It requires taking seriously the ethical obligations owed to the people whose data is being used, including robust privacy protections and meaningful accountability for misuse. And it requires designing the infrastructure not just for today's analytical methods but for the AI-driven methods that will increasingly define the frontier of what is possible.

None of this will happen overnight, and none of it will happen without sustained political commitment, institutional reform, and investment. But the costs of inaction — a country of 241 million people governing itself on the basis of incomplete, outdated, and fragmented information — are far greater than the costs of building the infrastructure that a modern statistical system requires.

The chapters that follow develop each element of this argument in detail. They examine how other countries have built modern data infrastructure and what Pakistan can learn from their experience. They set out the case for blended data — the systematic combination of survey, administrative, and other data sources — and the technical, institutional, and governance conditions it requires. They discuss the organisational models available for managing a national data infrastructure, the quality standards that must govern it, and the ethical frameworks that must protect the people whose data it uses. They address the emerging demands of artificial intelligence and the data foundations it requires. And they propose a sequenced set of short- and medium-term priorities for moving from the current system to the one the country needs.

Throughout, the argument rests on a simple premise: that the data Pakistan needs to govern itself effectively largely already exists, scattered across government agencies and systems that do not communicate with each other. The task is not primarily to collect more data. It is to build the infrastructure — technical, institutional, legal, and human — that connects what exists into a coherent, trustworthy, and usable picture of the country and its people.

References

NASEM (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good.* Washington, DC: National Academies Press.

NASEM (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources.* Washington, DC: National Academies Press.

PBS (2023). Announcement of Results of 7th Population and Housing Census-2023 'The Digital Census.' Islamabad: Pakistan Bureau of Statistics.

PBS (2025). Household Integrated Economic Survey (HIES) 2024-25. Islamabad: Pakistan Bureau of Statistics.

Understanding Data Infrastructure

What Do We Mean by Data Infrastructure?

When people hear the word "infrastructure", they usually think of roads, bridges, electricity grids or water supply systems. Physical things. But in the 21st century, data has become equally important as a foundational layer for how societies function. And just like physical infrastructure, data infrastructure is not a single thing. It is a system of many parts working together.

The Open Data Institute (ODI), which has done significant work on this concept, defines data infrastructure as consisting of "data assets, standards, technologies, policies and the organisations that operate and maintain them" (ODI, 2018). This is a useful starting point. But for a country like Pakistan, where the statistical system is still largely survey-dependent, we need to think about it in more practical terms.

At its core, data infrastructure is the whole arrangement that makes it possible to collect data, store it safely, process it, share it across organisations, analyse it, and turn it into information that people can actually use. If any part of this arrangement is missing or weak, the output suffers. You can have the best data in the world, but if there is no legal framework allowing its sharing, it stays locked in a silo. You can have excellent technology, but without skilled people to operate it, it remain underutilised.

The World Bank's Development Report 2021, titled "Data for Better Lives", made this point quite clearly. It argued that the value of data depends not just on its collection but on an entire ecosystem of governance, capacity, and trust that surrounds it (World Bank, 2021). Pakistan needs to take this seriously.

Data Assets: The Raw Material

Data assets are the starting point of any infrastructure. These are the actual datasets, records, and files that contain information about people, businesses, transactions, geography, and so on.

In Pakistan, the most obvious data assets are the ones produced by PBS — household surveys, labour force surveys, the population census, economic census. These are designed specifically for statistical purpose and they follow established methodology. But they are not the only data that exist.

There is a large amount of administrative data sitting with federal agencies. NADRA holds biometric and identity records for most of the population. FBR has tax records. BISP maintains records of social protection beneficiaries. SECP has data on registered companies and corporate filings. Regulatory bodies in telecom, banking, and energy also generate data as part of their normal operations.

Provincial governments are another major source. Education departments collect enrollment and attendance data. Health departments track disease surveillance and facility usage. Agriculture departments collect crop data. Much of this is not shared with PBS or with other provinces.

Then there is private sector data. Telecom companies have call detail records and mobility data. Banks process millions of transactions daily. E-commerce platforms, ride-hailing services, and digital payment systems all generate data that could, in principle, provide real-time indicators of economic

activity.

Finally, there is data from academic institutions, NGOs, and international organisations. These often conduct their own surveys or compile datasets that supplement official statistics.

The critical point is this: no single data source gives a complete picture. Each has gaps. Survey data is slow and expensive. Administrative data was not designed for statistics. Private sector data may not be representative. As the UNECE guidelines on using administrative data for statistics point out, "the statistical use of administrative data requires careful assessment of coverage, concepts, and quality" (UNECE, 2011). The value comes from combining them thoughtfully.

Technology: More Than Just Computers

When we talk about technology in the context of data infrastructure, we mean the tools and systems that allow data to be discovered, accessed, stored, processed, protected, and shared. This is a broad category.

It includes the obvious things — servers, databases, cloud computing platforms, data warehouses. But it also includes less visible components that are equally important. Application Programming Interfaces (APIs), for example, are what allow different software systems to talk to each other and exchange data automatically. Without APIs, data sharing between agencies requires manual processes which are slow, error-prone, and difficult to scale.

Data management platforms are needed for cataloguing what data exists, tracking its lineage (where it came from, how it was transformed), and maintaining metadata. Security systems — encryption, access controls, audit trails — are essential for protecting sensitive information.

One concept that is especially relevant here is **interoperability**. This means the ability of different systems to work together. If PBS uses one format for geographic codes and NADRA uses a different one, linking their data becomes a technical headache. Common standards, shared identifiers, and compatible formats are what make interoperability possible.

The European Interoperability Framework, developed by the European Commission, identifies four layers of interoperability: legal, organisational, semantic, and technical (European Commission, 2017). Pakistan's data infrastructure needs to address all four. It is not enough to just buy new hardware or install new software. The technology must be designed so that data can flow between organisations smoothly and securely.

A study by Kitchin (2014) on "The Data Revolution" emphasised that the real challenge is not generating data — we already generate enormous volumes — but organising, managing, and governing it in ways that make it usable. This resonates strongly with Pakistan's situation.

People and Expertise: The Human Factor

You can build the most advanced technology platform in the world. But if the people who supposed to use it lack the necessary skills, it will not deliver results. This is a lesson that many countries have learned the hard way.

A functioning data infrastructure requires people with diverse skills. Some of these are technical — data engineering, statistical programming, machine learning, database administration. But others are equally important and often overlooked. Data governance specialists. Legal experts who understand privacy law. Statisticians who can design multi-source surveys. People who can write metadata documentation. Communication experts who can explain statistical findings to non-technical audiences.

Pakistan faces a significant skills gap in this area. PBS and provincial bureaus have experienced survey statisticians, but fewer staff with skills in data science, record linkage, or privacy-enhancing technologies. The training infrastructure for these skills is limited, though some universities are beginning to offer relevant programmes.

The UN Statistical Commission has repeatedly emphasised that "modernisation of statistical production requires not only new technologies but new competencies" (UNSC, 2019). Countries that have successfully modernised their statistical systems — like Estonia, New Zealand, and the Netherlands — invested heavily in retraining their statistical workforce alongside technology upgrades.

Building human capacity takes time. It cannot be done overnight. But it can be started immediately through targeted training programmes, partnerships with universities, staff exchanges with other national statistical offices, and recruitment of specialists from the private sector. Without this investment, technology alone will not solve the problem.

Governance, Standards, and Legal Framework

This is perhaps the least visible but most critical component of data infrastructure. Governance refers to the rules, policies, processes, and institutional arrangements that determine how data is managed throughout its lifecycle.

The DAMA International Data Management Body of Knowledge (DAMA-DMBOK) defines data governance as "the exercise of authority, control, and shared decision-making over the management of data assets" (DAMA International, 2017). In practical term, this means answering questions like: Who is responsible for a given dataset? Who can access it? For what purposes? Under what conditions? How is quality ensured? How are disputes resolved?

For Pakistan, governance challenges are significant. There is no comprehensive legal framework that specifically addresses data sharing for statistical purposes across agencies. The Statistics Act provides PBS with certain authorities, but it does not fully cover the landscape of multi-source data that a modern infrastructure requires.

Standards are a closely related issue. Data standards define how data should be structured, labelled, classified, and documented so that it can be understood and used across organisations. Without common standards, you get a situation where one agency records income in monthly terms, another in annual terms, a third in different currency or using different categories. Linking such data becomes extremely difficult.

International standards like the Statistical Data and Metadata Exchange (SDMX) framework, maintained by international organisations including the UN, World Bank, and IMF, provide a starting

point (SDMX.org, 2023). But these need to be adapted and adopted at country level. Pakistan has made limited progress in this area so far.

The legal framework also needs attention. Laws governing privacy, data protection, and access to government records need to be reviewed and updated. Many countries have enacted specific legislation to enable the use of administrative data for statistics — for example, Australia's Data Availability and Transparency Act 2022. Pakistan needs similar legislative effort to provide legal clarity and protections for all parties.

As Floridi and Taddeo (2016) argued in their influential work on data ethics, "the governance of data is not just a technical matter but a deeply social and political one." This applies fully to Pakistan's situation.

Organisations and Institutional Arrangements

Data infrastructure does not operate in a vacuum. It is managed, maintained, and governed by organisations. These include statistical agencies (PBS and provincial bureaus), data-holding agencies (NADRA, FBR, BISP, etc.), oversight bodies, and coordination mechanisms.

Currently in Pakistan, these organisations operate largely in silos. There is no formal institutional mechanism for coordinating data access and sharing between PBS and other federal or provincial agencies. Each agency has its own mandate, its own systems, and its own concerns about sharing data.

This is not unique to Pakistan. Many countries faced similar fragmentation before undertaking reforms. The United Kingdom, for example, established the UK Statistics Authority as an independent body to oversee the statistical system and promote data sharing across government (UK Statistics Authority, 2007). Canada has the Canadian Centre for Data Development and Economic Research (CDER) which works to facilitate access to data from multiple sources for research.

Whatever organisational model Pakistan chooses, certain features are essential. There must be a clear mandate and legal authority. There must be accountability to the public. And there must be mechanisms for coordination that bring different data holders together rather than leaving each agency to operate on its own.

The institutional design must also recognise that data infrastructure is not only about government. Private sector, academia, civil society all have roles to play. A modern data infrastructure is a multi-stakeholder system. Its governance must reflect this reality.

Communities and Data Subjects

This component is often left out of technical discussions about data infrastructure, but it is among the most important. At the end of the day, much of the data we are talking about comes from people — citizens, households, businesses, communities. These are the data subjects.

Their trust is essential. If people do not believe that their data will be handled responsibly, they will resist providing it. Survey response rates will continue to decline. Public opposition to data linking will grow. And the whole system will lose credibility.

The OECD's Recommendation on Good Statistical Practice states that "public trust is a cornerstone of official statistics" and that statistical systems must operate with "transparency, accountability, and respect for individual privacy" (OECD, 2015). This is not abstract. It has direct implications for how data infrastructure is designed and communicated to the public.

In Pakistan, building this trust requires active effort. People need to understand what data is being collected about them, how it is being used, and what safeguards are in place. There should be clear, accessible communication — not just legal notices buried in government websites, but genuine public engagement.

Communities that are particularly vulnerable — low-income groups, minorities, women — may have additional concerns about how their data is used. The infrastructure must be designed with these concerns in mind. As Couldry and Mejias (2019) argued in "The Costs of Connection", data systems can reinforce existing inequalities if they are not designed with equity and inclusion as explicit goals.

Tying It All Together

Data infrastructure is not any single component described above. It is the way all of these components work together. Data assets need technology to be processed. Technology needs skilled people to operate it. People need governance frameworks to guide their work. Governance needs legal authority to be enforced. And all of it needs the trust and participation of the communities whose data makes the system possible.

Think of it like a road system. The road itself (data assets) is necessary, but so are traffic rules (governance), traffic police (organisations), drivers who know how to drive (people), road signs and markings (standards), and the communities that use the road and benefit from the commerce it enables. Remove any of these elements and the system breaks down.

Pakistan's challenge is to build this system in a coherent way. Not piece by piece in isolated projects, but as an integrated national infrastructure that serves the country's information needs for decades to come.

Think of It This Way Data infrastructure is to statistical information what road infrastructure is to commerce. It is the underlying system of assets, tools, rules, people, and institutions that makes the production and use of statistics possible. And like physical infrastructure, it gains value only through use.

References

1 Couldry, N. and Mejias, U. (2019). *The Costs of Connection: How Data is Colonizing Human Life and Appropriating It for Capitalism.* Stanford University Press.

1 DAMA International (2017). *DAMA-DMBOK: Data Management Body of Knowledge.* 2nd Edition. Technics Publications.

1 European Commission (2017). *New European Interoperability Framework.* Publications Office of the European Union.

1 Floridi, L. and Taddeo, M. (2016). "What is Data Ethics?" *Philosophical Transactions of the Royal Society A,* 374(2083).

1 Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* SAGE Publications.

1 ODI — Open Data Institute (2018). *What is Data Infrastructure?* Available at: https://theodi.org/

1 OECD (2015). *Recommendation of the Council on Good Statistical Practice.* OECD Legal Instruments.

1 SDMX.org (2023). *Statistical Data and Metadata Exchange: Technical Standards.* Available at: https://sdmx.org/

1 UK Statistics Authority (2007). *Code of Practice for Statistics.* London: UK Statistics Authority.

1 UNECE (2011). *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.* United Nations, Geneva.

1 UNSC — United Nations Statistical Commission (2019). *Global Assessment of the Modernisation of Statistical Production and Services.* Background document, 50th Session.

1 World Bank (2021). *World Development Report 2021: Data for Better Lives.* Washington, DC: World Bank.

The Case for Blended Data

The most significant shift in how national statistics can be produced in the 21st century is the move toward blended data. This means combining information from multiple sources — surveys, censuses, administrative records, private sector data, and digital data — to produce statistical outputs that are more timely, more detailed, and more accurate than any single source could provide on its own. The idea is not new in principle. Statistical agencies have always supplemented one data source with another when convenient. What is new is the scale at which this is now both possible and necessary.

The National Academies of Sciences, Engineering, and Medicine captured this shift in their landmark report on 21st century data infrastructure, arguing that "careful blending of data from multiple, complementary sources, such as statistical surveys and censuses, administrative agencies, and the private sector, offers a way to generate more detailed, timely, and useful statistical information than is currently available" (NASEM, 2023, p. 27). This is not merely a technical observation. It is a recognition that the traditional model — in which statistical agencies design, collect, and process their own data in relative isolation — is no longer sufficient to meet the information demands of modern societies.

For Pakistan, where the statistical system faces severe constraints of budget, coverage, and timeliness, the case for blended data is particularly compelling. But it is also particularly difficult to execute. The infrastructure, legal frameworks, and institutional habits needed to blend data from diverse sources are still in early stages of development. This chapter examines why blended data matters, what it requires, and what must change for it to work.

Why No Single Source Is Enough

Each data source has its own weaknesses, and these weakness are well documented in the statistical literature. Understanding them is the starting point for understanding why blending is necessary.

Surveys have historically been the backbone of national statistics. They provide structured, well-documented data collected using probability sampling methods. But surveys are under pressure everywhere. Response rates have been declining for decades across countries. The U.S. Current Population Survey, for instance, saw its response rate drop from over 95 per cent in the 1990s to below 85 per cent by the late 2010s. The American Community Survey fell from around 97 per cent to approximately 86 per cent over a similar period (NASEM, 2023, Table 2-1). Pakistan's household surveys face analogous challenges — security constraints in certain regions, limited field staff, poor infrastructure in remote areas, and respondent fatigue all contribute to growing nonresponse. Groves (2011) documented how survey research globally had entered what he called a "third era" characterised by falling participation, rising costs, and increasing reliance on supplementary sources.

Beyond response rates, surveys are expensive and slow. A typical household survey in Pakistan takes 12 to 18 months from fieldwork to publication, which limits its usefulness for policy decisions that require timely data. Sample sizes also impose limits on the granularity of estimates — most surveys cannot produce reliable district-level statistics for important indicators like poverty or unemployment. And as the NASEM panel noted, even when surveys achieve full coverage, they "are expensive, slow to produce information, and suffer from nonparticipation" (NASEM, 2023, p. 27).

Administrative records offer the advantage of large coverage and low marginal cost, since the data already exists as a byproduct of government operations. Tax records, social protection registries, civil registration, health facility records — these data cover large populations and are updated regularly. But they were not designed for statistical purposes. Wallgren and Wallgren (2007) emphasised this fundamental tension: administrative data reflects "administrative reality, not statistical reality." The definitions used in a tax system or a social programme may not match the concepts that statistical agencies need to measure. Coverage may be incomplete — Pakistan's tax base, for example, covers only a fraction of the working population. Data quality depends on the administrative process that generates it, which may vary across regions, change over time, or be affected by fraud and underreporting.

The U.S. experience illustrates both the promise and the limitations. The Census Bureau has used federal tax data in its economic censuses since the 1950s and in building its business register since the early 1970s (NASEM, 2023, p. 77). Yet even in that mature system, legal restrictions prevent the sharing of certain tax data between statistical agencies. The Internal Revenue Code still blocks the Census Bureau from sharing business tax data with the Bureau of Labor Statistics and the Bureau of Economic Analysis, despite repeated legislative attempts since 2002 to change this (NASEM, 2023, p. 78). If the United States, with its well-resourced institutional infrastructure, struggles with these barriers, the challenges for Pakistan are likely to be even more significant.

Private sector data — from telecom companies, banks, digital platforms, retailers — can be extraordinarily timely and granular. Transaction data can provide near-real-time indicators of economic activity. Mobile phone metadata has been shown to predict poverty levels at fine spatial resolution (Blumenstock, Cadamuro, & On, 2015). Scanner data from supermarkets has been used in the Netherlands to compile consumer price indices with greater accuracy and lower respondent burden than traditional price collection methods (Chessa, 2016). In the United States, all 13 designated federal statistical agencies except one were using private sector data assets at the time of the NASEM report, with the Bureau of Economic Analysis alone using some 142 private sector data sources (NASEM, 2023, p. 28).

But private sector data comes with serious complications. Companies have legitimate concerns about commercial confidentiality. Data quality is often unknown — as one workshop participant told the NASEM panel, private sector data is not like "gold dust" but rather like "sand, abundant and requiring significant effort to make them useful" (NASEM, 2023, p. 94). The data may not be representative of the population. Coverage depends on who uses the service, which typically skews toward wealthier, younger, and more urban populations. Data definitions can change without notice when companies update their systems. And data-use agreements tend to be one-off arrangements with no inherent sustainability (NASEM, 2023, p. 29).

Digital and emerging data sources — satellite imagery, social media, crowdsourced data, sensor data — add further possibilities but also further complications. These sources can be large in volume but noisy and difficult to interpret. They often lack the metadata needed to assess quality or fitness for statistical use.

No single source resolves all these weakness. But when carefully combined, different sources can compensate for each other's limitations. Survey data provide the statistical framework, quality controls, and conceptual definitions. Administrative data fill coverage gaps and reduce respondent

burden. Private sector data add timeliness and granularity. The result is a richer, more complete picture — one that none of these sources could produce alone.

What Blended Data Requires

Recognising the value of blended data is one thing. Actually producing it is quite another. The technical, institutional, and governance requirements are substantial, and they cut across multiple dimensions.

New Statistical Methods

Combining data from different sources is not straightforward, and it cannot be done by simply appending one dataset to another. Each source has its own coverage, its own definitions, its own error structure, and its own biases. Bringing them together requires a set of statistical methods that were not part of the traditional survey statistician's toolkit.

Lohr and Raghunathan (2017), in their comprehensive review in *Statistical Science*, identified four broad families of methods for combining data from multiple sources. The first is **record linkage** — matching individual records across datasets using identifying information like names, identity numbers, or addresses. This is the most direct form of data integration and has been used extensively in countries with well-developed statistical infrastructure. But as the NASEM panel noted, "linkage rates vary across studies and for subpopulations within studies" (NASEM, 2023, p. 96), and linkage requires that individual-level data with sufficient identifying information be available from each source. In Pakistan, where the CNIC number could serve as a universal linkage key, the potential for record linkage is considerable — but realising that potential depends on legal access to the relevant datasets.

The second family is **multiple frame methods**, in which independent samples from different sampling frames are combined to improve coverage or reduce costs. This is a technique with a long history in survey statistics (Lohr, 2021) and can be adapted to the multi-source environment by treating administrative registers or private datasets as additional frames.

Third, **imputation-based methods** treat the problem of combining sources as a missing data problem. Variables that are available in one dataset but not another can be imputed using statistical models that exploit the relationships among observed variables. Lohr and Raghunathan (2017) note that while imputation provides a transparent framework for combining incompatible sources, it requires considerable expertise and careful attention to the differences between data sources — differences in respondent types, interview modes, survey contexts, and measurement approaches.

Fourth, **modelling techniques** — including small area estimation, Bayesian hierarchical models, and machine learning approaches — can be used when direct linkage or imputation is not possible or appropriate. These methods have been used, for example, to combine satellite imagery with survey data to predict poverty at fine spatial resolution (Jean et al., 2016).

The NASEM report recommended that statistical systems "systematically coordinate agencies' efforts to blend multiple data sources" and "ensure that statistical agencies have the appropriate skills and expertise" for these methods (NASEM, 2023, p. 97). For Pakistan, where methodological capacity

within PBS and provincial bureaus is limited, investing in these skills is not optional — it is a prerequisite for any meaningful use of blended data.

New Statistical Designs

Once blended data become available, the design of statistical programmes itself can change. The traditional approach — in which a single survey is designed to capture all needed information — gives way to an approach in which different sources are assigned different roles, and survey designs are optimised to complement what other sources already provide.

This is perhaps the most exciting implication of blended data, and the one least discussed. As the NASEM panel put it, "after new blended statistics using multiple data sources are built, survey designs could likely be optimised, reducing original survey measurement in populations that are well measured and increasing survey measurement in populations not well covered by the various administrative record systems" (NASEM, 2023, p. 97). In practical term, this could mean smaller but more targeted surveys, focused on filling the specific gaps that administrative and private sector data cannot cover.

The U.S. Census Bureau's modernisation of its residential construction statistics programme illustrates this approach. Rather than collecting housing permit data from 9,000 permit-issuing organisations, the Bureau now receives data from third-party sources and supplements it with a small cutoff sample. Satellite imagery is used to identify construction starts instead of telephone interviews. The result is more granular statistics — permit data for every jurisdiction rather than just states — at lower cost (NASEM, 2023, p. 31).

For Pakistan, the implication is significant. The country cannot afford large-scale surveys at the frequency and granularity that modern policymaking demands. But if administrative data from agencies like NADRA, FBR, and BISP can be systematically accessed and combined with periodic surveys, the surveys themselves can be redesigned to be smaller, cheaper, and more focused — while the blended output is more comprehensive than either source alone.

New Capabilities

The NASEM panel drew on work by the United Nations Economic Commission for Europe (UNECE) to identify a set of capabilities that a modern data infrastructure would need for blending (NASEM, 2023, Box 4-4). These include the familiar ones — data transformation, security, and governance — but also several that are novel for many statistical agencies.

Data design, definition, and description for data not originally built for statistical analysis is perhaps the most critical capability gap. Administrative and private sector data typically lack the kind of metadata documentation that survey datasets carry. Variables may not be clearly defined. Coverage boundaries may not be documented. Without adequate metadata, meaningful blending is impossible. The NASEM panel was emphatic on this point: "to be responsibly discovered, combined, shared, used, and reused, data must be described. Limitations of data must also be readily accessible to ensure that biases in individual data assets do not ripple through any analysis" (NASEM, 2023, p. 94).

Data logistics — managing the supply chain of data from holders to users and back — is another capability that statistical agencies typically lack. In the 20th century, statistical agencies designed and collected their own data. The data sat behind the agency's firewall, under the agency's full control. In a blended data world, agencies must negotiate access to datasets held by other organisations, often on an ongoing basis, under conditions that respect the data holder's interests. This is fundamentally different from traditional data collection.

Data integration — the ability to link, combine, and align datasets from multiple sources — requires not just technical tools but also deep substantive understanding of what each dataset represents and where its limitations lie. This includes what the NASEM panel called **knowledge management**: documenting the meaning of individual measurements across diverse data assets so that analysts understand what they are combining (NASEM, 2023, p. 98).

The panel's assessment was blunt: "such data silos no longer serve the needs of modern society. Features of the 20th century data infrastructure must change" (NASEM, 2023, p. 98). This is equally true for Pakistan's statistical system, where data integration capabilities are still minimal across most agencies.

New Quality Frameworks

Blending data from multiple sources does not just add complexity — it fundamentally changes how quality must be assessed. Traditional quality frameworks were designed for single-source data, typically surveys, where concepts like sampling error, nonresponse bias, and measurement error had well-understood definitions and estimation methods. When multiple sources are combined, the quality of the blended output depends on the quality of each input, the method used to combine them, and the fitness of the result for its intended purpose.

The U.S. Federal Committee on Statistical Methodology (FCSM) addressed this challenge in its 2020 report, *A Framework for Data Quality*, which defined quality as "the degree to which data captures the desired information using appropriate methodology in a manner that sustains public trust" (FCSM, 2020, p. 6). The framework identifies 11 dimensions of data quality grouped across three domains: **utility** (relevance, accessibility, timeliness, punctuality, granularity), **objectivity** (accuracy, coherence, comparability), and **integrity** (credibility, transparency, confidentiality). Importantly, the framework applies to all data types — surveys, administrative records, blended data, and emerging sources — and emphasises fitness-for-purpose rather than adherence to any single standard of accuracy.

This is directly relevant for Pakistan. Any move toward blended data will require the development of quality assessment protocols that go beyond traditional survey error frameworks. The quality of an administrative dataset from NADRA or FBR cannot be evaluated using the same metrics as a probability survey. Different questions must be asked — about coverage completeness, definitional consistency, update frequency, and processing integrity. And when such data is combined with survey data, the quality of the blend must be assessed as a whole, not just source by source.

New Privacy Safeguards

Blending data from multiple sources increases the potential for re-identification of individuals. A dataset that is anonymised on its own may become identifiable when linked with another dataset. This is a well-known risk in the privacy literature, and it becomes more acute as the number and diversity of data sources increase.

The NASEM panel devoted considerable attention to this challenge, noting that "a 21st century national data infrastructure cannot succeed without ensuring ethical exchange of data; trust in institutions involved in data exchange; privacy-preserving techniques; and technical, organisational, and legal mechanisms supporting responsible data practices" (NASEM, 2023, p. 99). The panel identified four ethical values that must underlie data infrastructure: attention to how use of a subject's data affects their life; respect for autonomy and informed consent; concern for beneficence; and respect for human dignity.

On the technical side, advances in privacy-enhancing technologies offer new possibilities. Differential privacy, synthetic data generation, secure multiparty computation, and homomorphic encryption are all being explored in various countries. The U.S. Census Bureau began releasing data using synthetic data generation in 2006 as part of its Longitudinal Employer-Household Dynamics programme (NASEM, 2023, p. 101). Several U.S. agencies including NIH, NIST, and NSF are working on standards for homomorphic encryption to enable computation on encrypted data.

For Pakistan, where public trust in government handling of personal data is limited and legal frameworks for data protection are still developing, these safeguards are not secondary concerns — they are preconditions. Without credible privacy protections, neither citizens nor private sector data holders will support the data sharing that blended statistics require. The "Five Safes" framework (Desai, Ritchie, & Welpton, 2016) — safe projects, safe people, safe settings, safe data, safe outputs — offers a practical model for structuring access controls, one that several countries including the UK and Australia have adopted with success.

International Experience and Lessons

The move toward blended data is not happening in a vacuum. Many countries are already well advanced in this direction, and their experiences offer useful lessons for Pakistan.

Statistics Canada has adopted an "administrative data first" policy, meaning that it seeks to use existing administrative records before resorting to new data collection (NASEM, 2023, p. 80). Canada also developed the necessity and proportionality criteria for data intake — acquiring no more data than needed for the specified statistical purpose and considering the sensitivity and confidentiality of the data (Bowlby, 2021). This disciplined approach avoids the unbridled harvesting of all available data and focuses resources where they are most needed.

Statistics Netherlands has been a pioneer in using private sector data for official statistics. Its work on scanner data for consumer price indices, led by researchers like Chessa (2016), demonstrated that electronic transaction data from supermarkets could replace much of the traditional manual price collection. The Dutch approach shows that private sector data can not only supplement but in some cases improve upon traditional methods — but only after considerable investment in methodology and data quality assessment.

The United Kingdom created the UK Statistics Authority as an independent oversight body responsible for promoting and safeguarding official statistics that "serve the public good" (NASEM, 2023, p. 59). This kind of institutional accountability is important for building public trust when statistical systems move toward using more diverse and sensitive data sources.

For Pakistan, the lesson from international experience is not that it should replicate any single country's approach. Legal frameworks, institutional capacities, and data landscapes differ too much for that. The lesson is rather that blended data is not an abstract aspiration — it is a practical reality in many countries, and it is being achieved through a combination of legal reform, institutional investment, methodological development, and sustained partnership between statistical agencies and data holders. Pakistan can learn from these experiences while adapting them to its own context.

What Blended Data Does Not Mean

A few clarifications are important to avoid misunderstanding.

Blended data does not mean the end of surveys. Surveys will remain essential for measuring concepts that cannot be captured through administrative records or transactions — attitudes, perceptions, informal economic activity, unpaid care work, and many other topics. What changes is the role of surveys: from being the sole source to being one component in a multi-source system. As the NASEM panel recommended, blended data should lead to the redesign of surveys, not their abandonment.

Blended data also does not mean the unrestricted harvesting of all available digital data. The NASEM panel was explicit that "a new data infrastructure should not result in the unbridled harvesting of all digital data that exists in the country" (NASEM, 2023, p. 91). Data acquisition should be guided by necessity — the data must serve a pre-specified statistical purpose — and proportionality — the amount and detail of data acquired should be limited to what that purpose requires.

Finally, blended data does not reduce the need for quality assessment. If anything, it increases it. Each input source must be evaluated on its own terms, and the blended output must be assessed for fitness-for-purpose. Without rigorous quality frameworks, blended data risks producing statistics that appear comprehensive but rest on unexamined foundations.

A Critical Shift The NASEM panel's assessment is direct: "such data silos no longer serve the needs of modern society" (NASEM, 2023, p. 98). The features of 20th century data systems must change, and this demand enhanced capabilities across the board — in methods, technology, governance, and privacy protection. For Pakistan, where the statistical system still operate largely in the traditional single-source model, this shift represents both the biggest challenge and the biggest opportunity.

References

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076.

Bowlby, G. (2021). Private sector administrative data and the Canadian statistical system. Presentation to the National Academies' Panel on the Scope, Components, and Key Characteristics of a 21st Century Data Infrastructure, December 9, 2021.

Chessa, A. G. (2016). A new methodology for processing scanner data in the Dutch CPI. *Eurostat Review on National Accounts and Macroeconomic Indicators*, 1/2016, 49–69.

Desai, T., Ritchie, F., & Welpton, R. (2016). Five Safes: Designing data access for research. Working Paper, University of the West of England.

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality*. FCSM-20-04. September 2020.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5), 861–871.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794.

Lohr, S. L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47(2), 229–263.

Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2), 293–312.

National Academies of Sciences, Engineering, and Medicine. (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26688.

Wallgren, A., & Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. Chichester: John Wiley & Sons.

Key Data Holders in Pakistan

Building a new data infrastructure means first understanding who holds the data that can contribute to better national statistics. This might seem like a straightforward question. But in practice, Pakistan's data landscape is fragmented across dozens of organisations, each with its own mandate, its own systems, and its own reasons for collecting data. Some of these organisations produce data specifically for statistical purposes. Most do not. Yet all of them hold information that, if brought together thoughtfully, could transform the country's ability to understand itself.

The literature on national statistical modernisation consistently emphasises this point. As the UN Economic Commission for Europe noted in its guidelines on using secondary data sources, "the potential supply of data for official statistics extends far beyond what statistical offices themselves collect" (UNECE, 2011). The challenge is mapping this landscape, understanding what each holder brings to the table, and creating the conditions under which their data can be accessed and used.

Statistical Agencies

The Pakistan Bureau of Statistics (PBS) and the provincial Bureaus of Statistics are the primary producers of official statistics in the country. They conduct the population census, labour force surveys, household income and expenditure surveys, agricultural surveys, and a range of other data collection exercises. Their mandate is explicitly statistical — they collect data to produce estimates that inform public policy and planning.

This matters because it means their data follows established quality controls. Sampling frames are designed, questionnaires are tested, field work is supervised, and results go through validation and estimation procedures. The UN Fundamental Principles of Official Statistics, adopted by the General Assembly in 2014, lay out the standards that national statistical offices are expected to follow, including professional independence, scientific methods, and confidentiality of individual data (United Nations, 2014). PBS operates under these principles, at least in aspiration.

But PBS and provincial bureaus face well-known constraints. Their budgets are limited. The population census, which is the most comprehensive data collection exercise, happens only once a decade — and even then, it has faced significant delays in Pakistan's history. Between censuses, surveys provide useful but incomplete pictures. Labour force surveys, for instance, have sample sizes that do not allow for reliable estimates at the district level for many indicators. Household surveys typically take 12 to 18 months from fieldwork to publication, which limits their timeliness.

The declining quality of survey data is not unique to Pakistan. Groves (2011) documented how survey research globally has entered what he called a "third era" characterised by falling response rates, rising costs, and growing reliance on supplementary data sources. Pakistan's statistical agencies face all of these pressures, compounded by security challenges in certain regions, limited field staff, and weak infrastructure in remote areas.

Provincial Bureaus of Statistics deserve separate mention. Since the 18th Constitutional Amendment devolved significant responsibilities to provinces, provincial statistical capacity has become more important. But most provincial bureaus remain understaffed and under-resourced. Their relationship with PBS is also complicated — there is no uniform framework for how federal and provincial

statistical agencies coordinate data collection, share data, or harmonise standards.

Despite all these limitations, statistical agencies remain the backbone of the system. They have the mandate, the institutional knowledge, and the methodological expertise. What they lack is access to the wider data ecosystem that could supplement their surveys and make their estimates more timely, more granular, and more comprehensive.

Federal Administrative Agencies

Some of the richest data in Pakistan sits with federal agencies that collect it not for statistical purposes but as part of their normal administrative functions. This data is generated as a byproduct of government operations — registering citizens, collecting taxes, delivering social protection, regulating businesses.

NADRA (National Database and Registration Authority) is perhaps the most significant data holder in this category. It maintains the national identity database, which covers a very large proportion of Pakistan's adult population through the Computerised National Identity Card (CNIC) system. NADRA's data includes biometric information, addresses, family linkages, and demographic details. This database has been used for various purposes beyond its original registration mandate — including verification for BISP social protection, electoral rolls, and SIM card registration.

The potential of national identity systems for statistical purposes has been explored in other countries. Abraham et al. (2018) discussed India's Aadhaar system and its implications for governance, noting that universal identity databases can serve as a backbone for linking administrative records across agencies. Pakistan's CNIC system has similar potential, though it also raises significant privacy concerns that must be addressed.

FBR (Federal Board of Revenue) holds tax records for individuals and businesses. This includes income tax returns, sales tax data, customs records, and information from the withholding tax system. Tax data can provide detailed information about economic activity, income distribution, and the size of the formal economy. However, Pakistan's tax base is notoriously narrow — the number of registered taxpayers does not come close to covering the full working population — which limits coverage. Tax data also has well-known issues with underreporting and evasion, which statistical agencies would need to account for in any analysis.

BISP (Benazir Income Support Programme) maintains detailed records on social protection beneficiaries. Its National Socioeconomic Registry (NSER), which was compiled through a door-to-door poverty census, covers a very large number of households and contains information about demographics, assets, housing conditions, and vulnerability indicators. The NSER is one of the most comprehensive household-level databases in Pakistan and could be extremely valuable for producing poverty statistics and for calibrating survey estimates.

SECP (Securities and Exchange Commission of Pakistan) holds data on registered companies, corporate filings, financial statements, and securities markets. This data is useful for understanding the corporate sector, investment patterns, and economic concentration.

Other federal agencies that hold significant data include the State Bank of Pakistan (financial system data), Pakistan Telecommunication Authority (telecom subscriber data), NEPRA (electricity

consumption data), OGRA (oil and gas data), and various line ministries that maintain management information systems (MIS) for their programmes.

The common challenge across all these agencies is that their data was collected for specific operational purposes. The definitions, classifications, coverage, and quality controls were designed to serve those purposes — not statistical analysis. As Wallgren and Wallgren (2007), in their comprehensive guide on register-based statistics, noted, "administrative data reflect administrative reality, not statistical reality." Converting administrative data into statistical outputs requires careful assessment and adjustment. Coverage may not be universal. Definitions may not match survey concepts. Data quality may vary across regions or over time.

Despite these challenges, the volume and richness of administrative data in Pakistan is considerable. The key barrier is not the data itself but the lack of frameworks, legal provisions, and institutional mechanisms for accessing it for statistical purposes.

Provincial Governments

Pakistan's provinces hold enormous amounts of data across their various departments and agencies. Since the 18th Amendment, provinces have primary responsibility for health, education, agriculture, local governance, and many social services. This means that much of the data about how these services are delivered — and what outcomes they produce — sits at the provincial level.

Education departments in all four provinces and the federal areas maintain Education Management Information Systems (EMIS) that track school enrollment, attendance, teacher deployment, and infrastructure. Punjab and Sindh have invested more heavily in digitising these systems, while Balochistan and Khyber Pakhtunkhwa face greater challenges. The data can provide granular, real-time information about educational access that household surveys can only capture periodically.

Health departments collect data through the District Health Information System (DHIS), which tracks facility-based health indicators including disease surveillance, immunisation coverage, maternal health services, and facility utilisation. Pakistan adopted the DHIS2 platform, which is used in many developing countries, and this data is potentially very useful for producing health statistics between surveys. However, private health facilities, which account for a large share of healthcare delivery in Pakistan, are generally not covered.

Agriculture departments maintain crop reporting systems, livestock census data, and agricultural inputs data. Given that agriculture remains a large part of Pakistan's economy and employs a significant share of the workforce, this data is directly relevant for national accounts and food security analysis.

Revenue departments hold land records, property transaction data, and related information that is relevant for wealth estimation and economic analysis. Many provinces have been digitising their land records through projects like Punjab's Land Records Management Information System.

Social welfare departments, labour departments, local government bodies, and planning departments all generate data as part of their work. The problem is that this data is often not standardised, not easily accessible, and not shared with statistical agencies at either the provincial or federal level.

The fragmentation between provinces is also a significant issue. Each province may use different classifications, different software systems, and different data formats. There is no common framework for harmonising provincial data, which makes it very difficult to produce consistent national-level statistics from provincial sources.

This is not just a Pakistan problem. Groen (2012) documented similar challenges in the United States, where state-level administrative records needed significant harmonisation work before they could be used for federal statistical purposes. But it is a problem that Pakistan must solve if provincial data is to be leveraged effectively.

Private Sector

The private sector in Pakistan generates a massive volume of data that could, if made accessible under appropriate conditions, significantly improve the timeliness and granularity of national statistics. This is an area where international practice is evolving rapidly.

Telecommunications companies — Jazz, Telenor, Zong, Ufone — collectively serve the overwhelming majority of Pakistan's mobile subscribers. Their data includes call detail records, mobile money transactions (especially through platforms like JazzCash and Easypaisa), and location data derived from cell tower connections. This data can provide near-real-time indicators of population mobility, economic activity, and even disaster displacement. Blumenstock, Cadamuro, and On (2015) demonstrated in Rwanda that mobile phone metadata could be used to predict poverty levels at high spatial resolution, illustrating the statistical potential of telecom data.

Banks and financial institutions process millions of transactions daily. Transaction data can provide timely indicators of consumer spending, economic confidence, and financial inclusion. The State Bank of Pakistan already publishes aggregate financial statistics, but more granular data from individual banks — properly anonymised — could enhance economic measurement.

E-commerce and digital platforms — such as Daraz, Careem, Foodpanda, and various digital payment systems — generate data on consumer behaviour, service demand, pricing, and employment in the gig economy. These platforms are growing rapidly in Pakistan and their data could help capture economic activities that traditional surveys miss.

Large retailers, manufacturers, and agricultural commodity markets also hold data that could provide high-frequency economic indicators. International experience supports this. Statistics Canada, for instance, has developed programmes like the Canadian Survey on Business Conditions to collect high-frequency data from businesses across provinces, improving the timeliness of economic statistics considerably (Statistics Canada, 2020). The Dutch statistical office, Statistics Netherlands, has been using scanner data from supermarkets to compile consumer price indices since the early 2000s, reducing respondent burden while improving the accuracy and timeliness of price statistics (Chessa, 2016).

The American Economic Association Committee on Statistics (AEAStat) has also recognised the potential of high-frequency private sector data for modernising official statistics, calling for closer collaboration between statistical agencies and private data holders (Jarmin, 2019).

But private sector data comes with significant complications. Companies have legitimate concerns about commercial confidentiality. Sharing data with government agencies may expose proprietary business information. There are also questions about data quality, representativeness (not everyone uses digital services equally), and the stability of private data sources — companies can change their systems, change their data definitions, or simply go out of business.

Privacy is another major concern. Individual-level data from telecom or financial companies is highly sensitive. Any use of such data for statistical purposes must be governed by strict legal frameworks and technical safeguards. The "Five Safes" framework developed by Desai, Ritchie, and Welpton (2016) — safe projects, safe people, safe settings, safe data, safe outputs — provides a useful model for managing these risks. Several countries, including the United Kingdom and Australia, have adopted variations of this framework for managing access to sensitive data.

The key point is that private sector data is not a replacement for surveys or administrative records. It is a complement. Its value lies in its timeliness and granularity, but these advantages must be weighed against issues of access, quality, representativeness, and privacy.

Academic and Nonprofit Institutions

Universities, research organisations, think tanks, and international NGOs produce a substantial amount of data in Pakistan that can supplement official statistics. This data is often overlooked in discussions about national statistical systems, but it should not be.

Organisations like the Pakistan Institute of Development Economics (PIDE), the Lahore University of Management Sciences (LUMS), the Institute of Development and Economic Alternatives (IDEAS), and the Sustainable Development Policy Institute (SDPI) conduct surveys, compile datasets, and produce research that covers topics from poverty dynamics to urban development to labour market informality. International organisations like the World Bank, UNICEF, WHO, and FAO also fund and conduct data collection exercises in Pakistan — the Multiple Indicator Cluster Survey (MICS), Demographic and Health Survey (DHS), and various agricultural assessments being prominent examples.

These datasets sometimes fill gaps that official statistics do not cover. They may use innovative methodologies, cover hard-to-reach populations, or address emerging issues before statistical agencies catch up. Academic researchers also bring methodological expertise in areas like small area estimation, machine learning applications for statistics, satellite imagery analysis, and privacy-preserving techniques that may not yet be fully established within PBS.

Jean et al. (2016) published a widely cited study in *Science* demonstrating how satellite imagery combined with machine learning could predict poverty levels in African countries at fine spatial resolution. This kind of methodological innovation often originates in academic settings and can eventually be adopted by statistical agencies.

However, academic and NGO data has its own limitations. Sample sizes may be small. Methodologies may not be directly comparable with official standards. Data may not be collected regularly enough to provide trend information. And there are sometimes questions about data access and documentation — not all research datasets are publicly available or well-documented.

Strengthening the link between academic institutions and the national statistical system is important. This can take many forms: formal data-sharing agreements, joint research projects, secondments of researchers to statistical agencies, advisory committees, and collaborative capacity building. The goal is to create a relationship where academic innovation feeds into official statistical practice, and where official data is accessible to researchers who can add value.

Other Data Sources Worth Noting

Beyond the main categories above, there are several other data sources that deserve mention in Pakistan's context.

Satellite and geospatial data is increasingly important for statistics. Remote sensing can provide information on agricultural output, urbanisation, deforestation, flood damage, and nighttime economic activity. Pakistan already uses some satellite data for crop estimation, but the potential is much greater. The UN Global Pulse initiative has published several reports on how satellite and geospatial data can complement traditional statistics, particularly in developing countries (UN Global Pulse, 2012).

Social media and web data — from platforms like Twitter, Facebook, and Google Trends — have been explored as potential data sources for nowcasting economic indicators, tracking disease outbreaks, and measuring public sentiment. However, these sources are highly unrepresentative in the Pakistan context, where internet penetration and social media use vary dramatically by age, income, gender, and geography. They should be used with great caution and primarily as supplementary indicators, not as primary data sources.

Citizen-generated data — from mobile apps, crowdsourcing platforms, and community monitoring systems — is another emerging source. This data can be timely and locally relevant but typically lacks the quality controls and representativeness needed for official statistics.

Each of these sources has the same basic profile: potentially valuable, but with significant limitations that must be carefully managed.

The Limitations of Non-Statistical Data

It is important to be honest about the challenges. Other than statistical agencies which collect data specifically for statistical purposes, all other datasets have their own limitations when used for producing national statistics.

Administrative data was designed to serve operational needs, not statistical ones. Coverage may not be universal — FBR data, for example, only covers the formally registered tax base. Definitions and classifications may differ from those used in surveys. Data quality controls may be uneven, and errors that do not matter for administrative purposes may be significant for statistics. As Daas et al. (2015) documented in their work on using administrative data for official statistics in the Netherlands, "the fitness for statistical use of administrative data cannot be assumed but must be systematically assessed."

Private sector data raises questions of representativeness, stability, access, and privacy. Not everyone uses banks or mobile phones equally. Companies can change their data systems at any time. And sharing commercially sensitive or personally identifiable data with government agencies requires robust legal and technical safeguards.

Academic and NGO data may not be collected on a regular basis, may use non-standard methodologies, and may not cover the whole country.

Provincial data is often fragmented, with different provinces using different systems and standards.

Acknowledging these limitations is not a reason to avoid using non-statistical data. It is a reason to invest in the capabilities, legal frameworks, and quality assurance processes that make such use possible and trustworthy. The UNECE's suggested framework for the quality of big data provides a structured approach for assessing and managing these issues (UNECE, 2014).

The Incentive Problem

Perhaps the most fundamental challenge is not technical but institutional. Most data holders currently have no incentive to share their data for the common good. This applies to federal agencies, provincial governments, and private companies alike.

Government agencies worry about losing control over their data, about being exposed to criticism if the data reveals poor performance, and about the administrative burden of preparing data for external use. Privacy and legal concerns are also genuine — many agencies are unsure whether they are legally allowed to share their data.

Private companies worry about commercial confidentiality, competitive disadvantage, and regulatory risk. They also question what they get in return for sharing their data.

This incentive problem must be addressed head-on. A successful data infrastructure cannot rely on goodwill alone. It must create conditions where sharing is beneficial, safe, and legally clear for all parties. The World Bank's Data for Better Lives report (2021) emphasised this point, arguing that "the social contract around data" must ensure that data holders see tangible benefits from participation.

Data sharing is incentivised when all data holders enjoy tangible benefits valuable to their missions. For government agencies, this might mean getting access to better statistics for their own planning and evaluation. For private companies, it might mean access to aggregated benchmarking data or recognition as responsible corporate citizens. For all parties, the legal framework must provide clear protections and obligations, so that data sharing is not a discretionary favour but a structured, predictable process.

The societal benefits of better national statistics — improved policy, better targeting of resources, more effective governance, enhanced accountability — are ultimately proportionate to the costs and risks of data sharing. But these benefits need to be communicated clearly and consistently. People and organisations share data when they trust the system and understand why it matters.

The Incentive Challenge Most data holders currently have no incentive to contribute or share their data for the common good. A successful data infrastructure must address this by making data sharing beneficial, safe, and legally clear for all parties. Data sharing is incentivised when all data holders enjoy tangible benefits valuable to their missions

and when societal benefits are proportionate to possible costs and risks.

References

1 Abraham, R., Bennett, E.S., Sen, N., and Shah, N.B. (2018). "State of Aadhaar Report 2017-18." *IDinsight Working Paper.*

1 Blumenstock, J., Cadamuro, G., and On, R. (2015). "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science,* 350(6264), pp. 1073-1076.

1 Chessa, A.G. (2016). "A New Methodology for Processing Scanner Data in the Dutch CPI." *Eurostat Review on National Accounts and Macroeconomic Indicators,* 1/2016, pp. 49-69.

1 Daas, P.J.H., Puts, M.J., Buelens, B., and van den Hurk, P.A.M. (2015). "Big Data as a Source for Official Statistics." *Journal of Official Statistics,* 31(2), pp. 249-262.

1 Desai, T., Ritchie, F., and Welpton, R. (2016). "Five Safes: Designing Data Access for Research." *Economics Working Paper Series,* University of the West of England.

1 Groen, J.A. (2012). "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures." *Journal of Official Statistics,* 28(2), pp. 173-198.

1 Groves, R.M. (2011). "Three Eras of Survey Research." *Public Opinion Quarterly,* 75(5), pp. 861-871.

1 Jarmin, R.S. (2019). "Evolving Measurement for an Evolving Economy: Thoughts on 21st Century US Economic Statistics." *Journal of Economic Perspectives,* 33(1), pp. 165-184.

1 Jean, N., Burke, M., Xie, M., Davis, W.M., Lobell, D.B., and Ermon, S. (2016). "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science,* 353(6301), pp. 790-794.

1 Statistics Canada (2020). *Canadian Survey on Business Conditions.* Statistics Canada. Available at: https://www.statcan.gc.ca/

1 UN Global Pulse (2012). *Big Data for Development: Challenges and Opportunities.* United Nations Global Pulse.

1 UNECE (2011). *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.* United Nations, Geneva.

1 UNECE (2014). *A Suggested Framework for the Quality of Big Data.* UNECE Big Data Quality Task Team, United Nations Economic Commission for Europe.

1 United Nations (2014). *Fundamental Principles of Official Statistics.* General Assembly Resolution A/RES/68/261.

1 Wallgren, A. and Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes.* John Wiley & Sons.

1 World Bank (2021). *World Development Report 2021: Data for Better Lives.* Washington, DC: World Bank.

Core Principles for a Modern Data Infrastructure

It is tempting to jump straight to the technical architecture — the databases, the linkage systems, the access platforms. But the international experience makes clear that the most consequential decisions in building a data infrastructure are not technical. They are about values. What is the system for? Who does it serve? What does it protect? And who is accountable when things go wrong? Without clear answers to these questions, even well-designed technical systems tend to drift toward purposes they were not built for, or collapse under the weight of public mistrust.

The World Bank's 2021 World Development Report made this argument forcefully. It called for a "new social contract for data" — one grounded in the principles of value, trust, and equity — arguing that technical systems alone cannot generate the public legitimacy that data infrastructure requires to function (World Bank, 2021). A similar argument runs through the OECD's Recommendation on Enhancing Access to and Sharing of Data, which emphasises that governance frameworks must come before data flows, not after (OECD, 2021). And the UN Fundamental Principles of Official Statistics, adopted by the General Assembly in 2014, remain the most authoritative statement of what statistical systems owe to the societies they serve — including professional independence, scientific rigour, and strict confidentiality of individual data (United Nations, 2014).

Pakistan's challenge is to translate these international commitments into a set of operating principles that reflect local realities. The country does not have a comprehensive data protection law. It has no statutory framework for sharing administrative data across agencies for statistical purposes. Its statistical agencies lack the institutional authority to demand cooperation from data holders. And public trust in government handling of personal data is fragile, shaped by decades of experience in which citizen data was used more for surveillance and control than for service delivery. These conditions do not make principles irrelevant — they make them essential. Principles are not decorative statements of aspiration. They are the load-bearing walls of the infrastructure. Everything else — the technology, the governance, the legal reform — must be built on top of them.

Starting from People

The first commitment of any data infrastructure must be to the individuals whose data it uses. This is not just an ethical requirement. It is a practical one. If people do not trust that their information will be handled responsibly, they will resist providing it — whether through survey nonresponse, avoidance of administrative registration, or political opposition to data sharing arrangements. Trust is the foundation, and without it the infrastructure has nothing to build on.

The UN Fundamental Principles are explicit on this point. Principle 6 states that "individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes" (United Nations, 2014). This is not a suggestion. It is the bedrock compact between a statistical system and the people it measures. The data you give us will not be used to identify you, tax you, prosecute you, or embarrass you. It will be used only to produce aggregate statistics that help society understand itself.

In Pakistan, this compact is yet to be established in law or in practice. When NADRA shares CNIC data for voter verification or SIM registration, it is not a statistical use. When BISP's poverty registry is used to identify and exclude specific households from benefits, that is an administrative use, not a

statistical one. There is nothing inherently wrong with these uses. But they are fundamentally different from what a statistical data infrastructure is meant to do. And if the public cannot tell the difference — if they believe that data shared for one purpose will inevitably leak into another — then the entire infrastructure loses credibility before it begins.

The privacy literature offers useful frameworks for thinking about this. Helen Nissenbaum's theory of contextual integrity argues that privacy is not about keeping information secret. It is about information flowing appropriately — according to the norms and expectations that apply in a given context (Nissenbaum, 2004, 2010). When a citizen gives their address to NADRA for an identity card, the expectation is that it will be used for identification. When the same address shows up in a statistical poverty map, a different set of norms applies. Both uses may be legitimate, but the rules governing them — consent, access, purpose limitation — must be different. A data infrastructure that fail to maintain these distinctions will lose the trust it depends on.

What this means in practical term is that the infrastructure must establish clear boundaries around statistical use. Individual-level records should be accessible only to authorised analysts working on approved statistical projects. Outputs should be aggregated and checked for disclosure risk before release. And the entire system should be governed by a principle of data minimisation — acquiring only the data items needed for a specific statistical purpose, at the level of detail that purpose requires, and retaining them only as long as necessary. Statistics Canada's approach of "necessity and proportionality" — assessing whether the sensitivity and volume of data requested is proportionate to the statistical purpose — offers a sensible model for how this discipline can work in practice (Bowlby, 2021).

Purpose: Statistics as a Public Good

A data infrastructure can be built for many purposes — intelligence gathering, commercial exploitation, surveillance, or service delivery. The one proposed here has a specific purpose: the production of statistics for the common good. This needs to be stated clearly, because in the absence of a clear purpose, mission creep is inevitable.

Statistics are different from other information products. They describe populations and patterns, not individuals. They are meant to inform rather than to identify, to measure rather than to monitor. And critically, their value increases when they are shared. Unlike a commercial dataset whose value depends on exclusive access, a statistical estimate of district-level poverty becomes more useful the more widely it is used — by planners, researchers, journalists, civil society, and citizens. This is what economists call a public good: non-rivalrous and non-excludable in consumption.

The UN Fundamental Principles of Official Statistics capture this clearly. Principle 1 states that official statistics are "an indispensable element in the information system of a democratic society" and must be compiled and made available "on an impartial basis" to serve the government, the economy, and the public (United Nations, 2014). The outputs of the infrastructure, in other words, must be open. This is not simply a matter of posting datasets on a website. It means designing the entire production chain — from data acquisition to statistical estimation to dissemination — with the understanding that the end product is a public resource. User needs should drive what statistics are produced, when, and at what level of detail.

For Pakistan, this principle has direct implications. The country's statistical outputs are often delayed, difficult to access, and inadequately documented. Even when surveys are completed, the results are sometimes withheld or released only partially. The 2017 population census results, for example, generated significant controversy over delayed and partial publication. A data infrastructure built on the principle that statistics serve the common good would require full and timely release of statistical products, with adequate documentation for users to assess their quality and limitations. It would also require statistical agencies to actively engage users — researchers, policy makers, civil society organisations — in defining information priorities, rather than producing statistics in isolation and hoping someone finds them useful.

Balancing Access and Protection

Perhaps the most difficult practical challenge in building a data infrastructure is managing the tension between two legitimate goals: expanding access to data for statistical and research purposes, and protecting the confidentiality of the individuals and organisations whose data is being used. Every country that has built a serious data infrastructure has had to navigate this tension, and no country has done so perfectly.

The "Five Safes" framework, developed by Desai, Ritchie, and Welpton (2016) and now widely adopted in countries including the United Kingdom, Australia, and Canada, provides a structured way of thinking about this balance. It breaks the problem into five components: safe projects (is the use of data appropriate?), safe people (are the researchers trusted and trained?), safe settings (is the physical and digital environment secure?), safe data (has the data been treated to reduce identification risk?), and safe outputs (have the results been checked for disclosure?). The framework's strength is that it avoids all-or-nothing thinking. Rather than choosing between maximum access and maximum protection, it allows controls to be calibrated across multiple dimensions. A highly sensitive dataset can be made available if the project is well-justified, the researchers are accredited, the access environment is secure, and the outputs are reviewed.

For Pakistan, this framework is particularly useful because it acknowledge that different data assets carry different risks. Aggregated economic data from the State Bank of Pakistan poses little disclosure risk and can be shared widely. Individual health records from DHIS2, on the other hand, require much stricter controls. A population census microdata file sits somewhere in between — valuable for research but requiring careful anonymisation and access management. The infrastructure must be able to apply different levels of protection to different datasets, rather than treating all data the same.

The OECD's recommendation on data access makes a complementary point: that restrictions on access should be proportionate to demonstrated risks, not based on blanket prohibitions (OECD, 2021). Many data holders, in Pakistan and elsewhere, default to restricting access entirely — not because they have assessed the risks and found them unacceptable, but because saying no is easier than managing access responsibly. An infrastructure built on sound principles would reverse this default. The starting point would be that relevant data should be accessible for approved statistical purposes unless specific risks justify restriction — and those risks must be identified, documented, and mitigated using appropriate safeguards.

Governance That Works Across Boundaries

One of the hardest questions in any multi-source data infrastructure is governance. Who decides which data is included? Who approves access requests? Who monitors compliance? Who resolves disputes between data holders and data users? And critically, who is accountable when something goes wrong?

In a single-agency statistical system, these questions have relatively simple answers. The bureau of statistics makes the decisions, manages the data, and bears the accountability. But in a blended data environment — where data flows from NADRA, FBR, BISP, provincial departments, and private companies into a shared infrastructure — the governance challenge is fundamentally different. No single agency owns all the data. No single set of rules covers all the situations. And the incentives of different data holders are often misaligned.

The international literature on data governance has converged on several key insights. The British Academy and Royal Society, in their joint report on data management and use, argued that governance frameworks must be "as much about building relationships and trust between organisations as about establishing rules and regulations" (British Academy & Royal Society, 2017, p. 6). The World Bank's WDR 2021 called for "integrated national data systems" that explicitly build data production, protection, exchange, and use into planning and decision-making, and that bring diverse stakeholders into the governance structure (World Bank, 2021). And the U.S. Commission on Evidence-Based Policymaking recommended the creation of a dedicated coordination entity to facilitate data access across agencies while maintaining strict privacy protections (CEP, 2017).

What these frameworks share is a recognition that governance cannot be bolted on after the infrastructure is built. It must be designed in from the beginning. For Pakistan, this means establishing a governance body that includes representation from key data holders, statistical agencies, provincial governments, and independent experts. It means developing common standards for metadata, data quality, and interoperability — so that data from different sources can actually be combined. And it means creating clear processes for requesting access, approving projects, monitoring use, and sanctioning violations.

Interoperability deserve special emphasis here. One of the most basic obstacles to blending data in Pakistan is that different agencies use different classification systems, different identifier formats, different geographic coding, and different data structures. Without common standards, even willing agencies cannot share data effectively. The General Statistics Law of many countries require the national statistical office to set and enforce data standards across the statistical system. Pakistan has no equivalent provision. Building this function into the infrastructure governance framework is an essential early step.

The Legal Foundation

Everything described above — the privacy protections, the purpose limitations, the access frameworks, the governance structures — ultimately depends on law. Without a clear legal basis, the infrastructure cannot compel data sharing, cannot enforce privacy protections, cannot protect the statistical independence of the agencies that operate it, and cannot offer data holders the legal certainty they need to participate.

Pakistan's current legal framework was not designed for this. The General Statistics (Reorganisation) Act of 2011 gives PBS authority to collect statistics and coordinate the national statistical system, but its provisions on data sharing across agencies are vague at best. There is no comprehensive data protection legislation comparable to the European Union's General Data Protection Regulation or even India's Digital Personal Data Protection Act of 2023. The Electronic Transactions Ordinance of 2002 addresses certain aspects of electronic data but is not a data protection law in any meaningful sense.

The legal gaps are not trivial. Without statutory authority, PBS cannot compel NADRA, FBR, or other agencies to share data for statistical purposes. Without a data protection law, there is no legal standard against which privacy safeguards can be measured. Without legal provisions for accrediting researchers and governing data access, there is no basis for operating research data centres or managing controlled access to microdata. And without legal protections for the statistical independence of data infrastructure operators, there is a risk that political pressure could influence which statistics are produced or how they are released.

International experience suggests that legal reform is not a one-time event but an evolving process. The United States has passed multiple pieces of legislation over several decades to build its data infrastructure legal framework — from the Privacy Act of 1974 to the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 to the Foundations for Evidence-Based Policymaking Act of 2018 (U.S. Congress, 2019). Even so, significant legal obstacles remain, including restrictions on sharing tax data between statistical agencies that have resisted reform for over two decades (NASEM, 2023). Pakistan will need its own legislative journey, but it can learn from these experiences.

At a minimum, the legal framework should establish the statistical purpose doctrine — that data acquired through the infrastructure can only be used for statistical purposes and cannot be disclosed in identifiable form. It should provide clear authority for designated agencies to access administrative data held by other government bodies for approved statistical uses. It should establish legal protections for confidentiality that carry penalties for violations. And it should create the statutory basis for an independent governance body with authority to manage the infrastructure.

Transparency as an Operating Principle

Transparency in a data infrastructure is not about making everything open. Much of the data in the system will be confidential, and rightly so. What transparency means here is something more specific: that the rules governing the infrastructure are public, that decisions about access are documented and reviewable, that the methods used to produce statistics are disclosed, and that the public can see what data is being used, by whom, and for what purposes.

This kind of transparency serve multiple functions. It allows external scrutiny, which helps identify problems before they become crises. It builds public trust, because people are more willing to accept data use when they can see how it is governed. It enables accountability, because decision-makers can be held responsible for the choices they make. And it disciplines the system itself — when officials know their decisions will be reviewed, they tend to make better ones.

The UN Fundamental Principles make a related point. Principle 3 states that statistical agencies must "present information according to scientific standards on the sources, methods, and procedures of the

statistics" (United Nations, 2014). This is a commitment to methodological transparency — to letting users understand how the numbers were produced so they can assess their reliability. In a blended data environment, where statistics are produced by combining data from multiple sources using complex methods, this commitment is more important than ever. If a poverty estimate is produced by combining household survey data with satellite imagery and mobile phone records, the statistical agency must disclose how each source contributed to the estimate, what assumptions were made, and what the limitations are. Without this, users cannot make informed judgments about the quality of what they are being given.

Designing for Adaptation

A final principle concerns time. The data landscape is changing rapidly — new sources emerge, new methods are developed, new privacy threats are discovered, and new policy needs arise. An infrastructure designed to be permanent and fixed will quickly become obsolete. An infrastructure designed to adapt will remain useful.

This is easy to say but difficult to build. Adaptation requires investment in people, not just technology. It means statistical agencies need staff who are trained in modern data science, who can evaluate new data sources, who can apply new linking and modelling techniques, and who can assess new privacy risks. It means governance frameworks must be flexible enough to accommodate new types of data and new access arrangements without requiring legislative amendment every time. And it means the infrastructure must have mechanisms for learning — from its own experience, from international peers, and from the academic research community.

The OECD has emphasised this point in its digital government strategies, arguing that data governance frameworks should be "flexible and adaptive" rather than rigid and prescriptive (OECD, 2021). The NASEM panel similarly recommended that the U.S. data infrastructure should "incorporate state-of-the-art practices" and be designed for "continuous improvement" (NASEM, 2023). For Pakistan, where institutional reform tends to move slowly and where the gap between current capacity and international best practice is large, the temptation will be to design the infrastructure for today's reality. But the infrastructure that is needed is one that can grow — that can start with what is feasible now and evolve as capacity, trust, and legal frameworks develop.

Summary of Principles The principles that should guide Pakistan's data infrastructure can be summarised in seven commitments: 1. **People first.** The infrastructure must protect individuals from harm, preserve confidentiality, and earn public trust through responsible data practices. 2. **Statistical purpose.** Data accessed through the infrastructure must be used exclusively for producing statistics and evidence that serve the common good. 3. **Proportionate access.** Data sharing should be governed by a framework that balances access for approved purposes against protection of legitimate interests, calibrating controls to assessed risk. 4. **Clear legal authority.** The infrastructure must have a statutory basis that establishes its purpose, its powers, its governance, and the protections it affords to data subjects and data holders. 5. **Effective governance.** An inclusive governance framework with common standards, clear processes, and independent oversight must be in place before data begins to flow. 6. **Transparency.** The rules, methods, and decisions governing the infrastructure must be open to public scrutiny and external review. 7. **Continuous improvement.** The infrastructure must be designed for adaptation — capable of incorporating new data sources, new methods, and new safeguards as they become available.

References

Bowlby, G. (2021). Private sector administrative data and the Canadian statistical system. Presentation to the National Academies' Panel on the Scope, Components, and Key Characteristics of a 21st Century Data Infrastructure, December 9, 2021.

British Academy & Royal Society. (2017). *Data Management and Use: Governance in the 21st Century*. London.

Commission on Evidence-Based Policymaking. (2017). *The Promise of Evidence-Based Policymaking*. Washington, DC.

Desai, T., Ritchie, F., & Welpton, R. (2016). Five Safes: Designing data access for research. Working Paper, University of the West of England.

National Academies of Sciences, Engineering, and Medicine. (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good*. Washington, DC: The National Academies Press.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1), 119–158.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford University Press.

OECD. (2021). *Recommendation of the Council on Enhancing Access to and Sharing of Data*. Paris: OECD Publishing.

United Nations. (2014). Fundamental Principles of Official Statistics. General Assembly Resolution A/RES/68/261.

U.S. Congress. (2019). Foundations for Evidence-Based Policymaking Act of 2018. Public Law 115-435.

World Bank. (2021). *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank.

Repositioning PBS: From Data Collector to National Coordinator

The Pakistan Bureau of Statistics was created in 2011 through the General Statistics (Reorganisation) Act, which merged the Federal Bureau of Statistics, the Population Census Organisation, and the Agricultural Census Organisation into a single entity. The intent was to consolidate the country's fragmented statistical functions under one roof and improve both efficiency and coordination. Over a decade later, the consolidation has happened on paper, but the deeper transformation has not. PBS still operates primarily as a data collection agency. It conducts surveys and censuses, processes data, and publishes results. That work is essential and should continue. But it is no longer enough.

The reason it is not enough has to do with the changing nature of national statistics itself. As earlier chapters of this document have argued, the future of statistics lies in blending data from multiple sources — surveys, administrative records, private sector data, digital sources — to produce outputs that are more timely, granular, and comprehensive than any single source can deliver. This is not a task that any one agency can accomplish alone. It requires coordination across dozens of data holders, common standards that make data interoperable, quality frameworks that apply across the entire system, and governance arrangements that build trust among organisations with very different mandates and cultures. Someone has to play the coordinating role. In most countries that have modernised their statistical systems, that role falls to the national statistical office. In Pakistan, it should fall to PBS.

The UNECE Task Force on Data Stewardship, reporting in 2024, examined exactly this question — what role national statistical offices should play in the expanding data ecosystem. Their conclusion was clear: the transformation required is "an overall paradigm shift or evolution of the framework that guides how NSOs operate, moving from the production of statistics to the provision of data and data-related services" (UNECE, 2024, p. 5). Data stewardship, in this framing, means managing and coordinating interactions across the data system, "operating in service of — rather than in control of — the data ecosystem" (UNECE, 2024, p. 8). This is an important distinction. The goal is not for PBS to become a centralised data warehouse that owns all data. The goal is for PBS to become the institution that makes the system work — by setting standards, facilitating access, ensuring quality, and building the trust that data sharing requires.

What PBS Is Today

To understand what needs to change, it helps to be honest about where things stand. PBS is headed by the Chief Statistician, who functions as the principal executive officer and reports to the Ministry of Planning, Development and Special Initiatives. The bureau's mandate under Section 3 of the 2011 Act covers the collection, compilation, analysis, and publication of statistics across economic, social, demographic, and environmental domains. It conducts the Population and Housing Census, the Labour Force Survey, the Household Integrated Economic Survey, the Pakistan Social and Living Standards Measurement Survey, and various other data collection exercises.

These are significant responsibilities, and PBS carries them out under considerable constraints. Budgets are limited. Field infrastructure is weak in many regions. The gap between data collection

and publication is often 12 to 18 months or more. And the 2011 Act, while granting PBS operational autonomy, does not give the Chief Statistician the kind of system-wide authority that would be needed to coordinate data activities across other federal agencies, let alone provincial governments and the private sector.

This last point is critical. In many countries, the chief statistician has explicit legal authority not only to lead the national statistical office but also to coordinate the broader national statistical system. The UNECE Guidance on Modernising Statistical Legislation recommends that the chief statistician should "lead the strategic development" of the national statistical system and have authority over methods, standards, and procedures applied by all producers of official statistics (UNECE, 2018). In Canada, the Statistics Act gives the Chief Statistician personal responsibility for protecting confidentiality of individual records and full authority over programme priorities, including the power to approve all data collection forms used across government for statistical purposes (Statistics Canada, 2016). In Australia, the Australian Statistician has a legislated function "to provide statistical services for the state and territory governments" and an explicit leadership role in maximising the use of public data for statistical purposes (ABS, 2020).

Pakistan's Chief Statistician has no comparable authority. The 2011 Act does not give PBS the power to set standards that bind other agencies. It does not give the Chief Statistician authority to approve or review data collection instruments used by line ministries. And it certainly does not give PBS the legal basis to acquire administrative data from NADRA, FBR, BISP, or other federal agencies for statistical purposes — at least not with the kind of compulsory authority that statistical offices in Canada, Australia, and the Nordic countries enjoy. Without this authority, coordination depends entirely on goodwill and ad hoc arrangements, which are inherently fragile.

The Stewardship Model

The concept of data stewardship offers a useful framework for thinking about what PBS could become. As defined by the UNECE Task Force, a data steward is an entity that "manages data on behalf and for the benefit of the whole society" (UNECE, 2024). The stewardship role is distinct from both data ownership (holding data) and data production (creating statistics). It is about managing the system that connects them.

In practical term, data stewardship involves several interconnected functions. The UNECE framework identifies at least four broad areas where national statistical offices can operate as stewards: governance and standards-setting, quality assurance across the system, data integration and infrastructure, and advisory services to other data holders (UNECE, 2024). Not every NSO needs to perform all of these functions, and the scope will depend on the institutional context. But for Pakistan, where the data landscape is fragmented and no other institution is well positioned to play this role, PBS is the natural candidate.

This does not mean that PBS needs to do everything at once. The UNECE report is careful to emphasise that NSOs should proceed "at their own pace and take on a stewardship role that fits their purpose and environment" (UNECE, 2024). For a statistical office like PBS, which currently has limited capacity and no legal mandate for system-wide coordination, the transition would need to be gradual. But the direction should be clear from the start.

What does stewardship look like in countries that are further along? Australia offers one instructive example. The Australian Bureau of Statistics became the country's first Accredited Integrating Authority in 2012, which gave it formal responsibility for managing high-risk data integration projects involving Commonwealth government data. Under this framework, ABS does not hold all the data centrally. Instead, it acts as the trusted intermediary — linking datasets from health, education, taxation, social services, and census records into integrated assets like the Person Level Integrated Data Asset (PLIDA), while maintaining strict privacy controls and governance oversight through a multi-agency project board (ABS, 2020). The model is built on the principle that the statistical office's credibility — its long track record of protecting confidentiality and maintaining methodological rigour — is what makes it the natural home for this coordination role.

Canada provides another model. Statistics Canada has developed an Administrative Data Inventory — a central repository of metadata on all administrative datasets received by the agency — to understand what data it holds, how each dataset is used, and where there are opportunities for better integration (Trépanier, 2013). This sounds like a mundane administrative task. It is not. An inventory is the foundation of coordination, because you cannot manage, standardise, or integrate data assets you do not know exist. Canada has also adopted an "administrative-data-first" policy, meaning that statistical programmes must first investigate whether existing administrative or alternative data sources can meet their needs before designing a new survey (Statistics Canada, 2019). This policy reduces respondent burden, saves costs, and gradually shifts the agency's culture from one centred on primary data collection to one centred on data management and integration.

Knowing What Exists: The Inventory Function

One of the most basic but most neglected functions of a national coordinator is maintaining a comprehensive inventory of the country's data assets. In Pakistan, no such inventory exists. PBS knows what data it collects itself. But it has no systematic understanding of what data is held by NADRA, FBR, BISP, the State Bank, provincial health departments, NEPRA, PTA, or the dozens of other agencies and organisations that generate data as part of their operations.

This is not an unusual situation for developing countries. A PARIS21 survey of 59 national statistical offices in Africa, Latin America, and Asia-Pacific found that lack of information about existing data holdings was one of the most commonly cited barriers to coordination (PARIS21, 2021). You cannot share what you do not know you have. And agencies that do not know what data exists elsewhere in government will inevitably design new surveys to collect information that already exists in administrative systems — wasting resources and placing unnecessary burden on respondents.

Building an inventory is not a technology project. It is an institutional project. It requires someone — a dedicated working group or team — to systematically contact data-holding agencies across federal and provincial government, document what datasets they hold, what variables they contain, how they are collected and updated, what identifiers they use, what quality controls are applied, and under what legal authority the data is held. The Canadian experience suggests that this work should be led by a centralised function within the statistical office but involve extensive consultation with data-holding agencies, because much of the relevant knowledge resides with programme managers and IT staff in those agencies, not in any central register (Trépanier, 2013).

For Pakistan, this effort should be led jointly by the Chief Statistician and the Chief Economist, working with an inter-agency Data Inventory Working Group that include representatives from all major data-holding bodies. The inventory should cover federal agencies as well as provincial bureaus of statistics and key provincial departments. It should document not only the content of each dataset but also its potential fitness for statistical use — including coverage, frequency of update, granularity, and any known quality issues. The inventory itself should be treated as a living document, updated regularly as new data sources emerge and existing ones change.

The result would be, for the first time, a comprehensive map of Pakistan's national data landscape. This map is a precondition for everything else — for identifying opportunities for data blending, for assessing where gaps exist, for prioritising investments in data quality improvement, and for designing the governance arrangements that will govern data sharing.

Quality Across the System

When PBS produces a survey estimate, it can assess and report on the quality of that estimate because it controls the entire production process — from sample design to fieldwork to estimation. But in a blended data environment, many of the inputs to statistical production will come from outside PBS. Administrative data from tax authorities, health systems, or civil registration have their own quality characteristics — and their own quality problems. If PBS is to coordinate a system that blends these inputs, it must also take on a quality assurance function that extends beyond its own products.

This is a significant expansion of scope. Traditionally, statistical offices have applied quality frameworks only to their own outputs. The Federal Committee on Statistical Methodology (FCSM) in the United States published a comprehensive framework in 2020 that identifies 11 quality dimensions across three domains — the data source, the statistical process, and the statistical product (FCSM, 2020). But even the FCSM framework was primarily designed for data that statistical agencies control. The quality assessment of administrative data, private sector data, and digital sources requires additional considerations — including fitness for purpose, representativeness relative to the target population, stability over time, and sensitivity to changes in the administrative process that generates the data.

The UNECE has been developing guidance on quality frameworks for multi-source statistics, recognising that blended data environments require what might be called "input quality assessment" — evaluating the quality of data before it enters the statistical production process, not just after the final output is produced (UNECE, 2024). For PBS, this would mean developing a set of quality criteria that can be applied to any data source that enters the national statistical system, regardless of its origin. These criteria should cover at minimum the standard dimensions: relevance, accuracy, timeliness, accessibility, coherence, and interpretability. But they should also include dimensions specific to non-survey data, such as coverage relative to the target population, consistency of definitions with statistical concepts, and stability of the data-generating process.

In practical term, this function could begin with a Quality Assessment Working Group — a technically focused body that develops quality standards, creates assessment tools and templates, and works with data-holding agencies to evaluate the fitness of their data for statistical use. This group should draw on international frameworks, including the FCSM framework and the European Statistical System's quality guidelines, but adapt them to Pakistan's context. The key principle is that quality

assessment should not be a one-time exercise. It should be an ongoing function that monitors the quality of data inputs as the administrative processes that generate them evolve.

Making It Worth Their While

Perhaps the single most important insight from international experience is that coordination cannot be imposed from above alone. It must also be incentivised. Data-holding agencies will share their data with PBS, and tolerate the costs and risks of doing so, only if they see tangible value in return. This is what distinguishes a successful data infrastructure from one that exists on paper.

The principle of reciprocity is well established in the international literature. The NASEM panel on 21st-century data infrastructure argued that data sharing must be structured so that "all data holders enjoy tangible benefits from participation and that societal benefits are proportionate to the costs and risks involved" (NASEM, 2023, p. 29). Statistics Canada has long pursued what it calls a "win-win partnership" approach — seeking mutually beneficial arrangements with data holders rather than relying solely on legal compulsion, even though the Statistics Act gives the agency compulsory access powers (Statistics Canada, 2016). The reason is practical: legal authority is necessary but not sufficient. Agencies that feel coerced into sharing data will do so reluctantly, with minimal cooperation on quality improvement and metadata documentation. Agencies that see value in the relationship will actively support it.

What does value look like from the perspective of a data holder? It can take many forms. For NADRA, value might mean receiving back improved demographic estimates that help it plan registration drives. For FBR, value might mean receiving aggregated economic indicators at a level of detail that helps target enforcement efforts. For provincial health departments, value might mean receiving integrated health-and-poverty statistics that help them allocate resources more effectively. For BISP, value might mean receiving independent quality assessments of its registry data that improve targeting accuracy.

This requires information to flow in both directions — from data holders to PBS and from PBS back to data holders and the public. The traditional model, in which data flows one way (from respondents to the statistical agency, and from the agency out as published statistics), does not create the feedback loops needed to sustain a multi-party system. The infrastructure must be designed so that every contributor sees the output. When data holders see how their data contributes to better statistics, and when they receive useful analytical products in return, the case for continued sharing becomes self-reinforcing.

The Institutional Prerequisites

Repositioning PBS from a data collector to a national coordinator is not primarily a technical challenge. It is an institutional one. Several conditions need to be in place before the transformation can happen.

First, the Chief Statistician's role must be elevated. In countries with effective statistical coordination, the chief statistician is typically at the most senior level of the civil service — equivalent to a deputy minister or permanent secretary — with direct access to cabinet-level officials and inclusion in regular meetings of senior government leaders. In Canada, the Chief Statistician holds a rank equivalent to

Deputy Minister and attends all routine meetings of deputy ministers, giving the position both visibility and influence (Statistics Canada, 2016). This is not a matter of personal prestige. It is about ensuring that the person responsible for coordinating the national data system has the institutional standing to engage as an equal with the heads of agencies like NADRA, FBR, and the State Bank. PBS's current position as an attached department of the Ministry of Planning limits the Chief Statistician's authority and visibility.

Second, the legal framework must be reformed. The 2011 Act needs to be updated — or supplemented by new legislation — to give PBS explicit authority to coordinate the national statistical system, to set standards for data quality and interoperability that bind other agencies, to access administrative data held by government bodies for approved statistical purposes, and to protect the confidentiality of all data that enters the system. Without legal authority, coordination remains voluntary, and voluntary coordination in government rarely survives changes in leadership or political priorities. The UNECE Guidance on Modernising Statistical Legislation is clear on this point: system-wide coordination authority should have "an explicit basis in the law itself" (UNECE, 2018).

Third, PBS needs new capabilities. Coordinating a multi-source data system requires skills that are different from those needed for traditional survey work. PBS needs staff who can negotiate data-sharing agreements, assess the quality of administrative data, link records across different datasets, apply privacy-enhancing technologies, manage secure data environments, and communicate effectively with non-statistical agencies. Building these capabilities will take years and will require sustained investment in recruitment, training, and institutional development. International partnerships — with statistical offices in Canada, Australia, the Netherlands, and elsewhere — can help accelerate this process, but they cannot substitute for domestic capacity building.

Fourth, governance structures must be established. The repositioning of PBS should be accompanied by the creation of a National Statistical Coordination Committee — a high-level body chaired by the Chief Statistician and including senior representatives from all major data-holding agencies, provincial bureaus of statistics, and independent experts. This committee would advise on priorities, approve data-sharing arrangements, and monitor the implementation of data quality standards across the system. Many countries have such bodies. The PARIS21 guidelines on national strategies for statistical development emphasise that coordination committees are essential for building consensus around statistical priorities and for ensuring that the statistical system responds to the needs of both government and the public (PARIS21, 2007).

Starting Somewhere

The transformation described here is large. It will take years to complete. But it does not require all pieces to be in place before anything can begin. Some actions can be taken immediately, within PBS's existing mandate and resources.

The data inventory effort can begin now. PBS can start cataloguing the administrative data sources it already knows about, and gradually expand the scope as it establishes relationships with other agencies. The quality assessment function can start modestly — perhaps by developing quality profiles for the two or three administrative datasets most likely to be used in near-term blending exercises. Pilot data-sharing arrangements with willing agencies — BISP and NADRA are obvious candidates — can demonstrate value and build trust before comprehensive legal reform is achieved.

And the Chief Statistician can begin convening regular meetings with counterparts in other data-holding agencies, even informally, to build the relationships that coordination depends on.

The UNECE Task Force puts it well: data stewardship is not an all-or-nothing proposition. NSOs can "start with what is feasible and relevant given their current context, resources, and mandate" and expand their role incrementally as trust, capacity, and legal frameworks develop (UNECE, 2024). For PBS, the journey from data collector to national coordinator will be long. But the first steps are both clear and achievable.

The Core Shift The repositioning of PBS is ultimately about a change in institutional identity. PBS must come to see itself — and be seen by others — not as the agency that conducts surveys and censuses, but as the agency that makes Pakistan's national data system work. It retains its production role. But it adds to it a coordination role, a quality assurance role, a standards-setting role, and a facilitation role that together make it the steward of a data infrastructure far larger than anything PBS could build alone.

References

Australian Bureau of Statistics. (2020). The role of the Australian Bureau of Statistics in data governance and stewardship in Australia. Analytical Series. Canberra: ABS.

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality*. Washington, DC: FCSM.

National Academies of Sciences, Engineering, and Medicine. (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good*. Washington, DC: The National Academies Press.

PARIS21. (2007). *A Guide to Designing a National Strategy for the Development of Statistics (NSDS)*. Paris: PARIS21 Secretariat.

PARIS21. (2021). *Coordination Capacity in National Statistical Systems: Background Report*. Paris: PARIS21 Secretariat.

Statistics Canada. (2016). Leadership and coordination of the national statistical system. In *Management of Statistical Information Systems at Statistics Canada*. Ottawa.

Statistics Canada. (2019). *Statistics Canada Data Strategy*. Ottawa: Statistics Canada.

Trépanier, J. (2013). Administrative data initiatives at Statistics Canada. Paper presented at the Federal Committee on Statistical Methodology Research Conference, Washington, DC.

UNECE. (2018). *Guidance on Modernizing Statistical Legislation*. Geneva: United Nations Economic Commission for Europe.

UNECE. (2024). *Data Stewardship and the Role of National Statistical Offices in the New Data Ecosystem*. Geneva: United Nations Economic Commission for Europe.

Data Quality: Fitness for Use

The phrase "fitness for use" entered the quality literature through the work of Joseph Juran, who in the 1950s argued that quality should not be defined by abstract technical standards but by the degree to which a product actually serves the purpose for which it was intended (Juran, 1988). The idea was developed for manufacturing — a component is high-quality if it performs its function reliably — but it has since been adopted by statistical agencies worldwide. The reason is straightforward: not all data that can be collected should be used for national statistics, and no dataset is universally good or bad. Quality depends on the specific question being asked, the methodology used to generate the data, and the context in which the data will be used. A dataset that is perfectly adequate for one purpose — say, monitoring programme disbursements — may be entirely inadequate for another, such as estimating the number of poor households in a district.

This matters enormously for Pakistan's proposed data infrastructure. As earlier chapters have argued, the future of national statistics lies in blending data from multiple sources: surveys, censuses, administrative records, and possibly private sector and digital data. Each of these sources was created for a different purpose, by a different institution, using different methods and definitions. Some will be useful for statistical production. Others will not. And many will be useful only partially or conditionally — suitable for some purposes but not others, adequate in some dimensions but weak in others. The task, therefore, is not to apply a single pass/fail test to each dataset, but to develop a systematic way of assessing quality that is sensitive to purpose, transparent about limitations, and practical enough to be applied across a wide range of data sources.

Why Traditional Quality Frameworks Are Not Enough

For most of the twentieth century, the dominant approach to data quality in statistics was built around the survey. The Total Survey Error (TSE) framework, formalised by Biemer (2010) and others, provided a rigorous way to think about the errors that can affect survey-based statistics: sampling error, coverage error, nonresponse error, measurement error, and processing error. Each source of error was well understood theoretically, and statistical agencies developed sophisticated methods for estimating and controlling them. The quality of a survey could be evaluated by examining its response rate, sample design, fieldwork protocols, and estimation procedures.

This framework works well when the statistical office controls the entire production process — from questionnaire design to fieldwork to data processing. But it was not designed for administrative data, private sector data, or digital data, because these sources were not created through a statistical process at all. Administrative records are generated as a by-product of government operations: tax filing, hospital admissions, border crossings, benefit payments. The data reflects whatever the administrative process was designed to capture, which may or may not correspond to the statistical concept the analyst is interested in. There is no sampling frame, because the data covers whoever interacts with the administrative system — a population that is defined by institutional rules rather than by statistical design. Coverage depends on programme reach and compliance behaviour, not on sample theory. And the metadata that surveys routinely generate — documentation of concepts, definitions, response rates, estimation methods — often does not exist for administrative data, or exists only in fragmented form.

The NASEM panel on 21st-century data infrastructure recognised this gap directly. Their recommendation was that "federal statistical agencies should adopt a broader framework for statistical information than total survey error to include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability" (NASEM, 2017, p. 157). The Federal Committee on Statistical Methodology (FCSM) followed up in 2020 with a comprehensive data quality framework organised around three broad domains — utility, objectivity, and integrity — that is designed to apply to all types of data, "including data collected for nonstatistical purposes, such as data from administrative records and sensors" (FCSM, 2020, p. 3). And most recently, Berzofsky et al. (2025) proposed a Total Data Quality paradigm that explicitly extends the TSE framework to administrative and non-probability data sources, identifying error types specific to these sources — such as definitional mismatch, process change error, and reporting incentive error — that have no direct analogue in survey methodology.

For Pakistan, this shift in thinking is not academic. If PBS is going to incorporate administrative data from NADRA, FBR, BISP, the State Bank, or provincial health departments into statistical production, it needs a quality framework that can assess these sources on their own terms — not one that simply asks whether they meet the standards that apply to a well-designed probability survey.

What Quality Means for Different Data Sources

Quality, then, is not a single dimension but a collection of dimensions, and the relative importance of each dimension depends on the intended use. The European Statistical System (ESS), one of the most developed quality frameworks in official statistics, identifies six principal dimensions: relevance, accuracy, timeliness, accessibility, coherence, and comparability (Eurostat, 2019). The FCSM framework adds several more, including transparency, privacy protection, and interpretability (FCSM, 2020). These dimensions are widely accepted in international practice, and they provide a useful vocabulary for discussing quality. But the vocabulary alone is not enough. The challenge is applying these dimensions concretely to the diverse data sources that Pakistan's infrastructure will need to accommodate.

Consider some examples. When PBS conducts the Labour Force Survey, quality can be assessed against well-established survey criteria: Was the sample representative? Was the response rate adequate? Were the questionnaires administered consistently? Were the estimates properly weighted? These questions have standard answers in the survey methodology literature. But when the question is whether FBR's tax records can be used to estimate formal sector employment, the relevant quality dimensions shift. The key issues become: What population does the tax system actually cover? (Only formal sector workers and registered businesses — a fraction of Pakistan's largely informal economy.) Are the income categories in tax records consistent with the income concepts used in national accounts? (Often not, because tax definitions serve fiscal purposes, not statistical ones.) How stable is the data-generating process? (Tax compliance behaviour changes in response to policy announcements, amnesty schemes, and enforcement campaigns, creating instability in the data even when the underlying economic reality has not changed.)

Similarly, DHIS2 data from provincial health departments may be highly relevant for tracking healthcare utilisation — visits, vaccinations, facility deliveries — but its accuracy depends on whether facility staff enter data correctly and consistently. Coverage depends on which facilities report and which do not. Timeliness may be excellent (data is entered close to the point of care) but coherence

across provinces may be poor, because different provinces may use different definitions or reporting categories for the same indicator. And NADRA's CNIC database covers nearly the entire adult population, making it potentially valuable as a population frame, but its quality for statistical purposes depends on how frequently addresses are updated, whether deceased individuals are removed promptly, and whether the data structure allows linkage with other sources using common identifiers.

Each of these sources has real value for statistical production. But each also has specific quality limitations that must be understood before the data can be used. The critical point is that quality assessment must be source-specific and purpose-specific. A blanket statement that a dataset is "high quality" or "low quality" is almost always misleading.

The Metadata Problem

If there is a single prerequisite for quality assessment, it is metadata — structured information about how data was collected, what concepts and definitions were used, what the data covers, and what its known limitations are. Without adequate metadata, it is literally impossible to assess whether a dataset is fit for a particular statistical purpose. You cannot evaluate coverage if you do not know who was supposed to be in the data. You cannot assess accuracy if you do not know how variables were defined and measured. You cannot judge coherence if you do not know whether two apparently similar variables from different sources actually mean the same thing.

For survey data produced by PBS, metadata is generally available because it is a standard part of the statistical production process. Survey methodology reports document sample design, response rates, definitions, and estimation methods. But for administrative data, the situation is typically much worse. Administrative systems are built to support operational processes, not to produce statistics. Their documentation — if it exists — describes procedures for data entry, not concepts and definitions in the statistical sense. There is rarely any information about coverage gaps, measurement quality, or changes in the data-generating process over time.

This is not a problem unique to Pakistan. The Statistics New Zealand framework for administrative data quality — one of the most influential in the international literature — was developed precisely because the agency found that existing quality frameworks did not adequately address the metadata gaps that arise when administrative data is repurposed for statistics (Reid, Zabala & Holmberg, 2017). Their framework emphasises what they call "fitness for purpose assessment," which begins with a detailed investigation of how the data was generated, what rules and processes govern it, and how those rules and processes have changed over time. This investigation typically requires working closely with the data-holding agency, because much of the relevant information exists only as institutional knowledge — in the heads of programme managers and IT staff — rather than in formal documentation.

For Pakistan, addressing the metadata deficit should be an early priority. Any data-sharing agreement between PBS and an administrative data holder should include a requirement that the data holder provide — or work with PBS to develop — a standardised metadata report covering at minimum: the legal basis for data collection, the population covered, variable definitions, the frequency and method of data collection, known coverage gaps, and any recent or planned changes to the data-generating process. This is not onerous. But it is essential. Without it, PBS would be incorporating data into the national statistical system without knowing what that data actually represents.

Quality Assessment in Practice

How should PBS actually assess the quality of data sources it plans to use? The literature offers several approaches, and no single one is universally best. But a practical approach for Pakistan might combine elements from three established frameworks.

The first is the FCSM framework's three-domain structure. Under this approach, each data source is evaluated against three broad categories: utility (is the data relevant, timely, and accessible?), objectivity (is the data accurate, reliable, and coherent with other sources?), and integrity (was the data produced through a process that protects against manipulation, maintains confidentiality, and operates transparently?) (FCSM, 2020). This structure has the advantage of being comprehensive while remaining relatively simple to communicate.

The second is the input quality approach advocated by the UNECE for multi-source statistics. In this approach, quality is assessed at the point where data enters the statistical production process — before it is blended with other sources — rather than only at the point of final output (UNECE, 2024). This is critical because errors in input data propagate through the entire production chain. If FBR tax data systematically undercounts informal sector workers, no amount of sophisticated statistical modelling can fully correct for that coverage gap in the final estimates. Input quality assessment forces the analyst to identify and document these limitations early, so that they can be addressed — or at least disclosed — before the data is used.

The third is the concept of a quality profile — a standardised document that summarises the quality characteristics of a particular data source across all relevant dimensions. Quality profiles were pioneered by Statistics Canada and have been adopted by several other agencies (Statistics Canada, 2019). The profile for each data source answers a common set of questions: What population does this data cover? How is it collected and updated? What are the known quality strengths and limitations? How has the data-generating process changed over time? The profile is updated periodically and made available to all users of the data — both within the statistical agency and to external researchers.

For Pakistan, a practical starting point would be to develop quality profiles for the five or six administrative data sources most likely to be used in near-term data blending exercises. NADRA's CNIC database, FBR's tax records, BISP's beneficiary registry, SBP's financial sector data, and DHIS2 health data are obvious candidates. Each profile would be developed jointly by PBS and the data-holding agency, and would serve as the basis for decisions about how — and whether — to incorporate the data into statistical production.

A Quality Challenge Specific to Pakistan

All countries face data quality challenges when incorporating non-survey sources into official statistics. But Pakistan faces some additional obstacles that deserve specific attention.

The first is the sheer scale of informality. Pakistan's informal economy accounts for a very large share of economic activity and employment. This means that any administrative data source that depends on formal registration — tax records, social insurance records, corporate registries — will systematically exclude a large portion of the population and the economy. This is a coverage

problem, but it is also a conceptual problem: the population captured by administrative systems is not a random subset of the target population, but a systematically selected one. Ignoring this selection when using the data for statistics will produce biased results.

The second is the weakness of civil registration. WHO estimates indicate that only approximately 40 percent of births and 35 percent of deaths are registered in Pakistan, and virtually no death registrations include medically certified cause of death (Bashir, Mehmood & Samad, 2025). This means that vital statistics in Pakistan are derived almost entirely from household surveys and censuses, not from administrative records. It also means that one of the most fundamental building blocks of a population register — a comprehensive record of births and deaths — does not exist. Until civil registration coverage improves substantially, Pakistan's data infrastructure will lack the demographic foundation that countries with strong CRVS systems take for granted.

The third is inconsistency across provinces. After the 18th Amendment, many statistical and administrative function were devolved to the provinces. This has created a situation in which different provinces may use different definitions, classifications, and reporting formats for the same indicators. Health data is a good example: the indicators reported through DHIS2 may differ across provinces in terms of which conditions are tracked, how severity is classified, and how facility-based data is aggregated. For PBS, this means that achieving coherence — one of the core quality dimensions — requires not just assessing individual data sources but also harmonising definitions and classifications across provincial systems.

The fourth is the absence of a data quality culture across government. In countries with mature statistical systems, data-holding agencies understand that the quality of their administrative records matters not only for their own operations but for the national statistical system as a whole. In Pakistan, this awareness is largely absent. Agencies that generate administrative data do not typically think of themselves as contributors to national statistics, and they have little incentive to invest in the metadata documentation, quality monitoring, and standardisation that statistical use requires. Building this awareness — through engagement, training, and demonstration of value — is as important as any technical quality framework.

FAIR Principles as a Complementary Framework

The quality dimensions discussed above address whether data is fit for statistical use. A related but distinct question is whether data is managed in a way that makes it findable, accessible, interoperable, and reusable — the four attributes captured by the FAIR principles, first articulated by Wilkinson et al. (2016) in the context of scientific data management.

The FAIR principles were not developed for official statistics. They emerged from the scientific research community, where the challenge was enabling researchers to discover and reuse datasets created by other researchers. But the underlying problems are remarkably similar to those facing Pakistan's data infrastructure. Data that cannot be found — because it is not catalogued, not described with adequate metadata, or not registered in any central inventory — cannot be assessed for quality, let alone used. Data that is not accessible — because legal arrangements, technical systems, or institutional gatekeeping prevent access — remains locked in organisational silos regardless of its quality. Data that is not interoperable — because different agencies use different classification systems, identifier formats, or data structures — cannot be combined even when access

is granted. And data that is not reusable — because it lacks the documentation that would allow someone outside the originating agency to understand and use it correctly — will either go unused or be used incorrectly.

Adopting FAIR as a complementary framework does not mean that all government data should be openly published. The FAIR principles are explicit that "accessible" does not mean "open" — data can be FAIR while still being subject to access controls, authentication, and legal restrictions. What FAIR requires is that the existence and characteristics of data are documented and discoverable, that the conditions for access are clearly stated, that formats and standards enable combination across sources, and that documentation is sufficient to support correct reuse.

For Pakistan, FAIR principles translate into concrete infrastructure requirements: a national data catalogue (as discussed in Chapter 8), standardised metadata schemas, common classification systems, and interoperable identifier formats. These are not glamorous investments, but they are the plumbing without which a multi-source data system cannot function.

Building the Quality Function

Quality assessment is not a one-time activity. It is an ongoing institutional function that requires dedicated staff, clear procedures, and sustained investment. PBS should establish a dedicated data quality unit — or expand an existing unit — with the mandate to assess the fitness for use of all data sources that enter the national statistical system, to develop and maintain quality profiles for each source, to work with data-holding agencies on metadata documentation and quality improvement, and to publish regular quality reports that are available to users both inside and outside government.

This function should be guided by a formal quality framework — adapted from the FCSM and ESS models but calibrated to Pakistan's specific context and capacity constraints. The framework should be developed through a consultative process involving PBS, provincial bureaus of statistics, data-holding agencies, and external experts, and it should be published as a public document so that users of Pakistan's statistics can understand the standards against which data quality is assessed.

The ultimate goal is not perfection. No data source is perfect, and no country achieves perfect data quality. The goal is transparency — ensuring that the strengths and limitations of every data source are documented, disclosed, and taken into account when the data is used for statistical production. A system that is honest about its data quality is more trustworthy than one that claims its data is good without evidence to support the claim.

The Quality Imperative Data quality is not a technical afterthought to be addressed once the infrastructure is built. It is the infrastructure. A multi-source statistical system that cannot assess the fitness of its inputs is building on sand. For Pakistan, the priority is to develop the institutional capacity, frameworks, and habits that make quality assessment a routine part of statistical production — applied not only to PBS's own surveys but to every data source that enter the national system.

References

Bashir, F., Mehmood, M. T., & Samad, Z. (2025). A systems-change approach to addressing the mortality surveillance gap in Pakistan. *Journal of Global Health*, 15, 03027.

Berzofsky, M. E., Liao, D., Barnett-Ryan, C., & Smith, E. L. (2025). A total data quality paradigm for official statistics based on administrative data. *Big Data & Society*, 12(1).

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), 817–848.

Eurostat. (2019). *Quality Assurance Framework of the European Statistical System* (Version 2.0). Luxembourg: European Commission.

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality* (FCSM-20-04). Washington, DC: FCSM.

Juran, J. M. (1988). *Juran's Quality Control Handbook* (4th ed.). New York: McGraw-Hill.

National Academies of Sciences, Engineering, and Medicine. (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press.

Reid, G., Zabala, F., & Holmberg, A. (2017). Extending TSE to administrative data: A quality framework and case studies from Stats NZ. *Journal of Official Statistics*, 33(2), 477–511.

Statistics Canada. (2019). *Statistics Canada Data Strategy*. Ottawa: Statistics Canada.

UNECE. (2024). *Data Stewardship and the Role of National Statistical Offices in the New Data Ecosystem*. Geneva: United Nations Economic Commission for Europe.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

Organisational Options for the New Infrastructure

Organisational Options for the New Infrastructure

There is no single correct organisational model for a national data infrastructure. This is a statement that appears in nearly every major report on the subject, and it is worth taking seriously. The design of the institution — who coordinates data access, who sets standards, who makes decisions about linkage and sharing, who is held accountable — matters enormously. It determines whether the infrastructure actually functions or remains an aspiration documented in strategy papers. Different countries have arrived at different arrangements, shaped by their legal traditions, political structures, and the capacities of their existing institutions. Pakistan will need to find the model that fits its own context. But it will benefit from understanding what has been tried elsewhere, what has worked, and why.

The temptation in discussions of organisational design is to treat it as a technical question — draw an organogram, assign responsibilities, and move on. But in the context of data infrastructure, organisational form is deeply political. It involves questions about who gets access to whose data, under what conditions, with what safeguards, and subject to whose oversight. These are questions of power, trust, and institutional legitimacy. The National Academies of Sciences, Engineering, and Medicine (NASEM), in its landmark 2023 report on building a 21st-century national data infrastructure for the United States, emphasised that such infrastructure requires not only technical capabilities but also a legal and regulatory framework that clearly defines which data assets can be shared, with whom, and for what purposes (NASEM, 2023). In countries with decentralised statistical systems — and Pakistan's system is among the most fragmented — these questions are especially difficult. The 18th Amendment devolved many statistical functions to the provinces, while the Pakistan Bureau of Statistics (PBS) retains a federal coordination mandate that, in practice, lacks the authority and resources to be fully effective. Any organisational model for a new data infrastructure must be designed with this constitutional reality firmly in mind.

Lessons from Centralised and Distributed Architectures

Before examining Pakistan's specific options, it is instructive to observe the range of approaches that countries have taken, because these reveal something important: institutional design follows from context, not from theory.

At one end of the spectrum sits the fully centralised model. Statistics Netherlands (CBS) developed the System of Social Statistical Datasets (SSD), a centralised data library in which administrative registers from across government are linked at the individual level using unique personal identification numbers. This infrastructure enabled the Netherlands to conduct its 2001 census entirely from registers and surveys, without traditional field enumeration — the so-called "virtual census" (Bakker, van Rooijen and van Toor, 2014). The Dutch model works because CBS has clear legal authority to access administrative data, a mature population register infrastructure, and decades of experience in record linkage. It also works because the Netherlands is small, highly digitalised, and characterised by strong institutional trust. None of these conditions obtain in Pakistan, which makes the fully centralised model instructive as an aspiration but impractical as a starting point.

At the opposite end sits the distributed model, and Estonia's X-Road is the most cited example. Launched in 2001, X-Road is a decentralised data exchange layer connecting over 900 public and private sector institutions, enabling secure real-time data queries across government databases without centralising data storage. Each agency maintains its own systems and controls access to its own data. The architecture was deliberately designed to avoid creating a single database that could become a vulnerability — a decision shaped by a major data leak in 1996 when a government contractor compiled and sold a "superdatabase" of personal records from multiple government sources (Vassil, 2016). What makes X-Road work is not just technology but a combination of universal digital identity, legal mandates for the "once-only" collection of citizen data, and sustained political commitment over two decades. For Pakistan, the distributed principle is attractive — agencies that are reluctant to surrender data might be more willing to share it through a controlled platform — but the enabling conditions (universal digital ID, legal interoperability mandates, high digital literacy) are only partially present. NADRA's CNIC system provides a foundation, but much else remains to be built.

The honest conclusion from international experience is that most functioning data infrastructures are hybrids. They combine a lead coordinating agency with distributed data holdings, shared governance mechanisms, and specialised entities for particular functions. The question for Pakistan is what form the hybrid should take, given the country's particular institutional constraints.

Why Legal Authority Must Come First

The United Kingdom's experience is particularly instructive here, not because the UK model can be transplanted wholesale but because it illustrates a sequencing principle that Pakistan should take seriously: legal authority must precede institutional design.

The UK's Digital Economy Act 2017 created specific legal gateways for public authorities to share de-identified information with accredited researchers for public-good research. The Act also established the accreditation framework — administered by the UK Statistics Authority — for researchers, projects, and processing environments, structured around what is known as the Five Safes framework: safe projects, safe people, safe settings, safe data, and safe outputs (Ritchie, 2017). Only after this legal foundation was in place did the UK build its major data infrastructure investment, Administrative Data Research UK (ADR UK), established in 2018 with £44 million from the Economic and Social Research Council. ADR UK operates as a partnership between the Office for National Statistics (ONS) and three national partnerships in Scotland, Wales, and Northern Ireland. ONS serves as the main data infrastructure partner, operating the Secure Research Service where linked administrative datasets are made available to accredited researchers. Crucially, ADR UK chose to build on an existing trusted institution rather than creating an entirely new one. As the programme's leadership noted, "government and the public already trust ONS to handle administrative data safely and securely" (ADR UK, 2020).

Two lessons emerge. First, without legislation that creates clear data-sharing gateways with appropriate safeguards, any organisational model will operate in a legal grey zone that breeds caution and non-cooperation among data holders. Pakistan's General Statistics (Reorganization) Act 2011 empowers PBS to access records and documents maintained by government departments, but this is not the same as a comprehensive data-sharing framework with defined conditions, safeguards, and accountability mechanisms for the routine flow of administrative data into the statistical system.

Equivalent legislation — something that creates the Pakistani version of the Five Safes — is a prerequisite for whatever organisational form is ultimately chosen. Second, building on institutions that already possess public trust is more effective than starting from scratch. This has implications for whether PBS should be the anchor institution, and under what conditions.

The Question of PBS

The most natural starting point is to strengthen PBS itself as the central coordinating institution. This builds on existing legal authority, an established institutional identity, and relationships with international organisations. And it avoids the political complexity of creating new statutory bodies.

But the gap between PBS's formal mandate and its actual capacity is wide. The World Bank's 2025 assessment of Pakistan's statistical system noted that PBS needs significant infrastructure and capacity upgrades, that provincial bureaus of statistics remain weak with limited technical and institutional capacity following their devolution in 2011, and that coordination between federal and provincial nodes lacks frameworks for data sharing and quality assurance (World Bank, 2025). PBS operates with roughly 73 percent of its sanctioned posts filled, with acute shortages at senior technical grades — only 6 out of 17 BS-20 positions were occupied at the time of assessment. The 2019 decision to abolish the Statistics Division and place PBS under the Ministry of Planning, Development and Special Initiatives further complicated matters. Recent commentators have argued that this move, by placing the producers of statistics under a line ministry, violated the spirit of sound statistical governance and made PBS vulnerable to political pressure on everything from GDP figures to livestock counts (Javed, 2025). The call to restore the 2011 Act and strengthen the chief statistician's authority to resist interference and coordinate all official statistics-producing agencies is not merely an academic argument — it is a precondition for PBS to credibly serve as the institutional anchor for a multi-source data infrastructure.

India's experience is relevant here, given the similarities in scale, complexity, and federal structure. India's Ministry of Statistics and Programme Implementation (MoSPI) operates the National Statistical Office (NSO), which combines the Central Statistical Office with the National Sample Survey Office. An independent National Statistical Commission, established in 2006 on the recommendation of the Rangarajan Commission, provides oversight. Below the federal level, each of India's 28 states and 8 union territories maintain its own Directorate of Economics and Statistics. A 2025 conference of state ministers on strengthening statistical systems highlighted both the aspiration for better coordination and the continuing challenge: most states requested technical assistance and training from MoSPI, while calling for greater financial support through the Support for Statistical Strengthening Scheme (MoSPI, 2025). What India demonstrates is that in large federal systems, coordination depends not only on institutional design at the centre but on sustained investment in subnational capacity — and that even with strong central institutions, achieving coherence across states require continuous political and bureaucratic engagement. For Pakistan, the implication is that strengthening PBS alone is insufficient; there must be parallel investment in provincial bureaus, with incentives for voluntary participation in a coordinated system.

For PBS to serve as the institutional home for a multi-source data infrastructure, it would need at minimum: restoration and strengthening of its legal autonomy; a dedicated data integration unit with authority to negotiate data-sharing agreements across federal and provincial government; investment in secure computing infrastructure for handling linked administrative datasets; recruitment and

retention of staff with skills in data science, record linkage, and modern statistical methodology; and a principles-based governance framework that provides data-holding agencies with confidence that their data will be handled appropriately.

Creating Separate Specialist Functions

An alternative — or more accurately a complement — is to establish separate entities for specific functions that PBS is not well-positioned to perform in its current state.

The concept of a dedicated data integration service has gained currency internationally. The US Commission on Evidence-Based Policymaking recommended establishing a National Secure Data Service (NSDS) to facilitate the secure blending of government administrative records, survey data, and private sector data for statistical and evidence-building purposes. This concept was legislated as a demonstration project in the CHIPS and Science Act of 2022 (NASEM, 2023). The NSDS would not replace existing statistical agencies but complement them by providing the technical infrastructure and governance mechanisms for cross-agency data linkage. A Pakistani equivalent — perhaps a statutory National Data Integration Service — could be established with its own governing board, mandate, and budget, responsible for negotiating data access, conducting record linkage, maintaining secure environments for research, and enforcing quality standards, while PBS continues its core functions in surveys, censuses, and national accounts.

Australia's institutional separation offers a real-world demonstration that this approach can work. The Australian Institute of Health and Welfare (AIHW), established by its own Act of Parliament in 1987, operates as an independent statutory body with its own ethics committee and accreditation as a Commonwealth Integrating Authority. AIHW brings together data from across the health and welfare systems using a distributed linkage model in which state and territory data linkage units connect to a common spine (AIHW, 2022). This is functionally separate from the Australian Bureau of Statistics, which handles surveys and censuses. Governance is provided through Australia's Data Availability and Transparency Act 2022, which adopted its own version of the Five Safes as "Five Data Sharing Principles" (Australian Productivity Commission, 2017). The lesson is that data integration can be separated from the primary statistical agency without losing coherence, provided governance frameworks are clear, consistently applied, and backed by legislation.

A less ambitious but potentially more feasible variant for Pakistan is a trusted research environment — a secure facility, physical or virtual, where accredited researchers access linked datasets under controlled conditions. In the United States, 31 Federal Statistical Research Data Centers (FSRDCs) located at universities and Federal Reserve banks facilitate collaborative agreements between academic institutions and statistical agencies. Creating even one such facility in Pakistan — at a university with strong quantitative research capacity, operating under PBS oversight and with data shared under negotiated agreements — could serve as a pilot for the broader infrastructure while simultaneously building trust among data-holding agencies and demonstrating the value of integrated data to policymakers. This is a low-risk entry point that does not require new legislation or a new statutory body, only willingness from PBS and at least one data-holding agency to cooperate on specific, bounded projects.

ADR UK itself is, at its core, a research-government partnership funded by a research council and operating through universities and statistical agencies. In Pakistan, a similar arrangement could bring

universities into the infrastructure — not merely as users of data but as institutional partners responsible for analytical capacity, methodological development, and the training of data professionals. Universities can conduct the methodological research on record linkage techniques, privacy-preserving methods, and quality assessment for administrative data that underpins the whole enterprise. And they can provide an environment where experimentation is possible in ways that government agencies, with their accountability structures and risk aversion, may find it difficult to allow. The limitation is sustainability. University-based partnerships in developing countries depend frequently on project funding with limited time horizons. When the project ends, the partnership dissolve and capacity dissipate. If this model is pursued, it must be designed from the outset with a sustainability plan — ideally involving a legislative mandate, core government funding, and institutional structures that outlast any particular funding cycle.

Governing Sensitive Domains Through Stewardship Arrangements

For some categories of data, conventional organisational models may need to be supplemented with governance mechanisms that provide stronger protections and broader stakeholder representation.

The Open Data Institute (ODI) in the United Kingdom has explored a range of "data institutions" — organisations whose primary purpose is the stewardship of data on behalf of others. These include data trusts providing independent fiduciary stewardship, data cooperatives governed by their members, and data commons maintained collaboratively for shared use (Hardinges, 2020; Dodds et al., 2020). In a strict legal sense, a data trust involves trustees who hold a fiduciary duty — a legal obligation of impartiality and loyalty — to the beneficiaries of the data. This is a powerful governance mechanism because it subordinates the interests of any particular agency or user to those of data subjects and the public. However, the concept remain largely experimental globally, and the ODI itself has acknowledged that data trusts represent just one approach among many.

For Pakistan, the data trust concept is most relevant not as the primary organisational model for the entire infrastructure but as a governance mechanism for specific sensitive domains. A health data governance board — with representation from health ministries at federal and provincial levels, patient advocacy groups, medical researchers, and independent ethics experts — could govern the terms under which DHIS2 health facility data is linked with vital registration and other administrative records. Similarly, a social protection data governance arrangement could oversee the integration of BISP beneficiary data with education, health, and employment records, ensuring that the linking of such data serves the interests of the vulnerable populations whose information it contains, rather than becoming a tool for surveillance or exclusion. This domain-specific governance approach complements whatever primary organisational model is chosen for the infrastructure as a whole.

What Is Actually Feasible

Any discussion of organisational options must reckon honestly with Pakistan's institutional constraints. The 18th Amendment has fundamentally altered the distribution of authority, and any organisational model must respect the constitutional reality that many statistical and administrative functions are now provincial responsibilities. A purely federal entity that attempts to compel provincial compliance will face resistance. The infrastructure must be designed as a cooperative arrangement in which provinces participate voluntarily, motivated by the tangible benefits they receive — access to

integrated data products, technical assistance, training, and financial support. India's Support for Statistical Strengthening Scheme offers one template: federal investment in provincial capacity, conditional on adoption of common standards and participation in coordination mechanisms.

PBS's current institutional position presents a further complication. Placed under the Ministry of Planning in 2019, it lacks the independent authority envisioned by the 2011 Act. The chief statistician does not currently have the standing to negotiate with powerful agencies like NADRA, FBR, or the State Bank as an equal. Restoring PBS's statutory autonomy is not an abstract governance ideal — it is a practical requirement for the data-sharing negotiations that must underpin any infrastructure.

Given these constraints, the most realistic approach is probably a phased strategy rather than a single organisational choice made at the outset.

In the near term, the priority should be legal and institutional reform: restoring PBS's statutory autonomy, enacting data-sharing legislation with safeguards analogous to the Five Safes, and establishing a national statistics advisory body with representation from both federal and provincial governments. Simultaneously, PBS should establish or expand a data integration unit — even a small one — with the mandate to develop agreements, conduct pilot linkage projects with willing agencies, and build secure data handling capacity. NADRA's CNIC database and FBR's tax records are obvious starting points for pilot work.

In the medium term, a trusted research environment established jointly by PBS and one or more universities could provide accredited researchers with access to linked datasets under controlled conditions, generating analytical products that demonstrate value to policymakers while building broader institutional confidence in the infrastructure.

In the longer term, the model may evolve toward a more elaborate partnership — with PBS as the central infrastructure partner, provincial bureaus as regional nodes, university-based centres providing analytical capacity, and domain-specific governance arrangements for sensitive data in health, social protection, and other area where public sensitivity is highest. A distributed technical architecture — in which agencies maintain their own data but connect through a shared platform with common standards — may be more appropriate for Pakistan's federal structure than full centralisation.

The important point is that organisational design is not a one-time decision made on paper and then implemented. It is an evolving process that responds to developing legal frameworks, growing capacities, and accumulating trust among participants. What matters most at the outset is not choosing the theoretically perfect model but getting the fundamentals right: legal authority, institutional autonomy, governance principles, and the willingness to begin with modest pilot projects and learn from what they teach. The infrastructure will be built incrementally, and the organisational form will mature alongside the capacities of the institutions that operate it. What Pakistan cannot afford is to wait for the perfect design before taking the first practical steps.

References

ADR UK (2020). Administrative Data Research UK. *Patterns* 1(2), 100044.

AIHW (2022). *Data Governance Framework*. Canberra: Australian Institute of Health and Welfare.

Australian Productivity Commission (2017). *Data Availability and Use: Inquiry Report*. Canberra: Productivity Commission.

Bakker, B. F. M., van Rooijen, J. and van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS* 30(4), 411–424.

Dodds, L., Szász, D., Keller, J., Snaith, B., Duarte, S., Hardinges, J. and Tennison, J. (2020). *Designing Sustainable Data Institutions*. London: Open Data Institute.

Hardinges, J. (2020). Data trusts in 2020. London: Open Data Institute.

Javed, U. (2025). Reforming the PBS. *The News*, 27 July 2025.

MoSPI (2025). Conference of State Government Ministers on Strengthening of Statistical Systems. Press Information Bureau, Government of India, 5 April 2025.

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good*. Washington, DC: The National Academies Press.

Ritchie, F. (2017). The 'Five Safes': A framework for planning, designing and evaluating data access solutions. *Data for Policy Conference 2017*.

Vassil, K. (2016). *Estonian e-Government Ecosystem: Foundation, Applications, Outcomes*. Washington, DC: World Bank.

World Bank (2025). *Pakistan Country Partnership Framework FY25–FY35*. Washington, DC: World Bank Group.

Ethics, Privacy, and the Rights of Data Subjects

Ethics, Privacy, and the Rights of Data Subjects

A national data infrastructure built on the integration of administrative records is, at its core, a project that depends on the willingness of citizens to trust that their information will be handled responsibly. This is not a secondary concern to be addressed once the technical architecture is in place. It is the foundation on which the entire enterprise stands or falls. The United Kingdom's experience with care.data offers a stark illustration: in 2014, NHS England launched a programme to extract patient records from general practitioner systems and link them to hospital data for research and planning purposes. The programme was abandoned in 2016 after sustained public and professional opposition, driven by inadequate communication, unclear consent mechanisms, and revelations that patient data had been shared with insurance companies and commercial organisations without proper governance (Carter et al., 2015; Sterckx et al., 2016). The technical infrastructure was sound. The ethical infrastructure was not. As Floridi and Taddeo (2016) observed, the care.data experience demonstrated that "social acceptability, or even better, social preferability, must be the guiding principles for any data science project with even a remote impact on human life."

For Pakistan, the stakes may be even higher. The country does not yet have a comprehensive data protection law. The Personal Data Protection Bill, first drafted in 2023, has not been enacted as of mid-2025. The National Database and Registration Authority (NADRA) holds biometric records on over 200 million citizens — one of the world's largest centralised identity databases — yet a Joint Investigation Team confirmed in 2024 that data on 2.7 million citizens was stolen from NADRA systems between 2019 and 2023, with records transferred to Dubai and eventually sold in Argentina and Romania (Privacy International, 2024). The Federal Board of Revenue suffered a separate data breach in 2021. These are the agencies whose administrative data any new statistical infrastructure would seek to integrate. Without a credible ethical framework, the public will have every reason to resist.

The Re-identification Problem Is Real

The starting point for any ethical discussion of data linkage is the recognition that "de-identified" data is not the same as anonymous data. In 2000, Latanya Sweeney demonstrated that 87 percent of the United States population could be uniquely identified using just three pieces of information: a five-digit zip code, date of birth, and gender (Sweeney, 2000). This finding — replicated and extended in numerous subsequent studies — established a fundamental principle: removing names and identity numbers from a dataset does not, by itself, protect individuals from re-identification. When multiple datasets are combined, the risk increases, because each additional source provides new quasi-identifiers that can be cross-referenced to narrow down individuals.

In the context of a Pakistani data infrastructure that might link NADRA identity records with tax data from FBR, health facility records from DHIS2, education data from NEMIS, and social protection records from BISP, the re-identification risk is not theoretical. The CNIC number — a 13-digit unique identifier held by virtually every adult citizen — provides a direct linkage key. Even without CNIC, combinations of district, age, gender, household size, and programme participation status could

identify individuals in sparsely populated areas or among minority groups with distinctive demographic profiles. The infrastructure must be designed on the assumption that re-identification is always technically possible, and that the protection of individuals therefore depend not only on statistical disclosure control but on institutional safeguards, legal constraints, and governance mechanisms that control who can access what data, for what purpose, under what conditions.

This is the logic underlying the Five Safes framework, developed by Ritchie (2017) and now widely adopted across the United Kingdom, Australia, and other countries with advanced data infrastructure. The framework recognises that no single safeguard is sufficient and that protection comes from the combination of: safe projects (research must serve a legitimate public purpose and be approved through ethical review); safe people (researchers must be trained and accredited); safe settings (data must be accessed only through secure environments); safe data (datasets should be de-identified to the extent possible without destroying analytical utility); and safe outputs (all results must be checked before release to ensure they cannot identify individuals). The Five Safes approach explicitly reject the idea that de-identification alone is adequate. It treats privacy protection as a system property — the product of multiple layered safeguards — rather than a feature of any individual technical step.

Why Consent Alone Cannot Solve the Problem

Traditional research ethics, rooted in the Belmont Report's principles of respect for persons, beneficence, and justice, place informed consent at the centre of ethical practice. Individuals should know what data is being collected about them, how it will be used, and should have the right to refuse participation. These principles remain essential. But they face severe practical limitations when applied to the integration of administrative records.

The data in administrative systems was collected for operational purposes — registering births, collecting taxes, distributing social protection payments, recording hospital visits — not for statistical research. Citizens provided their information to receive a service, not to participate in a study. Seeking retrospective consent from millions of individuals whose records are already in government databases is logistically impossible and would introduce devastating selection bias: those who decline consent would disproportionately be the most vulnerable, the most distrustful of government, and the most difficult to reach, which means the resulting datasets would systematically misrepresent the very populations that statistical evidence is most needed to serve. This is not a hypothetical concern. Studies of consent bias in data linkage consistently find that non-consenters differ systematically from consenters on key sociodemographic variables, including income, education, and health status (Sakshaug et al., 2012).

The alternative to individual consent is not the absence of consent. It is a different model of authorisation — one based on legal frameworks, institutional governance, and what scholars have termed "social licence." The concept of social licence, borrowed from the extractive industries and adapted for data use, refers to the ongoing acceptance by the public that their data is being used in ways that are legitimate, transparent, and beneficial (Carter et al., 2015). Social licence is not granted once and forgotten. It must be continuously maintained through transparent communication about what data is being used, by whom, for what purposes, with what safeguards, and to what public benefit. When social licence is violated — as it was with care.data — the damage to public trust can take years to repair.

For Pakistan, this means that the ethical framework for data integration cannot rely solely on the legal authority granted to PBS under the General Statistics (Reorganization) Act 2011 to access government records. Legal authority is necessary but not sufficient. The infrastructure must also demonstrate, through visible and accountable governance mechanisms, that data about citizens is being used for their benefit and not to their detriment. This is especially important in a context where public trust in government institutions is fragile and where citizens have seen their data compromised through security breaches at the very agencies whose records will form the backbone of the infrastructure.

The Problem of Group Privacy

Conventional data protection frameworks focus on protecting the identity of individual persons. The European Union's General Data Protection Regulation (GDPR), for instance, is concerned primarily with "identifiable natural persons." But as Floridi (2014) argued, this individualistic focus misses an important dimension of the problem. In data-intensive environments, harms can occur at the group level even when no individual is identified. If a linked dataset reveals that residents of a particular neighbourhood have high rates of certain health conditions, or that members of a particular ethnic or religious community are disproportionately represented in social protection programmes, the information can be used to stigmatise, discriminate against, or target those communities — without any single individual being identified.

This concern has particular salience in Pakistan. The country's social fabric is characterised by deep ethnic, sectarian, and linguistic divisions. Administrative data, when linked across sources, could reveal patterns that, in the wrong hands or without proper governance, might be used to reinforce existing inequalities or to facilitate surveillance and discrimination. Data on caste, tribe, ethnicity, and religious affiliation — all of which are collected in various administrative systems — is classified as "sensitive personal data" even under Pakistan's draft data protection bill. The ethical framework for the data infrastructure must address group privacy explicitly, not merely as an extension of individual privacy but as a distinct concern requiring its own safeguards. This might include restrictions on the types of analysis that can be performed on certain variables, requirements for community consultation before research involving identifiable population subgroups, and prohibitions on the release of outputs that could facilitate discrimination against vulnerable groups.

Taylor (2017), writing about the ethics of data use in developing countries, has noted that data collected ostensibly for humanitarian or development purposes can be repurposed for surveillance, border control, or political targeting. In contexts where the state's relationship with certain communities is characterised by tension or conflict — and this includes parts of Pakistan — the risk that linked administrative data could be used for purposes far removed from statistical analysis is not negligible. The ethical framework must include mechanisms to prevent such repurposing, including legal restrictions on secondary use, independent oversight of data access decisions, and transparency requirements that enable civil society to monitor how the data is being used.

Proportionality and the Minimum Necessary Principle

The ethical use of administrative data requires adherence to the principle of proportionality: the benefits of data use must be proportionate to the risks, and the data collected or accessed should be

the minimum necessary to achieve the stated purpose. This principle, embedded in both the GDPR and in the privacy frameworks of most jurisdictions that have adopted data protection legislation, has practical implications for how a Pakistani data infrastructure should operate.

First, not all data needs to be linked at the individual level. Many important statistical questions can be answered using aggregate data, tabulations, or statistical techniques that do not require researchers to access individual records. The infrastructure should be designed so that individual-level linkage is reserved for research questions that genuinely require it, and even then, the linked data should be de-identified to the maximum extent consistent with analytical validity. The principle of data minimisation — collecting and processing only the data that is strictly necessary for a specified purpose — should be built into the technical architecture, not merely stated as a policy aspiration.

Second, access to linked data should be tiered. Some users may need only aggregate statistics or pre-computed indicators. Others may require access to microdata within a secure research environment. The most sensitive linkage projects — those involving health records, criminal justice data, or information about vulnerable populations — should require additional layers of approval, including review by an independent ethics committee with representation from affected communities. The Australian model, in which the Australian Institute of Health and Welfare (AIHW) operates its own ethics committee and conducts independent review of all data linkage proposals involving health and welfare data, provides one template (AIHW, 2022). Pakistan's infrastructure should develop an equivalent mechanism, adapted to the local context.

Third, the outputs of any research using linked data must be subject to disclosure control before they are released. This mean that all tables, figures, and statistical results should be reviewed to ensure that small cell counts, unusual combinations of variables, or other features do not permit the identification of individuals or small groups. The UK's Office for National Statistics operates a systematic output checking process for all research conducted in its Secure Research Service, and similar processes should be a non-negotiable component of any Pakistani data research environment.

The Question of Benefit

An ethical framework that focuses only on preventing harm, while necessary, is incomplete. The use of citizens' data must also demonstrate positive benefit — and specifically, benefit that flows back to the communities whose information the infrastructure holds. This is the principle of beneficence, which the Belmont Report articulated as a dual obligation: to do no harm, and to maximise possible benefits while minimising possible harms.

In the context of a statistical data infrastructure, beneficence means that the research conducted using linked data should address questions that matter for public welfare. It should improve the evidence base for policies that affect the lives of the people whose data is being used. If administrative data from BISP is linked with education records to study the impact of cash transfers on children's school attendance, the findings should inform policy in ways that benefit the programme's beneficiaries — the poorest households in the country. If health facility data from DHIS2 is linked with vital registration to study maternal mortality, the results should contribute to interventions that reduce maternal deaths.

This requires that the governance framework include mechanisms for prioritising research that serves the public interest. Not every proposed use of linked data will be equally valuable, and the infrastructure should have the capacity to assess whether proposed research projects are likely to generate knowledge that benefits society — and particularly the most vulnerable members of society — rather than serving primarily commercial or narrow institutional interests. The UK's Digital Economy Act 2017 explicitly requires that data shared under its provisions be used for purposes that serve the "public good." Pakistan's legal framework for data sharing should include a similar requirement, and the governance body overseeing data access should have the authority to reject proposals that do not meet this standard.

Equally important is the communication of benefits. The public must be able to see, in concrete terms, what societal good has come from the use of their data. This requires proactive transparency — regular public reporting on what research has been conducted, what findings have emerged, and what policy changes have resulted. If citizens can see that their data contributed to evidence that improved maternal health services, or that identified gaps in educational access, or that strengthened the targeting of social protection programmes, they are more likely to maintain their support for the infrastructure. If the benefits remain invisible, locked in academic journals or government reports that no one reads, the social licence on which the infrastructure depends will erode.

Pakistan's Ethical Landscape

Pakistan's constitutional and legal framework provides some foundation for data ethics, but significant gaps remain. Article 14(1) of the Constitution guarantees "the dignity of man and, subject to law, the privacy of home, shall be inviolable." The courts have interpreted this provision expansively; the Lahore High Court in *M.D. Tahir v. State Bank of Pakistan* held that unauthorised collection of personal data constituted an "extraordinary invasion" of liberty. But constitutional provisions are general principles, not operational frameworks. Without implementing legislation that specifies how data should be collected, stored, shared, and protected — and that creates enforcement mechanisms with meaningful sanctions for violations — the constitutional right to privacy remains more aspirational than protective.

The absence of a comprehensive data protection law is perhaps the most significant ethical gap that the infrastructure will need to confront. The Personal Data Protection Bill 2023 has not yet been enacted. Even the draft bill has been criticised by digital rights organisations for exempting government agencies from key compliance obligations — precisely the agencies whose data would be integrated in a national data infrastructure (GenderIT, 2025). If the government exempts its own data handling from the rules it imposes on the private sector, the credibility of any ethical framework built around the infrastructure will be fundamentally compromised.

NADRA's track record illustrates the problem concretely. Despite holding biometric records on virtually the entire adult population, NADRA has experienced multiple security breaches, and the 2024 Joint Investigation Team report revealed that data theft was facilitated by insiders at NADRA offices in multiple cities (Privacy International, 2024). The Federal Board of Revenue's 2021 breach further demonstrated systemic vulnerability. For citizens, these experiences shape their perception of government data handling. An ethical framework for the new infrastructure must acknowledge this history honestly and take concrete steps to demonstrate that the new system will operate under fundamentally different standards of security and governance.

Several practical measures could help establish ethical credibility. First, the data-sharing legislation that enables the infrastructure should apply equally to government and non-government entities, with no blanket exemptions for national security or administrative convenience. Second, an independent oversight body — distinct from both PBS and the data-holding agencies — should review and approve all data linkage projects, with authority to reject proposals that do not meet ethical standards and to audit compliance after approval. Third, a public register of all approved data linkage projects, including their purposes, the data sources involved, and the safeguards applied, should be maintained and accessible to citizens. Fourth, the infrastructure should adopt the Five Safes framework or a comparable principles-based approach that provides a clear, communicable standard for ethical data use. Fifth, meaningful sanctions for data misuse — including criminal penalties for deliberate breaches of confidentiality and administrative penalties for negligent handling — should be established by law and enforced consistently.

Dignity as a Design Principle

The ethical considerations discussed in this chapter are sometimes presented as constraints on the infrastructure — as obstacles that slow down the work and limit what can be done with data. This framing is mistaken. Ethics is not an obstacle to building a data infrastructure. It is a design requirement for building one that works.

An infrastructure that respect the dignity of data subjects — that treats citizens not as sources of extractable information but as rights-holders whose data carries obligations — will earn the trust it needs to function. An infrastructure that ignores these considerations may be built, but it will operate in a climate of suspicion, non-cooperation, and resistance that will ultimately limit its usefulness far more than any ethical safeguard ever could. The care.data experience is only the most prominent example. Around the world, data-intensive programmes that failed to invest adequately in ethical governance have faced public backlash, legal challenges, and political abandonment.

Pakistan has the opportunity to learn from these experiences and to build ethics into the infrastructure from the beginning, rather than attempting to retrofit it after trust has been lost. The challenge is significant, given the legal gaps, institutional vulnerabilities, and the history of data breaches. But it is precisely because the context is difficult that the ethical commitment must be explicit, visible, and credible. Citizens whose data powers the infrastructure must be able to see how their data are used, by whom, for what purposes, and to what societal benefit. This transparency is not a secondary feature. It is the foundation on which public trust in the entire system rests, and without it, the system cannot be sustained.

References

AIHW (2022). *Data Governance Framework*. Canberra: Australian Institute of Health and Welfare.

Carter, P., Laurie, G. T. and Dixon-Woods, M. (2015). The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics* 41(5), 404–409.

Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy and Technology* 27(1), 1–3.

Floridi, L. and Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A* 374(2083), 20160360.

GenderIT (2025). Between privacy and power: the fine line in Pakistan's data protection bill. GenderIT.org.

Privacy International (2024). State of Privacy: Pakistan. London: Privacy International.

Ritchie, F. (2017). The 'Five Safes': a framework for planning, designing and evaluating data access solutions. *Data for Policy Conference 2017*.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B. and Weir, D. R. (2012). Linking survey and administrative records: mechanisms of consent. *Sociological Methods and Research* 41(4), 535–569.

Sterckx, S., Rakic, V., Cockbain, J. and Borry, P. (2016). "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. *Medicine, Health Care and Philosophy* 19(2), 177–190.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Carnegie Mellon University Data Privacy Working Paper* 3.

Taylor, L. (2017). The ethics of big data as a public good: which public? whose good? *Philosophical Transactions of the Royal Society A* 374(2083), 20160126.

UK Digital Economy Act (2017). *Digital Economy Act 2017*. London: Her Majesty's Stationery Office.

Preparing for the Future: AI-Ready Data

Preparing for the Future: AI-Ready Data

It would be a mistake — and not a small one — to design a new data infrastructure that serves only today's analytical needs. The world is changing rapidly in how data is produced, managed, and used, and the countries that fail to anticipate these changes will find themselves building systems that are already outdated by the time they become operational. Artificial intelligence and machine learning are not futuristic possibilities. They are present realities, transforming everything from how national accounts are compiled to how satellite imagery is classified, from how survey non-response is imputed to how administrative records are linked across agencies. None of these applications would be possible without high-quality, well-structured, properly documented data. The infrastructure Pakistan builds now must therefore be designed not merely for traditional statistical tabulation but for a future in which advanced computational methods are the standard tools of evidence production.

This does not mean that Pakistan should rush to adopt every fashionable technology. It means something more fundamental: that the basic properties of data — its structure, its documentation, its consistency, its accessibility — must meet standards that enable not only human analysts but also machines to find, read, interpret, and use it. Getting these foundations right is the most important investment the country can make. Without them, even the most sophisticated algorithms will produce unreliable results, and the promise of AI for public policy will remain unfulfilled.

The Gap Between Collecting Data and Making It Usable

Pakistan collects an enormous amount of data. NADRA processes biometric records for over 200 million citizens. The District Health Information System (DHIS2) captures data from thousands of health facilities across the country. The National Education Management Information System (NEMIS) tracks school enrolments and teacher deployments. The Federal Board of Revenue hold tax records. BISP maintains records on millions of beneficiary households. The Pakistan Bureau of Statistics conducts censuses, labour force surveys, household income and expenditure surveys, and a range of other data collection exercises.

But collecting data is not the same as making it usable. Much of Pakistan's existing data exists in formats that are difficult to access, hard to combine with other sources, and poorly documented. Survey microdata files are sometimes released without complete data dictionaries or codebooks. Administrative records follow agency-specific coding systems that are incompatible with one another. Metadata — the descriptive information that tells a user what each variable means, how it was measured, when it was collected, and what limitations it carries — is frequently incomplete or altogether absent. A dataset without proper metadata is like a library book without a catalogue entry: it may contain valuable information, but no one can find it, and even those who stumble upon it cannot easily determine whether it is relevant to their needs or how to interpret it correctly.

This problem is not unique to Pakistan. Across the world, researchers report spending up to 80 percent of their time preparing data into usable formats before any analysis can begin (Bipartisan Policy Center, 2022). The gap between raw data and analytical-ready data is enormous, and bridging

it consumes resources that could otherwise be devoted to producing insights. In developing countries, where technical capacity is already scarce, this inefficiency is especially costly. The infrastructure Pakistan builds must therefore treat data preparation, documentation, and quality assurance not as afterthoughts but as core functions — as essential as data collection itself.

What FAIR Actually Requires

The international standard for making data usable is captured in the FAIR principles, first articulated by Wilkinson and colleagues in 2016 and since adopted by research funders, statistical agencies, and international organisations around the world. FAIR stands for Findable, Accessible, Interoperable, and Reusable. Each principle addresses a specific barrier to data use.

Findable means that datasets must be registered in searchable catalogues with rich, standardised metadata, and that each dataset must have a unique, persistent identifier — the digital equivalent of an ISBN for a book. Without this, users cannot discover what data exists. In Pakistan today, there is no comprehensive catalogue of the data held by different government agencies. A researcher wanting to know what education data is available, at what geographic level, for what time periods, and with what variables, would need to contact individual agencies and hope for a response. A national data catalogue — a searchable register of all major datasets held by government, with standardised descriptions of their content, coverage, and access conditions — would be a relatively low-cost, high-impact first step.

Accessible means that once a user finds a dataset, there should be a clear, documented process for obtaining access to it. This does not mean that all data must be openly downloadable. Some data — particularly individual-level administrative records — must be restricted for privacy reasons. But the access procedures must be transparent and standardised, not ad hoc. Users should know what data is available, what conditions apply to its use, and how to apply for access. The current situation in Pakistan, where access to government data often depend on personal connections and informal negotiations rather than on published procedures, is neither efficient nor equitable.

Interoperable means that data from different sources must be able to work together. This requires common coding standards, shared classification systems, and compatible file formats. If DHIS2 records a patient's district of residence using one coding scheme and NADRA uses a different one, linking the two datasets becomes a labour-intensive exercise in manual reconciliation. If PBS uses one definition of "urban" and provincial planning departments use another, combining their data produces confusion rather than clarity. Interoperability is not achieved by decree. It requires sustained technical work to develop and maintain common standards, and institutional commitment to adopting them consistently across agencies. The Statistical Data and Metadata Exchange (SDMX) standard, developed by a consortium of international organisations including the IMF, World Bank, and OECD, provides one framework for achieving interoperability in statistical data. Pakistan's adoption of SDMX standards — already recommended in various reform proposals — would be a significant step toward making its data infrastructure interoperable.

Reusable means that data must be well-documented enough that it can be understood, replicated, and used in new contexts by people who were not involved in its original collection. This requires detailed metadata describing how variables were defined, how samples were drawn, what quality checks were applied, and what known limitations the data carries. It also require clear licensing that

specifies what users are permitted to do with the data. In many countries, including Pakistan, the legal status of government data — who owns it, who can use it, under what conditions — is ambiguous. This ambiguity discourages reuse and forces potential users to navigate a fog of uncertainty that many, especially those outside government, simply choose to avoid.

It is worth emphasising that FAIR does not mean "open." The FAIR principles are sometimes conflated with the open data movement, but they are conceptually distinct. A dataset can be FAIR — well-documented, discoverable, structured, and accessible through a transparent process — without being publicly downloadable. This distinction is critically important for administrative data, which often contains sensitive individual-level information that cannot and should not be released publicly. The FAIR framework accommodates restricted access: what it requires is that the *conditions* of access are clear and standardised, not that access is unrestricted. For Pakistan, this means that applying FAIR principles to NADRA records, tax data, or health facility records does not require making them public. It requires documenting what the data contains, establishing who can access it, under what governance procedures, and ensuring that the data is structured in ways that permit authorised linkage and analysis.

Why AI Changes the Stakes

The FAIR principles were developed primarily with human researchers in mind, but they have become even more important in the age of artificial intelligence. Machine learning systems are powerful pattern-recognition tools, but they are also unforgiving consumers of data. A human analyst can often work around inconsistencies in a dataset — recognising, for instance, that "Faisalabad" and "Lyallpur" refer to the same city, or that a column labelled "inc" probably means "income." A machine cannot make these inferences unless the data is structured and documented in ways that eliminate ambiguity.

This is what "AI-ready" means in practice: data that is not only high-quality by traditional statistical standards but also structured for machine consumption. The UK government's 2025 guidelines on making government datasets AI-ready identify four pillars: technical optimisation (data is structured and formatted for efficient machine use); data and metadata quality (data is accurate, complete, consistent, and maintained through effective processes); organisational and infrastructure context (data governance and stewardship arrangements are in place); and legal, security, and ethical compliance (data use is aligned with applicable laws and managed responsibly) (DSIT, 2025). These four pillars are not separate from good statistical practice. They are the same principles that make data useful for any purpose, carried to the standard of consistency and documentation that machines require.

For Pakistan, the practical implications are significant. Consider just one example: the potential use of machine learning to improve the targeting of social protection programmes. BISP currently uses a proxy means test based on household survey data to determine eligibility for cash transfers. Machine learning methods could potentially improve targeting accuracy by combining survey data with administrative records from multiple sources — tax data, utility consumption, land ownership records, school enrolment — to build more nuanced predictions of household welfare. But this would require all of these data sources to be structured in compatible formats, documented with consistent metadata, linked through common identifiers, and accessible through governed procedures. Without these foundations, the machine learning application is not feasible, regardless of how sophisticated

the algorithm might be.

The same logic applies to other domains. Predictive analytics for disease surveillance using DHIS2 data. Satellite imagery classification for agricultural crop monitoring. Natural language processing of court records for justice system analysis. Text mining of parliamentary proceedings. In every case, the quality, structure, and documentation of the underlying data determine whether the analytical method can produce reliable results. The data infrastructure is the binding constraint, not the algorithm.

The National Data Library as Institutional Architecture

The concept of a National Data Library has gained significant momentum internationally, most prominently in the United Kingdom, where the government announced plans in 2024 to create such an institution. The UK's National Data Library, as envisioned by the Department for Science, Innovation and Technology (DSIT), is not a single centralised database. It is a service layer — a governed infrastructure that enables curated, de-identified, research-ready datasets to be discoverable, accessible, and linkable within secure environments (GOV.UK, 2025). The Tony Blair Institute estimated in 2025 that a fully developed National Data Library could generate returns of £5 for every £1 invested in data linkage, with wider societal benefits potentially reaching £319 billion by 2050 (TBI, 2025).

The UK model builds on the work of Administrative Data Research UK (ADR UK), which since 2018 has invested over £105 million in creating secure access to linked administrative data for research. A key insight from ADR UK's experience is that the traditional "create and destroy" model of data access — where each research project negotiates its own access to raw data, conducts its own linkage, and destroys the linked dataset after use — is enormously inefficient. The ADR UK approach instead does the governance, cleaning, and linkage work upfront, so that de-identified, research-ready, curated datasets can be maintained over time and accessed by multiple approved researchers. This allows knowledge to accumulate: researchers can share code, derived variables, and analytical methods, building on what has come before rather than starting from scratch each time (ADR UK, 2024).

For Pakistan, the National Data Library concept is worth adapting to local conditions. The country does not need to build a replica of the UK system. It need a mechanism that performs several essential functions: cataloguing what data exists across government; establishing and enforcing common standards for data structure, documentation, and quality; creating governed procedures for data access that are transparent, consistent, and proportionate to the sensitivity of the data; providing a secure environment for approved research using linked administrative data; and building institutional capacity for data curation and stewardship.

This last function — capacity for data curation — is perhaps the most neglected and most important. Data does not curate itself. Converting raw administrative records into research-ready datasets requires skilled professionals who understand both the domain (health, education, taxation) and the technical requirements of data management. These skills are in short supply in Pakistan, and the infrastructure will not function without sustained investment in building them. A National Data Library is not primarily a technology project. It is an institutional project — one that requires people, governance structures, and sustained funding at least as much as it require servers and software.

From Static Systems to Adaptive Infrastructure

A data infrastructure designed only for the analytical methods of today will be inadequate for the methods of tomorrow. The pace of technological change means that new types of data, new sources, and new analytical approaches will emerge continuously. An infrastructure that cannot accommodate these changes will quickly become obsolete.

Several developments illustrate what an adaptive infrastructure must prepare for. First, the growing availability of geospatial data — satellite imagery, GPS traces, mobile phone location data — offers enormous potential for understanding economic activity, agricultural production, urbanisation patterns, and environmental change. Pakistan's agricultural sector, which employs a large share of the workforce and is increasingly affected by climate variability, could benefit enormously from the integration of satellite-derived crop estimates with ground-level agricultural survey data. But this requires the infrastructure to handle raster data, vector data, and tabular data within a common framework.

Second, the proliferation of real-time and high-frequency data streams — from mobile phone usage, digital payment systems, web traffic, and sensor networks — creates opportunities for more timely indicators of economic and social conditions. During the COVID-19 pandemic, countries that could rapidly deploy alternative data sources to supplement traditional surveys were better positioned to monitor the crisis in near real-time. Pakistan's infrastructure should be designed to incorporate such data sources as they become available, even if their integration is not immediate.

Third, the development of privacy-preserving analytical techniques — including differential privacy, federated learning, and secure multi-party computation — offers new ways to extract insights from sensitive data without exposing individual records. Federated learning, for example, allows machine learning models to be trained on data distributed across multiple institutions without the data ever leaving its original location. This could be particularly valuable in Pakistan, where institutional reluctance to share raw data is a major barrier to integration. If the infrastructure supports federated approaches, agencies can contribute to analytical products without surrendering control of their data.

Fourth, the emergence of large language models and other generative AI tools creates new possibilities for working with unstructured data — text, images, audio — that have traditionally been outside the scope of statistical infrastructure. Court records, parliamentary debates, citizen complaints, and media coverage all contain information relevant to policy analysis. An infrastructure that can accommodate unstructured data alongside traditional structured datasets will be significantly more valuable than one that cannot.

None of these developments require Pakistan to adopt them immediately. They require that the infrastructure be designed with sufficient flexibility to accommodate them as they mature and as the country's technical capacity grows. This means adopting open standards rather than proprietary formats, modular architectures rather than monolithic systems, and governance frameworks that can evolve with changing technology and changing needs.

The temptation in any large infrastructure project is to build for the specifications of the moment. A system designed in 2025 to process the data types and volumes that exist in 2025 will struggle when confronted with the data types and volumes of 2030. The infrastructure should therefore be built on the principle that change is not an exception to be managed but a constant to be expected.

Application programming interfaces (APIs) should be designed to accommodate new data sources without requiring re-architecture. Data storage systems should handle structured, semi-structured, and unstructured data. Governance protocols should be flexible enough to apply to data types that have not yet been imagined. This is not a call for unlimited spending on future-proofing. It is a call for architectural decisions that prioritise openness and modularity over closure and rigidity.

The Political Economy of Data as a Public Asset

There is one final argument that must be made, and it is not a technical one. Data produced through the use of public resources — through government surveys, administrative processes, and publicly funded research — is a public asset. It belong to the citizens whose taxes funded its collection and whose information it contains. Treating it as such has implications for how the infrastructure is governed.

In too many countries, including Pakistan, government data is treated as the property of the agency that collected it. Agencies are reluctant to share their data with other agencies, let alone with researchers or the public. This reluctance is sometimes justified by legitimate concerns about privacy, quality, or misinterpretation. But it is often driven by institutional culture — a sense of ownership over "my data" — or by bureaucratic inertia. The result is that data collected at public expense sits unused in agency silos, generating no value beyond the narrow purpose for which it was originally collected.

Overcoming this requires a shift in mindset that treats data sharing as the default rather than the exception, subject to appropriate safeguards for privacy and confidentiality. The UK's Digital Economy Act 2017 established a legal framework for sharing government data for research purposes, with specific provisions requiring that shared data be used for the "public good." Similar legislation in Pakistan — establishing a legal obligation to share administrative data for statistical and research purposes, with appropriate governance mechanisms — would provide the institutional foundation for a more productive data ecosystem.

The World Bank's Development Data Group has articulated this vision in terms of "AI-ready development data," arguing that the challenge is not a scarcity of high-quality data but rather the absence of standardised frameworks and infrastructure to make existing data consistently findable, accessible, and usable (World Bank, 2025). The same argument applies to Pakistan. The data exists. The gap is in the infrastructure — technical, institutional, and legal — that would make it useful. Building that infrastructure is not merely a technical modernisation exercise. It is an investment in the country's capacity to understand itself, to make evidence-based decisions, and to participate in the global economy of knowledge and innovation.

The decisions made in the next few years about how Pakistan's data infrastructure is designed will shape the country's analytical capacity for decades. If the infrastructure is built on open standards, with strong governance, comprehensive metadata, and the flexibility to accommodate new data sources and methods, it will serve as a foundation for continuous improvement. If it is built as a rigid, closed, and poorly documented system, it will become another legacy burden — expensive to maintain, difficult to adapt, and ultimately unable to deliver the evidence that the country needs. The choice is not between a perfect system and no system. It is between a system designed to learn and grow, and one that is not. Pakistan should choose the former.

References

ADR UK (2024). The new UK Government wants a National Data Library: a brilliant aspiration, if built on solid foundations. London: Administrative Data Research UK.

Bipartisan Policy Center (2022). AI-Ready Open Data. Washington, DC: Bipartisan Policy Center.

DSIT (2025). Guidelines and best practices for making government datasets ready for AI. London: Department for Science, Innovation and Technology.

GOV.UK (2025). National Data Library. London: HM Government.

TBI (2025). Governing in the Age of AI: Building Britain's National Data Library. London: Tony Blair Institute for Global Change.

UK Digital Economy Act (2017). *Digital Economy Act 2017*. London: Her Majesty's Stationery Office.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E. and others (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018.

World Bank (2025). From Open Data to AI-Ready Data: Building the Foundations for Responsible AI in Development. Washington, DC: World Bank Development Data Group.

Moving Forward: Priorities and Conclusion

Moving Forward: Priorities and Conclusion

The preceding chapters have set out a case for fundamental change in how Pakistan produces, manages, and uses data for public policy. The argument, in brief, is this: the country's current statistical system — built around periodic large-scale surveys conducted by a single bureau — is no longer adequate for the demands placed upon it. A modern data infrastructure must blend survey data with administrative records and other sources, must be governed by institutions with the authority and capacity to coordinate across agencies, must meet rigorous standards of quality and ethical practice, and must be designed for a future in which artificial intelligence and advanced analytics are standard tools. None of this is controversial in principle. The challenge is in execution.

Vision documents are easy to write and, in Pakistan as in many countries, rarely in short supply. What is in short supply is the disciplined sequencing of concrete actions that can move a system from where it is to where it needs to be. The temptation is always to attempt everything simultaneously — to draft comprehensive legislation, build sophisticated technology platforms, recruit hundreds of specialists, and launch dozens of pilot projects all at once. This approach almost invariably fails. Resources are spread too thin, institutional resistance overwhelms fragmented reform efforts, and early failures discredit the entire enterprise before it has a chance to demonstrate value. The alternative is to sequence reforms carefully, beginning with actions that are achievable in the short term and that create the conditions for more ambitious changes later.

Mapping What Exists Before Building What Is Needed

The first priority is deceptively simple: establishing a comprehensive inventory of what data the government already holds. As discussed in earlier chapters, Pakistan's federal and provincial agencies collectively maintain vast quantities of administrative data — NADRA's biometric records, the Federal Board of Revenue's tax files, BISP's beneficiary databases, DHIS2's health facility records, NEMIS's education data, the Punjab Land Record Authority's digitised cadastral records, and many others. But there is currently no systematic catalogue that tells researchers, policymakers, or even other government agencies what data exists, at what geographic and temporal resolution, in what format, and under what conditions it might be accessed.

A national data inventory — a structured register of government data assets with standardised descriptions of content, coverage, quality, and access procedures — is the essential first step. Without it, every subsequent action is built on guesswork. The inventory need not be perfect or complete from the outset. It should begin with the major federal agencies and expand progressively to include provincial governments and other data holders. The exercise itself will reveal gaps, inconsistencies, and opportunities that are currently invisible. It will also, importantly, begin the process of normalising the idea that government data is a shared national asset rather than the property of the agency that collected it.

This inventory should be led by a cross-agency data working group, convened under the authority of the Chief Statistician and including senior representatives from the principal data-holding agencies.

The working group's mandate should extend beyond the inventory to encompass the development of common data standards, metadata requirements, and governance protocols. Its establishment signals institutional commitment to coordination and provides a forum for resolving the inter-agency tensions that have historically impeded data sharing.

Demonstrating Value Through Pilot Projects

Reform efforts that begin with years of planning before producing any tangible output lose momentum and political support. Pakistan's data infrastructure transformation should instead follow the principle of "start small, learn fast, scale what works." This means identifying two or three pilot projects where blending survey data with administrative records can produce immediate, visible improvements in the quality or timeliness of statistical outputs.

The candidates for such pilots are not difficult to identify. Consider poverty measurement: BISP's household registry contains detailed information on millions of families, collected through door-to-door surveys and progressively updated. Linking this data with FBR tax records, utility consumption data, and school enrolment records from NEMIS could produce more timely and granular estimates of household welfare than the periodic Household Integrated Economic Survey alone can provide. The technical challenges — establishing common identifiers, reconciling different coding schemes, managing privacy risks through the Five Safes framework discussed in Chapter 11 — are real but not insurmountable. And the policy payoff is significant: better-targeted social protection programmes that reach the people who need them most.

A second pilot might focus on health. The DHIS2 system already captures facility-level health data across the country. Linking this with civil registration records (where available) and NADRA's population database could improve the accuracy of health indicators that currently rely on sample surveys with substantial sampling error at the district level. A third pilot might address education, linking NEMIS school data with examination board records and labour force survey data to trace educational pathways and employment outcomes.

The purpose of these pilots is not merely to produce better statistics, though they should do that. It is to demonstrate that blended data approaches work in practice, to build the technical skills required for data linkage within PBS and partner agencies, and to create a track record that justifies the larger investments needed for system-wide transformation. Each pilot should be designed with explicit evaluation criteria, so that its successes and failures can inform subsequent decisions about scaling and institutional design.

Investing in People Before Investing in Technology

No data infrastructure can function without skilled people to operate it. This is perhaps the most important lesson from the international experience reviewed in earlier chapters. The countries that have built successful data linkage systems — the United Kingdom, the Netherlands, Australia, the Nordic states — have invested heavily and consistently in human capacity. They have trained data engineers, data scientists, metadata specialists, privacy officers, and data governance professionals. Pakistan must do the same.

The current skills base within PBS and provincial statistical agencies is heavily oriented toward survey design, fieldwork management, and tabulation — the competencies required by the traditional model. The new infrastructure requires additional competencies: record linkage techniques, statistical disclosure control, data quality assessment for administrative sources, machine learning methods, and the governance skills needed to manage complex multi-agency data sharing arrangements. These skills do not develop overnight, and they cannot be acquired solely through short training workshops.

A serious capacity-building programme should include several elements. First, partnerships with universities — both domestic and international — to develop specialised curricula in data science for official statistics. Second, placement programmes that embed PBS staff in international statistical agencies with established data linkage operations, allowing them to learn through hands-on experience. Third, the recruitment of new staff with backgrounds in computer science, data engineering, and related fields, which will require reforming the civil service recruitment processes that currently make it difficult for statistical agencies to compete for technical talent. Fourth, investment in training for staff at data-holding agencies beyond PBS, since effective data sharing requires that the agencies producing administrative data understand the standards and procedures that make their data usable for statistical purposes.

Establishing the Legal and Institutional Framework

The pilot projects and capacity-building efforts described above can begin under existing institutional arrangements. But sustaining and scaling them will require legal and regulatory reform. Pakistan currently lacks a comprehensive data protection law — the Personal Data Protection Bill has been under discussion for years without enactment. It also lacks a clear legal framework for sharing administrative data for statistical and research purposes. The Statistics Act gives PBS certain authorities, but these were designed for a world in which the bureau collected its own data through surveys, not one in which it coordinate data flows across dozens of agencies.

The legal reform agenda should address several specific needs. First, a data protection law that establishes clear rules for how personal data may be collected, stored, shared, and used, with meaningful enforcement mechanisms and an independent oversight body — but one that does not exempt government agencies from its core obligations, as some earlier drafts have proposed. Second, amendments to the statistical legislation that give PBS (or whatever coordinating body is established) explicit authority to access administrative data from other agencies for statistical purposes, subject to appropriate safeguards. Third, legal provisions establishing that government data shared for research and statistical purposes must be used for the public good, following the model of the UK's Digital Economy Act 2017.

These legislative changes should be informed by the experience of the pilot projects. One of the advantages of starting with practical demonstrations is that they reveal the specific legal barriers that need to be addressed, allowing legislation to be drafted with precision rather than in the abstract. They also build the political constituency for reform: agencies that have seen the benefits of data sharing in practice are more likely to support the legal framework that enables it.

Choosing the Right Organisational Model

Chapter 10 reviewed a range of organisational models for managing data infrastructure, from centralised approaches like the Netherlands' CBS to distributed networks like Estonia's X-Road to intermediate models like Australia's AIHW. The right model for Pakistan will depend on factors that can only be fully assessed after the initial phase of mapping, piloting, and capacity-building. It would be premature to commit to a specific organisational structure before understanding the practical realities of inter-agency coordination, the political dynamics of federal-provincial relations as they affect data sharing, and the absorptive capacity of the institutions involved.

What can be said with confidence is that the organisational model must satisfy certain requirements identified throughout this document: it must have sufficient authority to coordinate data sharing across agencies; it must be operationally independent enough to maintain public trust; it must accommodate Pakistan's federal structure, in which much of the relevant administrative data is held by provincial governments; and it must be sustainable, with funding and institutional arrangements that do not depend on the enthusiasm of individual champions or the priorities of a single political cycle.

The medium-term goal should be to select and begin implementing an organisational model within three to five years of the initial reforms, drawing on the evidence accumulated through the pilot projects and the institutional learning generated by the data working group. This is not a delay. It is a recognition that institutional design decisions made without adequate evidence tend to produce institutions that look impressive on paper but fail in practice.

The Sequencing Principle

The priorities outlined above are not a checklist to be completed in parallel. They are a sequence, in which each step creates the conditions for the next. The data inventory reveals what is available and what is missing. The pilot projects demonstrate what is possible and build technical capacity. The capacity-building programme ensures that institutions have the people needed to sustain and expand the work. The legal reforms remove the barriers that limit scaling. And the organisational model provides the institutional home for the mature system.

This sequencing is itself a form of risk management. By starting with low-cost, reversible actions — mapping data, running pilots, training staff — Pakistan can learn what works before committing to expensive and difficult-to-reverse institutional changes. If a pilot project reveals that a particular data source is unusable, or that a particular agency is unwilling to cooperate, these lessons can be incorporated into the design of subsequent phases without having wasted large investments on assumptions that proved wrong.

The country has spent decades discussing statistical reform. It has produced numerous reports, strategies, and action plans, many of which contain sensible recommendations that were never implemented. The difference this time must be in execution: in the willingness to begin with concrete, modest, achievable steps rather than waiting for a comprehensive plan that satisfies everyone, and in the discipline to learn from each step and adjust course accordingly. The data infrastructure Pakistan needs will not be built in a single leap. It will be built incrementally, through sustained effort, institutional learning, and the gradual accumulation of capability, trust, and demonstrated value.

Why This Matters

It is worth stepping back, in closing, to recall why any of this matters. Statistics are not an end in themselves. They are the means by which a society understands its own condition — how many children are in school and whether they are learning, how many people are in work and whether their wages are rising, how disease burden is distributed and whether health services are reaching those who need them, how public money is being spent and whether it is achieving its intended purposes. A society that cannot answer these questions reliably is a society governing in the dark.

Pakistan faces enormous challenges: a young and rapidly growing population that needs education and employment, a health system under strain, an economy that must become more productive and more equitable, a climate that is becoming more hostile to agriculture and human settlement. Addressing these challenges requires not only political will and financial resources but also evidence — evidence about what is happening, where, to whom, and why. The data infrastructure described in this document is the foundation on which that evidence must be built.

This is not just a technical challenge, though the technical demands are real. It is a governance challenge, because data must flow across institutional boundaries that have historically been impermeable. It is a legal challenge, because the frameworks governing data collection, sharing, and protection must be modernised. It is an ethical challenge, because the rights of data subjects must be protected even as their information is put to new uses. And it is, ultimately, a political challenge, because it require commitment from the highest levels of government and a willingness to change how things have been done for decades.

But the potential rewards are proportionate to the difficulty. A functioning data infrastructure — one in which trusted, well-governed, interoperable data can be combined across sources to produce timely and accurate evidence — would transform not only the statistical system but the quality of governance itself. It would enable social protection programmes that actually reach the poorest households, health interventions that target the communities most at risk, education policies informed by actual learning outcomes rather than enrolment figures alone, and economic strategies grounded in evidence about what is working and what is not.

Data in silos serves no one well. The information exists, scattered across agencies and systems that do not communicate with each other. The task is to build the infrastructure — technical, institutional, legal, and human — that connects these fragments into a coherent picture of the country. Pakistan's people deserve a statistical system equal to the complexity of their lives and the urgency of the challenges they face. Building it is not optional. It is a precondition for the kind of governance that a country of 240 million people requires.