Data & Policy (2024), 6: e38
doi:10.1017/dap.2024.30

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

What does a record have to do with it? Re-situating records management within Indian public data governance and policy

Srijoni Sen and Trishi Jindal

Law, Technology, and Society Initiative, National Law School of India University, Bengaluru, India
Corresponding author: Srijoni Sen; Email: srijonisen@nls.ac.in

Received: 19 July 2023; **Revised:** 29 November 2023; **Accepted:** 18 May 2024

Keywords: data gaps; data governance; open data; record management; state capacity

Abstract

A number of data governance policies have recently been introduced or revised by the Indian Government with the stated goal of unlocking the developmental and economic potential of data. The policies seek to implement standardized frameworks for public data management and establish platforms for data exchange. However, India has a longstanding history of record-keeping and information transparency practices, which are crucial in the context of data management. These connections have not been explicitly addressed in recent policies like the Draft National Data Governance Framework, 2022. To understand if record management has a role to play in modern public data governance, we analyze the key new data governance framework and the associated Indian Urban Data Exchange platform as a case study. The study examines the exchange where public records serve as a potential source of data. It evaluates the coverage and the actors involved in the creation of this data to understand the impact of records management on government departments' ability to publish datasets. We conclude that while India recognizes the importance of data as a public good, it needs to integrate digital records management practices more effectively into its policies to ensure accurate, up-to-date, and accessible data for public benefit.

Policy Significance Statement

In this study, we focus on urban data exchanges as an example of public data governance and argue that there is a risk of overlooking administrative records-based data coverage. Urban data includes various types of information, such as demographic statistics, financial data, infrastructure details, environmental indicators, and socio-economic trends. To ensure the reliability, consistency, and comparability of this data, standardized protocols, and best practices should be followed during the data generation process. Neglecting these fundamentals could result in the accumulation of data hordes that serve no meaningful purpose. Therefore, when prioritizing initiatives like data sharing and exchange, it is essential to invest in state capacity by reforming public record management and integrating the principles outlined in this paper into digitalization efforts in public administration. The challenges identified in this paper regarding data coverage and quality are relevant to many developing countries. It is crucial to address these challenges by improving the capacity of the government to collect and maintain reliable data.

1. Current Indian approaches in data governance policy

The Indian government, over the last 3 years, has released a number of policies on data governance. These policies cover various aspects ranging from data sharing, data empowerment and usability, to data

protection and privacy, to harness its rapidly growing digital economy.[1] Along with overarching national policies, sectoral and regional data governance policies have also multiplied, particularly in health and finance, and at the sub-national level for regional priorities. These policies encompass personal, non-personal,[2] commercial, and government data. They range across draft policy documents, proposed and enacted laws and regulations, and data architectures and protocols. A common foundation for these policies is the "Digital India" mission, which aims to "transform India into a digitally empowered society and knowledge economy." (MEITY, n.d.)

A preliminary review of the policies and surrounding documents and statements reveals some common priorities. Chief among them is the assertion that data sharing is key to economic empowerment. This is in consonance with developments in the EU and in China—across these countries, "unlocking" data for economic development has emerged as an explicit priority, incentivizing data sharing and reducing barriers for public and private domestic actors (Kak and Sacks, 2021).

In 2022, the Indian government asserted its evolving policy to encourage data sharing under the draft National Data Governance Framework (NDGF). This document outlined the central government's updated vision for sharing government data for public benefit. It asserts that "The power of this data must be harnessed for more effective Digital Government, public good, and innovation." Similarly, the policy on non-personal data aims at "unlocking economic benefit from non-personal data for India and its people." Across these efforts, a common belief is that valuable data exists in silos or corporate vaults, and policy should prioritize freeing this data up for more economic actors, public and private, to realize its value. The key challenge, in this framing, is the lack of data *flows*.

Where is this data to come from? The data governance policies make it clear that sharing of both public and privately owned data is critical. With private entities understandably reluctant to engage in regulated data sharing, real-world implementations of these policies have commenced with a focus on government data. From government portals to newer public data exchanges, this approach has prioritized platform-based facilitation of government data sharing. With cities being characterized as "engines of the new data economy" (Barns, 2018), urban data generated by government agencies has become a key priority testing ground for these efforts. India's flagship "Smart Cities Mission" repeatedly cites the use of data and information to improve urban infrastructure and services, characterizing data "at the core of this new thinking around technology as an enabler to drive growth" (Ministry of Housing and Urban Affairs, 2015, 2020).

These data governance policies and related initiatives have met with a variety of critiques: industry representatives have highlighted intellectual property rights and business realities, and scholars and civil society actors question the potential for "empowerment," and highlight the risks to privacy and security. One core assumption in these policies, however, remains relatively under-examined: that of valuable data already in existence, ready to be shared and utilized. This overlooks critical challenges of data availability and quality.[3]

Giest and Samuels (2020) define data gaps as primary (known to the government), secondary (poor quality data in certain categories), and hidden (unknown). Lerman (2013) articulates the risks to those excluded by big data and argues for a right against data exclusion. Herrera and Kapur (2007) acknowledge the data quality problem as they remind us that actors that produce data, including data collection agencies, public authorities, NGOs, and academics, do so while facing problems of agency, capacity, and misaligned incentives.

In India, these issues have been raised from time to time, particularly regarding the availability and quality of government statistics (Agrawal and Kumar, 2020; Rukmini, 2021; BBC News, 2023). In the

[1] We adopt Kak and Sack's understanding of data governance, in the context of public policy, as "a rapidly expanding body of law and other (softer) policy frameworks that regulate access to and transfer of data between different entities in the digital economy" (Kak and Sacks, 2021).

[2] Non personal data (NPD) is generally data that does not contain personally identifiable information. This includes both data which never related to a natural person (e.g., weather data, data from sensors in industrial machines), and data which has been anonymized (Ministry of Electronics and Information Technology, 2020).

[3] We adopt the elements discussed by Herrera and Kapur (2007), from a political science perspective, of *validity, coverage* and *accuracy* as key elements of data quality.

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

data governance context, however, more attention is paid to more downstream elements such as data sharing or exchange. As a result, a study of India's open data in 2018 revealed significant asymmetries in data coverage, with certain states, or departments within states, contributing far more data than others (Saxena, 2018). The focus on data quality reveals significant challenges across sectors. India's GDP data quality is graded a "D-Poor" by an international data organization (World Economics, 2023). In the healthcare sector, concerns regarding the official health and mortality statistics have been raised consistently and with increased urgency during the COVID-19 pandemic (Mulye, 2021; Vasudevan et al., 2021).

While there are several aspects to addressing the coverage and quality of government data, this paper uses a combination of quantitative data analysis and analysis of applicable laws and policies to argue that robust records-management values and practices form a key pillar of successful data empowerment, which has hitherto been neglected in India to the detriment of its open data ambitions. For the data analysis, metadata records of the Indian Urban Data Exchange platform are analyzed (Section 3.2) to categorize and summarise the composition of data sources on the IUDX platform. The legal and policy analysis (Section 4) examines key policies, laws, and documents to identify gaps between data governance approaches and public records management principles. These include the draft National Data Governance Framework Policy 2022, the Public Records Act 1993, the Right to Information Act 2005, as well as operational guidelines and announcements.

2. Records, data, and information

Records management and data governance have been described as two different discourses with different conceptual frames but similar concerns (Borglund and Engvall, 2014). Depending on the primary disciplinary perspective, some scholars and practitioners argue that records management is only one aspect of information management, or in other words, while all records are data, all data are not records (Mcleod and Hare, 2006; Öberg and Borglund, 2006). The International Organization for Standardization defines records as "information created, received and maintained…by an organisation or person, in pursuance of legal obligations or in the transaction of business." In this definition, they seem to follow this approach of regarding records as a subset of information. Others in the domain of records-management maintain that information should in fact be managed as records. The gradual takeover of digital technologies in both data management and record-keeping professions has made matters more complex, arguably at a greater cost to record-keeping practices (Yeo, 2018).

Practitioners working in developing economies largely view records as critical source documents from which data are derived, and the quality of record-keeping has a major impact on the quality of derived data and statistics (Thurston, 2012). They would argue, for example, that accurate data on the COVID-19 pandemic has much to do with the strength of record-keeping practices of a government's civil registry system, just as progress in gender representation in the government has to do with its employment records. In practice, record keeping and data quality have been closely linked; for example, in the World Health Organisation's guide to improving data quality (World Health Organisation, 2003).

In most developing countries, however, the advent of digital governance mechanisms has been accompanied by a shift of focus away from records management. Lemieux (2016) asserts that while in 1996, the national archives played an active role in records management in 71% of developing countries, the landscape transformed in the next 20 years, with the primary responsibility for managing records and related data falling to ICT authorities.

Others have argued that the volume of born-digital data, coupled with modern data capabilities, makes current record-keeping practices too rigid and unsustainable (Ranade, 2016). For the purposes of this paper, however, we follow Geoffrey Yeo (2018), who argues that because modern, powerful data systems are still subject to political influence and human values, norms and concerns found in record-keeping practices, such as persistence and integrity, are relevant in shaping the data we place so much faith on.

Even where access to more advanced technologies is possible, for example, transparency and accountability movements have at times chosen to rely on in-person meetings and written reports,

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

because the priority is the generation of usable *information*, not the application of a particular technology. The reliability of certain older modes and practices makes the generation of usable information more predictable, while the introduction of new technologies without accompanying updates of norms and guidelines may have the reverse effect (Lemieux, 2016). New forms of record and data production bring with them new lines of accountability and oversight, but also new challenges to data quality—rapid data production through increasingly system-driven modes requires increased attention to the reliability of the data and the responsibility for its quality (Agostino et al., 2022). While the government's focus may, therefore, be on digitization alone, the risks of erosion of trust, integrity of information, and even loss of data may result from rapid shifts to electronic information processes.

Regardless of whether one takes a records-first or data-first approach, certain key practices that flow from records management are critical to address issues of data quality, coverage, and reliability (Jaeger and Bertot, 2010; Lemieux, 2016; Casadesús de Mingo and Cerrillo-i-Martínez, 2018; Yeo, 2018). We enumerate a few such critical processes and practices that have been identified in the literature—First, processes to ensure the authenticity and integrity of documents, particularly in the face of intangible and malleable digital sources, coupled with poor integration, duplication, and discrepancies. Second, the contextualization of information, that is, the authorship, provenance, and ability to trace changes and decisions over time, which expresses itself, to some extent, in data governance as "metadata management." Third, preservation of information beyond immediate needs, and fourth, identification of what to retain among the massively expanding volume of born-digital elements. These administrative practices can be seen as foundational to the generation of high-quality, reliable data that is consistently available. We turn now to the frameworks and practices on public data governance in India to see if they operationalize these principles in the Indian context.

3. Public data: policies and portals

3.1 Public data policies

Propelled by global open data movements,[4] the Indian government introduced the National Data Sharing and Accessibility Policy (NDSAP) in 2012, aimed at facilitating the sharing of government data with various stakeholders. At the time of publication of this paper, with the proposed National Data Governance Framework Policy, 2022 (NDGF) still in draft form, the NDSAP continues to be the operational policy on public data sharing in India. Several states have also created their own data policies, which reflect a similar prioritization of the economic potential of open data (Panjiar and Waghre, 2022).

The introduction of the NDSAP in 2012 was followed by the launch of an open data portal, data.gov.in, where government agencies were expected to upload at least five "high-value"[5] datasets on the platform in the first 3 months, and remaining datasets within a year. Datasets were to be periodically updated and contain comprehensive metadata to enable discovery and access. Similar efforts, either with a sectoral focus on agricultural or urban data or more comprehensive but state-specific portals, were launched by state governments as well. While the data.gov.in portal remains the main effort by the central government on open data, more recent proposals and efforts indicate a shift in focus to analytics (for example, the National Data and Analytics Platform by the NITI Aayog) and data exchanges in the proposed National Data Platform as a unified data exchange for India.

In 2017, an assessment of the data.gov.in portal indicated that 40,000 datasets had been published on the platform, with close to 40% being sourced from the 2001 and 2011 Census (Misra et al., 2017). Most datasets were already publicly available elsewhere, and added value in terms of quality, searchability, and

[4] For a history of the development of the NDSAP, and its relationships to the Right to Information Act, see Chattapadhyay (2014).

[5] While different departments could arrive at their own understanding of high-value datasets, a common articulation can be found in the Report by the Committee of Experts on Non-Personal Data Governance Framework (2020) which broadly defines high-value datasets as "beneficial to the community at large," and useful either in policy-making, for research and education, or in job creation and expansion of business opportunities (clause 7.6). The understanding is that such data is more likely to be requested for its quality, aggregate nature and higher utility (Barbosa et al., 2014; Yin, 2023).

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

openness continued to be an issue. This meant that use of the data portal was low, and those who did use it, found it unsatisfactory (Agarwal, 2015; Larquemin et al., 2016). A key challenge, according to Government departments, was the conversion of available data into open data formats (Misra et al., 2017), indicating that the innovation centered around openness and sharing rather than data availability and quality. As a result, datasets have been found to be incomplete, static, and out-of-date, with missing metadata.

This is recognized in the draft NDGF 2022, which has been introduced as a revision of the NDSAP framework. It frames the current scenario as follows: "…Digital Government data is currently managed, stored, and accessed in differing and inconsistent ways across different government entities, thus attenuating the efficacy of data-driven governance." Accordingly, the Policy articulates the ambition "to enable and catalyze vibrant AI and Data led research and Start-up ecosystem, by creating a large repository of Indian datasets."

In order to understand the ways in which the NDGF approach is working in practice, we turn now to a data exchange initiative that seeks to demonstrate the approaches outlined in the NDGF 2022 in the urban data context.

3.2 Data practices in public urban data

Urban data as a category is a good exemplar of the approach to data governance outlined in Section 1 of this paper. The launch of the ambitious Smart Cities Mission in 2015, pledging 90 billion dollars for technology-driven development projects in 100 cities in India, was accompanied by an increasing emphasis on the economic and developmental potential of urban data (Ministry of Housing and Urban Affairs, 2015). Projects funded under this mission tended to prioritize digital monitoring, automation, and overall, a digitalization of urban governance. Data-driven urban management and policy-making fall squarely within this framing.

Globally, it is recognized that sources of urban data come from legacy systems across a range of government and private actors at multiple organizational layers, which in turn contributes to institutional tensions (Dawes et al., 2016; Gupta et al., 2020). In New York City, for example, there is a mandate for all city government agencies to publicly list datasets under their control. This includes agencies such as the health department's inspection records, as well as the location and costs of housing projects from the housing department. While some of these data sources are automatically generated and updated by the agencies' system, others require coordination and manual intervention with each agency (Dawes et al., 2016).

In India, the policy push on data governance, and on Smart Cities, has resulted in new digital initiatives on urban data, including a Smart Cities Open Data Portal, which focuses on the publication of datasets, as well as the India Urban Data Exchange (IUDX), launched in 2021. Data exchanges, in contrast to portals, enable multiple agents to publish, request, and consume data and potentially create a "many to many" data relationship among urban agencies, research organizations, commercial entities, and non-profits. The emphasis of the IUDX, as with the national data governance policies, is the movement of data across current departmental "silos." The government companies created as Special Purpose Vehicles for the implementation of Smart Cities projects in every city are key stakeholders in the collection and publication of this urban data.

In this paper, we take a closer look at data availability in the IUDX platform, which exemplifies current approaches to public data governance in India. Exchanges and portals are not tasked with the generation of complete and high-quality data—instead, they articulate their task as *finding* data and breaking silos to enable sharing (see Figure 1). Sources of the data can be either public or private, as indicated in Figure 2, and have also been classified by source, as emerging from sensors or devices on the one hand, and repositories or databases, on the other, for data relating to property, utilities and legal processes. The IUDX Overview further elaborates on its data sources as "Some of the data consists of streams of IOT data from installed sensors (e.g., Air Quality, Traffic), some of the data is demographic or geographical, some may be from municipal tax or property records, some from legal documents or registrations, and some may be historical data from archival sources" (MoHUA and MEITY, 2021).

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

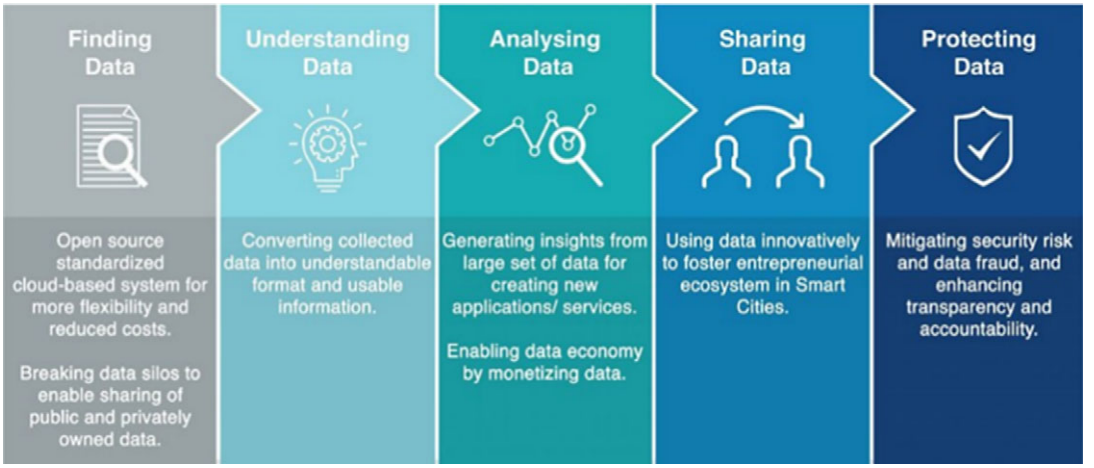

Figure 1. *Data management lifecycle (Source: IUDX: Overview (2021)).*

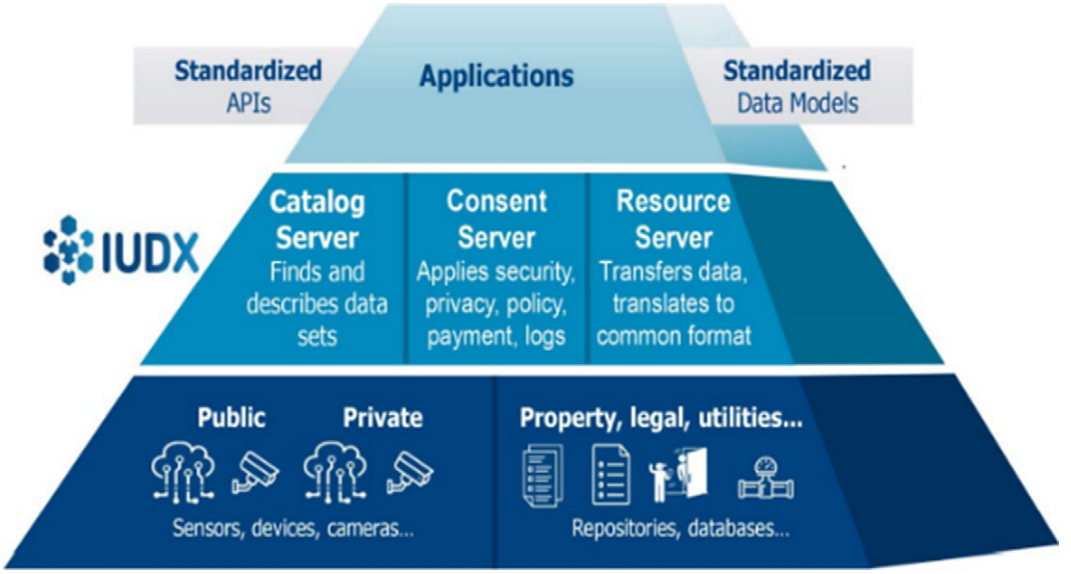

Figure 2. *Architecture and data sources (Source: IUDX: Overview (2021)).*

To assess data coverage in the IUDX platform, we constructed a dataset from the publicly available metadata, as on July 2023, from 381 published datasets on the IUDX portal from 36 different Indian cities. Based on the literature review, the scope described by the IUDX, and the observation of the "name" and "parameters" labels of each dataset, we classified the IUDX datasets into three categories based on the source of the data: "GPS-Based Data" for spatial or geographic data derived from the GPS satellite-based navigational system combined with GIS systems; "Sensor-Based Data," for datasets derived from installed sensors; and "Records-Based Data" for datasets derived from administrative systems and processes. A limitation of this analysis is that it only speaks to aspects of source and coverage, but not of quality or accuracy, that would require further assessment. Figure 3 indicates the overall composition of these three categories in our dataset. Figure 4 indicates the specific types of data included in each of these three categories.

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

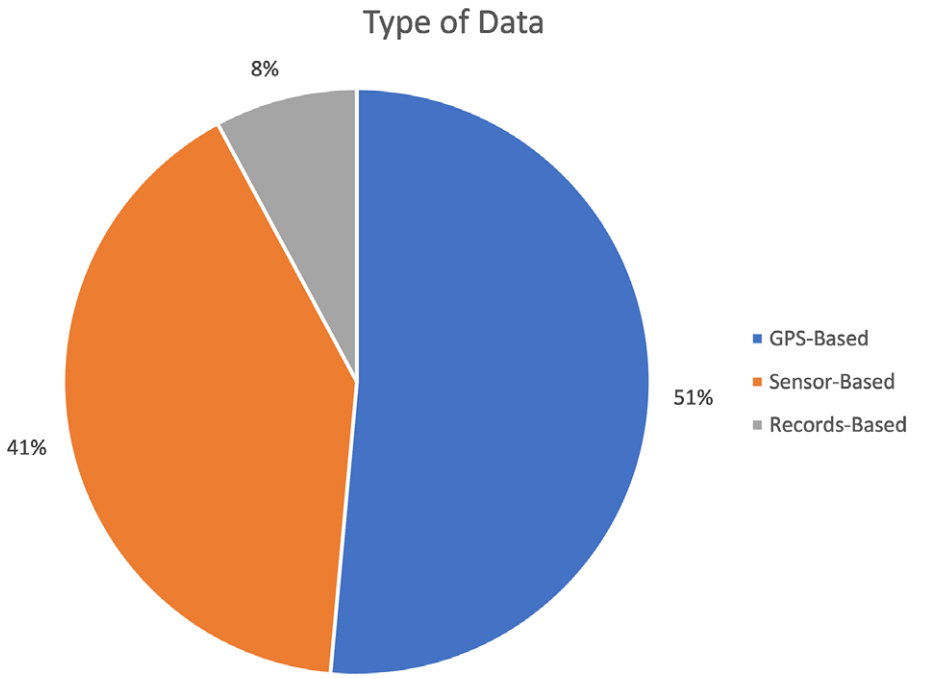

Figure 3. *Composition of data sources in the IUDX Urban Data Exchange (Source: Authors' analysis).*

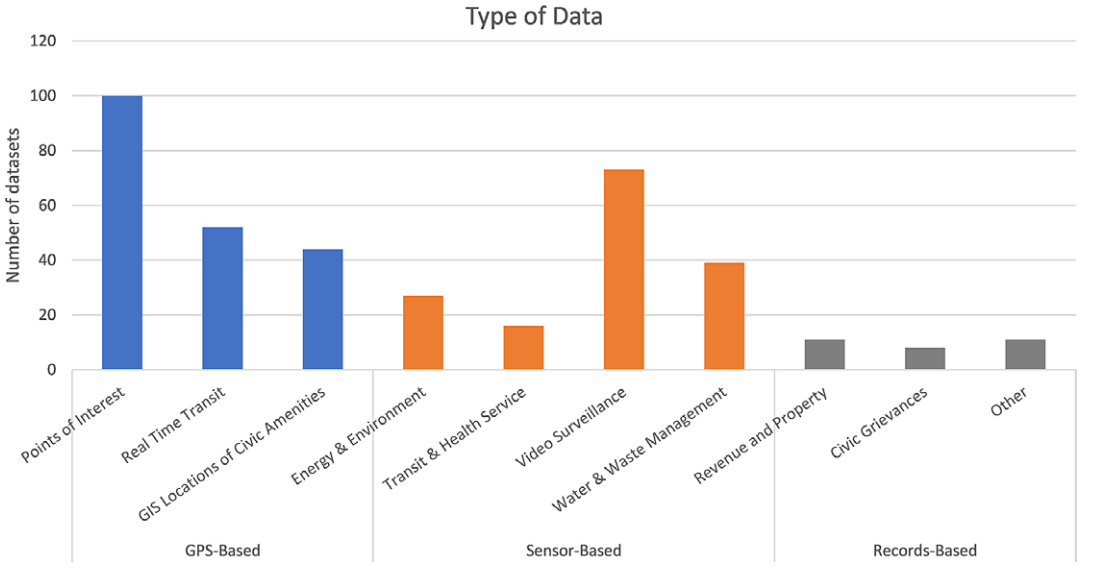

Figure 4. *Types of datasets on IUDX in each categorized data source (Authors' analysis).*

From Figure 3, we observe that GPS and Sensor-Based data, where both the capture and upload of data are likely to be automated, comprises >90% of the datasets in the IUDX platform. Records-Based data, which depends more on administrative procedures across different urban authorities, as well as typically on manual data capture and upload, form only 8% of the 381 datasets analyzed.

Figure 4 provides a more granular picture. Each of our categorizations contains data types that are potentially valuable to a number of different users, as intended by public data exchanges. GPS data conveys information about transit and traffic systems, road networks and places of interest for tourism

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

purposes—data that can be used by public authorities in policy implementation, as well as commercial players and researchers. Similarly, sensor-based data sheds light on a range of critical civic amenities managed by different public authorities, from air quality monitoring, solar panel performance, waste disposal, and water and flood management.

Figure 4 also demonstrates, however, potential data gaps in urban data collection and exchange. Only 30 records in the dataset, from 14 cities, related to public administrative information such as municipal revenue collection and budgets, redressal of civic grievances, or toll collections. The Surat Municipal Corporation was the only source to have shared its revenue collection data, which included collections of property tax and professional tax. It is also interesting to note that where records-based data is available, it seems to be sourced from agencies that have transitioned to automated record management practices: the "Civic Grievances" source, for example, comprised data collected from various citizen grievance portals and apps, that allow for automated data generation. Missing from this set, however, are other kinds of important records-based data, from crime and safety records (a frequent, though contested inclusion in urban data portals around the world), traffic enforcement data, financial information of public agencies, or data on municipal records and urban development schemes.

We argue that the availability of critical categories of urban data depends on robust administrative systems and processes that can generate "high-value" data of the sort contemplated by the new Indian data governance policies. With the lower availability of such data categories in the IUDX dataset, we look at whether the operation of record-management legislation in India explains this outcome, and whether new data governance policies such as the NDGF address these issues.

4. Public records legislation and systems in India

Discussions on data availability and quality have tended to overlook the statutory frameworks that govern them (Kumar, 2020). We address this gap in the context of public records in India by looking at its governing legislation, namely the Public Records Act, 1993 (PRA) and the Right to Information Act, 2005 (RTI). These laws define "public records," not based on whether they are publicly available or slated to be. Instead, the definition is based on the source of the record, that is, a government department, agency, or undertaking.

Although the objective of the PRA is cast broadly (to "regulate the management, administration and preservation of public records"), its purpose in fact was to introduce legislation on the National Archives, which had been in operation from the 1930s without legislative backing. Nevertheless, the Act did in fact deal with the entirety of the public record management lifecycle: creation, classification, storage, retrieval, and archiving or destruction. It mandates the creation of a Records Officer and records room in every department. However, as with the appointment of Data Officers under the NDSAP, compliance with this provision remains low (Prasad, 2013; Sen and Jindal, 2022). The Records Officer is mandated to arrange, maintain and preserve public records, create retention schedules, and compile indices in accordance with standards and guidelines developed by the National Archives (Section 6, PRA, 1993). At an operational level, the Public Records Rules, 1997 and the Central Secretariat Manual of Office Procedure (CSMOP) detail procedures for both paper and digital management, retention, and sharing of government records and files. These standards speak to the practices on authenticity, traceability, integrity, and preservation identified in the literature in Section 2. For example, the Manual establishes procedures for numbering, classification systems and movement of government files, and conformity to the Records Retention Schedule based on the sub-category of records (financial, personnel-related, substantive decision-making, and so on). Equivalent laws, rules, and manuals are to be found at the state level as well.

Unclassified public records under the PRA that are more than 30 years old are to be transferred to the National Archives (Section 12, PRA, 1993). The RTI Act, on the other hand, focuses on contemporaneous access to information. The RTI Act enables access to contemporary and past information through (i) proactive disclosures by departments, including a mandate to departments to disclose a list of their data sets, and (ii) the release of information in response to RTI queries. The RTI Act was enacted to fulfill the distinct right to "know," which has been guaranteed under the right to freedom of speech and the right to freedom of the press under the Indian Constitution.

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

Both these laws define public records broadly. The catch-all nature of the definition helps include all communications, executive decisions, laws, and regulations, as well as information collected from individuals and private entities held by government departments, agencies and undertakings, regardless of whether they are open to public access, or freely available in the public domain. All of this can be, and has been, viewed as crucial building blocks to public data and information. While the two Acts work in tandem with each other, both set out distinct frameworks for record management and access. At the same time, laws like the Official Secrets Act, 1923, the laws on intellectual property, information privacy and security circumscribe records that may be accessed by the public.

In spite of multiple legal and policy frameworks, scholars, public policy practitioners, and government actors have all repeatedly cited limited capacity, time, budgetary outlay and incentives as challenges to effective information practices (Gautam, 2007; Prasad, 2013). These issues manifest across functions—inconsistent indexing and tagging metadata on records; maintaining accurate, up-to-date, and synchronized records; financial reporting, publication of legal instruments and enforcement actions; fulfilling legal obligations for proactive disclosures and access to information requests, etc.

The task of balancing privacy, national security, intellectual property, and related interests while making reliable and accurate information and records available for public or restricted access is not an easy one. In addition to the multiplicity of governing frameworks, the responsibility of implementing policies is also unclear. An analysis by the authors revealed five different Ministries that are tasked with crucial aspects of data and record management, including—confusingly—the Ministry of Culture, which is responsible for the oversight of the Public Records Act, due to its close relationship with the National Archives (Sen and Jindal, 2022).

Moreover, the PRA definition of public records is based on the analog and paper-based communications that were of chief concern at the time of its enactment. While the definition is broad enough to include computer-generated records, the methods of management, and indeed even an understanding of public records for "digital" records of governments, has not been updated since (Gautam, 2007).

In practice, Indian government agencies employ a wide range of practices, from paper-based workflows similar to 19th century practices to born-digital record and data generation practices using cutting-edge technologies. A significant proportion of records exist first as physical paper that is signed, scanned, uploaded, and digitally signed. The process adds discrepancies, which have been particularly noted in land record management. Even with born-digital records, manual entry of data is an inevitable step that is witness to a high degree of inconsistency and error, arising from lack of technical capacity among data entry operators, inconsistent procedural guidelines, and deliberate discrepancies motivated by corruption (Sen and Jindal, 2022). From a data policy perspective, this weakens the generation of reliable, complete, and accurate data, which is a critical first step in the data governance lifecycle.

As mentioned earlier, the shortcomings of public data management are recognized in the new policy framework, that is, the draft NDGF 2022. It addresses this with a broad goal to "transform and modernise Government's **data collection and management processes and systems** through standardised guidelines, rules and standards for the collection, processing, storage, access, and use of Government data" (emphasis supplied). The scope of the policy is not limited to data access or exchange but includes all aspects of the data governance lifecycle. Frameworks of how such data collection or management is to be conducted, however, are not mentioned. The NDGF envisages a new India Data Management Office and India Data Council as part of the institutional infrastructure. The data management office is tasked with creating the capacity for standardized data management practices across government departments. In turn, government departments are directed to set up Data Management Units.

The policy, as an overall vision document, does not go very far beyond the articulation of its goals and the responsibilities of the proposed data management offices. It does not refer to extant legislation such as the PRA and the RTI Act. While agencies have been given a broad mandate to create "detailed, searchable data inventories," the work of standard-setting has been left to the proposed Data Management Office. This includes standards for identification of datasets, data quality and metadata standards, access and availability, and usage. While it is appropriate to leave the work of substantive standard-setting to a dedicated unit, what is also missing is an articulation of key principles that are fundamental to "data

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

quality," or the creation of "high-value datasets" such as the ones identified in Section 2 of this paper, of accuracy, integrity, and traceability. Neither does it set out any departmental obligations for the identification and publication of data, even though lack of departmental response has been identified as one of the major impediments to the success of the 2012 NDSAP policy.

We conclude, therefore, that initiatives such as the NDGF 2022 and the IUDX Portal depend on the existence of high-quality data in government departments. In turn, these datasets rely, at least in part, on effective record management, where specific data elements and indicators are established by organizing and indexing files and records before the data collection process. In this context, efficient records management plays a crucial role in the government's digital transformation efforts. Indian law and policy mandates record-keeping practices that are important for public accountability and data access. However, legislative standards set out in the PRA are overlooked, and new-age frameworks like the NDGF have not yet referred to, or elaborated on, these standards. Moreover, state capacity in the implementation of these policies remains of significant concern.

Conclusion

The challenges identified in this paper with respect to data coverage and quality are of common concern across most developing countries (World Health Organisation, 2003; Zhao et al., 2022). Also common is the need for enhancement of state capacity across the board. In this paper, we demonstrate, in the context of urban data, that in the new era of automated public systems, records-based data coverage and quality may be at particular risk of being overlooked.

These critical categories of urban data encompass a wide range, including demographic and financial statistics, infrastructure details, environmental indicators, and socio-economic trends. The administrative processes involved in such data generation should adhere to standardized protocols and best practices to guarantee the reliability, consistency, and comparability of the produced data that records-management processes have earlier set out to do. Without these fundamentals, data platforms and exchanges may be inundated with "zombie data" that is easy to release but serves no meaningful purpose (Gurin, 2014).

We argue in this paper, therefore, that when policy priorities are focussed on new efforts such as data sharing and exchange initiatives, new data governance policies, and consequent investment in state capacity should not overlook foundational processes in favor of new technologies and systems. This may take the path of reform of public record management per se, or merely an adoption of the principles of record-keeping identified in this paper, to new digitalization efforts in public administration. Often, such efforts reveal the tensions between open data rhetoric and meaningful transparency, and demonstrate that the obstacles are as political as they are about building state capacity (Kaufmann and Bellver, 2005). Adherence to enacted legislative principles and accompanying protocols is critical, therefore, to ensure sustained transparency and developmental goals.

Data availability statement. Replication data can be found at this Kaggle link: https://www.kaggle.com/datasets/srijonisen/indian-urban-data-sources

Acknowledgements. The author is grateful for the administrative support provided by the National Law School of India University. They would also like to thank the participants of the 2022 Plenary Workshop on Digital Public Records conducted at NLSIU and the participants of the Seattle edition of the Data for Policy Conference, 2022 for their insightful comments on early versions of this paper. A preprint of this paper is available on https://ssrn.com/abstract=4505408.

Author contribution. Conceptualization: S.S., T.J.; Data curation: S.S., T.J.; Formal analysis: S.S., T.J.; Funding acquisition: S.S.; Investigation: S.S., T.J.; Methodology: S.S., T.J.; Project administration: S.S.; Visualization: S.S.; Writing—original draft: S.S.; Writing—review & editing: S.S., T.J. All authors approved the final submitted draft.

Funding statement. This research was supported by grants from the Thakur Family Foundation. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interest. S.S. is a member of the Data Ethics and Policy Committee of the IUDX platform as an independent expert. All data and information about the IUDX platform used in this article are from publicly available sources.

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

References

Agarwal N (2015) Open government data: An answer to India's growth logjam. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.2822676.

Agostino D, **Saliterer I and Steccolini I** (2022) Digitalization, accounting and accountability: A literature review and reflections on future research in public services. *Financial Accountability & Management 38*(2), 152–176. https://doi.org/10.1111/faam.12301.

Agrawal A and Kumar V (2020) *Numbers in India's Periphery: The Political Economy of Government Statistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108762229.

Barbosa L, *et al.* (2014) Structured open urban data: Understanding the landscape. *Big Data 2*(3), 144–154. https://doi.org/10.1089/big.2014.0020.

Barns S (2018) Smart cities and urban data platforms: Designing interfaces for smart governance. *City, Culture and Society 12*, 5–12. https://doi.org/10.1016/j.ccs.2017.09.006.

BBC News (2023) Census in India: Baffling lack of data is hurting Indians. *BBC News.* Retrieved from https://www.bbc.com/news/world-asia-india-64282374.

Borglund E and Engvall T (2014) Open data? Data, information, document or record? *Records Management Journal 24*(2), 163–180. https://doi.org/10.1108/RMJ-01-2014-0012.

Casadesús de Mingo A and Cerrillo-i-Martínez A (2018) Improving records management to promote transparency and prevent corruption. *International Journal of Information Management 38*(1), 256–261. https://doi.org/10.1016/j.ijinfomgt.2017.09.005.

Chattapadhyay S (2014) Opening government data through mediation: Exploring the roles, practices and strategies of data intermediary organisations in India. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2549734.

Dawes SS, **Vidiasova L and Parkhimovich O** (2016) Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly 33*(1), 15–27. https://doi.org/10.1016/j.giq.2016.01.003.

Gautam, M (2007) Electronic records management: Challenges and issues—A case study—National Archives of India. *Atlanti 17* (1–2). Retrieved from https://www.dlib.si/stream/URN:NBN:SI:DOC-A52013ZE/49d793e7-40c9-45aa-9a11-3ea607334dc1/PDF.

Giest S and Samuels A (2020) 'For good measure': Data gaps in a big data world. *Policy Sciences 53*(3), 559–569. https://doi.org/10.1007/s11077-020-09384-1.

Gupta A, **Panagiotopoulos P and Bowen F** (2020) An orchestration approach to smart city data ecosystems. *Technological Forecasting and Social Change 153*, 119929. https://doi.org/10.1016/j.techfore.2020.119929.

Gurin J (2014) Open governments, open data: A new lever for transparency, citizen engagement, and economic growth. *The SAIS Review of International Affairs 34*(1), 71–82. https://doi.org/10.1353/sais.2014.0009.

Herrera YM and Kapur D (2007) Improving data quality: Actors, incentives, and capabilities. *Political Analysis 15*(4), 365–386. https://doi.org/10.1093/pan/mpm007.

Jaeger PT and Bertot JC (2010) Transparency and technological change: Ensuring equal and sustained public access to government information. *Government Information Quarterly 27*(4), 371–376. https://doi.org/10.1016/j.giq.2010.05.003.

Kak A and Sacks S (2021) *Shifting Narratives and Emergent Trends in Data-Governance Policy.* Yale Law School-Paul Tsai China Center.

Kaufmann D and Bellver A (2005) Transparenting transparency: Initial empirics and policy applications. *SSRN Electronic Journal* https://doi.org/10.2139/ssrn.808664.

Kumar V (2020) Census laws and the quality of census data: The limits of punitive legislation. *Statistical Journal of the IAOS 36*(4), 1143–1160. https://doi.org/10.3233/SJI-200651.

Larquemin A, **Mukhopadhyay JP and Buteau S** (2016) Open government data and evidence-based socio- economic policy research in India: An overview. *The Journal of Community Informatics 12*(2).

Lemieux VL (2016) *One Step Forward, Two Steps Backward? Does E-Government Make Governments in Developing Countries More Transparent and Accountable?* Washington, DC: World Bank. https://doi.org/10.1596/23647.

Lerman J (2013) Big data and its exclusions. *Stanford Law Review Online 66*, 55–64.

MEITY (n.d.) Introduction—Digital India. Retrieved May 15, 2023, from https://digitalindia.gov.in/introduction/.

Ministry of Housing and Urban Affairs, Government of India (2015) *Smart Cities: Mission Statement & Guidelines.* Retrieved from https://smartcities.gov.in/sites/default/files/SmartCityGuidelines.pdf.

Ministry of Housing and Urban Affairs, Government of India (2020) *Data Maturity Assessment Framework.*

Misra D, **Mishra A**, **Babbar S, and Gupta V** (2017) Open government data policy and Indian ecosystems. In *Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, New Delhi, India*, New York, NY: ACM, p. 218–227. https://doi.org/10.1145/3047273.3047363.

MoHUA and MEITY (2021) *IUDX: Overview.* Retrieved from https://iudx.org.in/portfolio/iudx-overview-february-2021/.

Mulye, P (2021) The lapses in India's COVID-19 data are a result of decades of callousness towards statistics. Retrieved November 22, 2023, from https://qz.com/india/2002374/covid-19-data-india-hasnt-cared-about-statistics-for-decades.

Panjiar T and Waghre P (2022) The State of Data Protection across Indian States: A Comparison of State-Level Data Policies. Internet Freedom Foundation. Retrieved November 29, 2023, from https://internetfreedom.in/a-comparison-of-state-level-data-policies/.

Prasad A (2013) Two Decades of the Public Records Act (1993): A Critical Re-appraisal. *Proceedings of the Indian History Congress 74*, 1025–1033.

https://doi.org/10.1017/dap.2024.30 Published online by Cambridge University Press

Ranade S (2016) Traces through time: A probabilistic approach to connected archival data. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3260–3265. https://doi.org/10.1109/BigData.2016.7840983.

Rukmini S (2021) *Whole Numbers and Half Truths: What Data Can and Cannot Tell Us About Modern India*. Chennai: Westland Publications Private Limited.

Saxena S (2018) Asymmetric Open Government Data (OGD) framework in India. *Digital Policy, Regulation and Governance 20* (5), 434–448. https://doi.org/10.1108/DPRG-11-2017-0059.

Sen S and Jindal T (2022) Workshop Report: Plenary Workshop on Digital Public Records., SSRN Scholarly Paper, Rochester, NY. https://doi.org/10.2139/ssrn.4187790.

Thurston AC (2012) Trustworthy records and open data. *The Journal of Community Informatics 8*(2). https://doi.org/10.15353/joci.v8i2.3047.

Vasudevan V, **Gnanasekaran A**, **Sankar V**, **Vasudevan SA and Zou J** (2021) Disparity in the quality of COVID-19 data reporting across India. *BMC Public Health 21*(1), 1211. https://doi.org/10.1186/s12889-021-11054-7.

World Economics (2023) India. Retrieved May 16, 2023, from https://www.worldeconomics.com/DataQualityRatings/India.aspx.

World Health Organisation (2003) *Improving data quality: a guide for developing countries*. Manila: World Health Organisation.

Yeo G (2018) Records, information and data: Exploring the role of record keeping in an information culture. *Facet*. https://doi.org/10.29085/9781783302284.

Yin H (2023) An overview of urban data variety and respective value to urban computing. In Zhang H (Ed), *Handbook of Mobility Data Mining*. Elsevier, pp. 1–13. https://doi.org/10.1016/B978-0-443-18428-4.00001-3.

Zhao L, **Cao B**, **Borghi E**, **Chatterji S**, **Garcia-Saiso S**, **Rashidian A**, **Doctor HV**, **D'Agostino M**, **Karamagi HC**, **Novillo-Ortiz D**, **Landry M**, **Hosseinpoor AR**, **Noor A**, **Riley L**, **Cox A**, **Gao J**, **Litavecz S and Asma S** (2022) Data gaps towards health development goals, 47 low- and middle-income countries. *Bulletin of the World Health Organization 100*(1), 40–49. https://doi.org/10.2471/BLT.21.286254.

Cite this article: Sen S and Jindal T (2024). What does a record have to do with it? Re-situating records management within Indian public data governance and policy. *Data & Policy*, 6: e38. doi:10.1017/dap.2024.30