**Sam Asher**[1]
*Johns Hopkins SAIS,*
*Development Data Lab*

**Aditi Bhowmick**[2]
*Development Data Lab*

**Alison Campion**[3]
*Development Data Lab*

**Tobias Lunt**[4]
*Development Data Lab*

**Paul Novosad**[5]
*Dartmouth College,*
*Development Data Lab*

# Big, Open Data for Development: A Vision for India[6]

**Abstract.** Government generates terabytes of data directly and incidentally in the operation of public programs. For intrinsic and instrumental reasons, these data should be made open to the public. Intrinsically, a right to government data is implicit in the right to information. Instrumentally, open government data will improve policy, increase accountability, empower citizens, create new opportunities for private firms, and lead to development and economic growth. A series of case studies demonstrates these benefits in a range of other contexts. We next examine how government can maximize social benefit from government data. This entails opening administrative data as far upstream in the data pipeline as possible. The majority of administrative data can be minimally aggregated to protect privacy, while providing data with high geographic granularity. We assess the status quo of the government of India's data production and dissemination pipeline, and find that the greatest weakness lies in the last mile: making government data accessible to the public. This means more than posting it online; we describe a set of principles for lowering the access and use costs close to zero. Finally, we examine the use of government data to guide policy in the COVID-19 pandemic. Civil society played a key role in aggregating, disseminating and analyzing government data, providing analysis that was essential to policy response. However, key pieces of data, like testing rates and seroprevalence distribution, were unnecessarily withheld by the government, data which could have substantially improved the policy response. A more open approach to government data would have saved many lives.

**Keywords**: Open Data, Governance, India, Economic Growth, Public Goods Provision
**JEL Classification**: C8, I15, I25, O1, R11

---

[1] sasher@jhu.edu (Corresponding author)
[2] bhowmick@devdatalab.org
[3] acampion@devdatalab.org
[4] lunt@devdatalab.org
[5] paul.novosad@dartmouth.edu

[6] We are grateful to Poonam Gupta, Karthik Muralidharan, and Shekhar Shah for the invitation to write this paper. We thank Raj Chetty, Bob Cull, Ricardo Dahis, Joel Gurin, Daniel Mahler, PC Mohanan, Nachiket Mor, Malavika Raghavan, Ajay Shah, Rukmini S, eGov foundation and the Development Monitoring and Evaluation Office at NITI Aayog for their guidance and feedback. This paper grew out of many years of research using government data in India, supported by many mentors, friends, and donors too numerous to name individually.

## 1. Introduction

In 1881, the first recognizably modern census was conducted in India, covering both British India and the princely states, with the exception of Kashmir and areas controlled by other European powers. Over a two month period from December 1880 to February 1881, a standard twelve question survey was asked of every one of the 253,982,595 inhabitants of the subcontinent, the results of which were published in dozens of volumes that provided detailed descriptions and tabular data on the demographic, economic, linguistic, religious, educational, and geographic characteristics of India's massive population. More than half a million enumerators traveled to 714,707 villages and towns. They faced problems ranging from logistical challenges accessing mountainous and forested regions, to concerns that the census was preparation for a major forced displacement or recruitment for war. In some parts, enumerators were preceded by rumors that they would bring bad luck or injury, motivating people to respond behind closed doors or hide in family members' houses when enumerators were present. In others, the questions about age and marital status were entertained and said to cause "much amusement" (Plowden, 1883).

Today, more data is generated by the Government of India in a single day than in the entire Census of 1881. Every payment in MGNREGS, every health insurance claim under PMJAY, and minute details of every rural road constructed under PMGSY are recorded and stored across a sprawling network of disparate databases. Some of these data are analyzed to inform policy, and some are released publicly for use by a wide range of actors across government, civil society, and the private sector. But the vast majority sits idle in virtual warehouses, behind restricted logins and arcane websites, inaccessible even to those within government who could use them for the public good. This paper lays out a vision for setting those data free, to power India's development through better policy design, greater accountability, and more efficient markets.

India is in many ways already a leader in the broad field of data for development. The Indian Statistical and Economic Services are a deep pool of expertise in the collection and use of data. The Population Census, National Sample Survey, and Annual Survey of Industries are among the many rich data sources that have been collected for decades. In contrast to many other developing countries, most government programs have management and information systems that record detailed administrative microdata. A wide assortment of government websites make data available to the public: data.gov.in, microdata.gov.in, ecourts.gov.in, etc. India's world class tech sector has helped to build much of the public data infrastructure and finds myriad ways to generate economic growth from government data. A vibrant civil society, from academic researchers to watchdog

NGOs and a free press, uses government data to improve the understanding of India's economy, evaluate policies, and advocate for better governance.

All this notwithstanding, India's government is hampering growth and development through poor service delivery in the realm of government data. Just as India's economic growth and poverty alleviation would be enhanced if a high level of education were accessible to the entire population, so would development be accelerated by more widespread access to government data. Currently, much of the data generated by the government is either not released or is put out in a way that makes it impossible to use effectively. Resources are wasted recreating imperfect copies of databases that ministries operate but only partially release through their websites. Data collected at village and neighborhood level are often released only as state-level aggregates which have limited value for decision-making.

In this paper, we present a vision for the use and dissemination of public data that can unlock far more of its potential, based on four principles:

1. **Government data should be free.** Borrowing terminology from the open source software movement, it should first be "free" in that access should be unrestricted except to prevent harm. Data belongs to the people of India, not to their government. Second, data should also be "free" in that the cost of accessing it should be zero, not only in terms of monetary costs, but in terms of all other costs: search, cleaning, harmonizing, etc. Too many of India's "public" datasets are for all practical purposes not in fact accessible at scale, with data stuck behind web portals with attractive layouts but minimal data access.
2. **Government's primary role in the data pipeline is to generate and disseminate data.** India's vibrant civil society and private sectors have repeatedly demonstrated that they have the capability to generate original insights and add value to government-generated data. Making government data open can thus benefit society both directly and indirectly by improving government policymaking and accountability. Government needs capacity to conduct its own analyses, but this should never crowd out the delivery of data to those outside the halls of power.
3. **Data production requires clear quality standards.** Data quality is essential to its effective use. There currently exist many sensible standards for the production of government data, but implementation of these standards can be much improved. Concerns over the quality of data are legitimate but should not prevent the opening of data to the public: openness contributes to data quality through scrutiny.
4. **Data must be delivered effectively to maximize its social value and prevent abuse.** If posted data is inaccessible to users, it is not open in practical terms. For non-sensitive data, citizens should have unrestricted access to raw data via APIs and

other mechanisms that enable all members of society to use it for their myriad purposes. For sensitive data, minimal geographic aggregation can protect privacy while maintaining usefulness. Protocols for accessing personally identifiable information (PII), following well-established international guidelines, can allow researchers and others pursuing the common good to use such data without risking harm through privacy violation.

This paper explains why virtually all government data should be open, and how to go about the process of delivering that open data to all of India. It contains many proposals on how to maximize the value of data to Indian society while respecting the hard constraint of privacy protection. However, this paper is not a manual of the exact regulations that would accomplish such a goal. We do not pretend to examine every possible privacy risk or technical challenge; rather, we seek to show how a broad consensus is possible around opening much of the government's data, even as the debates rightly continue on how to respect privacy and prevent abuse.

Two broad themes run through our argument. The first is that given the investments already made in the generation and dissemination of data, achieving this vision entails high returns but relatively low marginal. The second is that public access to data is valuable because its potential uses are so varied. It is impossible for public officials (or anyone else) to anticipate the myriad uses to which the data generated by their programs may be put to use. Many billion dollar "unicorns" in Silicon Valley are built upon a foundation of free access to data on real estate, geospatial information, satellite imagery, and other standardized data layers; their Indian equivalents are far behind not because of a lack of talent or skills, but because of lack of access to data, the raw material of the information sector. The primary responsibility of government, once it has generated the data, is to deliver it to the entirety of society at zero cost, monetary or otherwise.

Government data should be open by default, and restricted only where there is a clear case in doing so for the public interest. Indeed, the Indian government already has a commitment to share its internal data with the public through the Right to Information (RTI) and various open data policies. But the RTI mechanism implies access restriction by default: only through significant work can the public obtain data that was collected from them, and even then not always. A more complete right to information would require that government data is open, usable, and available even without requiring a heroic effort by the public to unlock it. In 2021, there is no technological or other constraint on making the entirety of non-sensitive government data open and easily accessible.

This paper proceeds by describing the principles behind wider access to government data, and demonstrating some of the potential benefits through a series of case studies

describing downstream effects of open data from around the world (Section 2). We then elucidate how government should change its data production pipeline such that it is no longer the chief bottleneck in access to public data (Section 3). In Section 4, we outline the key steps required to improve the quality of government data and argue that appropriate concerns about data quality are no reason to keep data from the public. We discuss in Section 5 how data dissemination can maximize usability subject to appropriate safeguards by meeting three standards: frictionless access, appropriate delivery, and conceptual clarity. This builds on some of our own work at Development Data Lab in creating the SHRUG, a data platform designed explicitly to broaden access to otherwise inaccessible government data. In Section 6, we assess India's performance thus far in delivering government data for use by policymakers, civil society, and the private sector.

Finally in Section 7, we examine the case study of the ongoing COVID-19 crisis in India. We highlight the dynamism of policymakers, researchers, the open data community, journalists, and businesses, who worked together to use data to fight the pandemic. We also highlight the tragedy of missed opportunities caused by a lack of detailed, high quality and timely data. More open data would have enabled government to better understand the spread of the disease, to better target non-pharmaceutical interventions, and to better prioritize scarce resources by age and health conditions. Opening government data can help make India better prepared for the next major crisis.

It is clear to us that India can be the world leader in open data for development. Kapur (2020) points out that the Indian state has often excelled in creating islands of excellence but has struggled with the final mile delivery of services such as electricity, education, and health. Thanks to the internet, which transports data effortlessly across space to anyone with signal and a device, the final mile of service delivery for data is much shorter than in other domains. Relative to the huge resources that have gone into the digitization of government, small investments are required to push virtually all of government data out into the public domain, where it can be used to improve governance and propel economic growth. The larger shift required is philosophical: government must recognize that government data belongs to all of the people of India, and as such it must be made available at zero cost to anyone who wants to use it, with restrictions only to prevent harm. A now-common refrain is that data is the new oil, but instead of fueling economic growth at great cost to bank balances, health, and the environment, data has the potential to drive widespread development in India through better governance and more efficient markets. But only if it is truly set free.

## 2. Set Government Data Free

This section describes some of the many benefits that can arise from the creation of a more open ecosystem around government data in India. We argue that increasing access to government data is both intrinsically and instrumentally important.

Government data is information collected from the people of India, but it intrinsically belongs to the Indian people. They should have a right not only to access it, but to access it seamlessly and costlessly, both locally to understand how government data represents the place where they live, but also in aggregate, to understand the impacts of policy choices on a national scale.

In addition to its intrinsic ownership by the Indian people, we highlight numerous instrumental advantages to India of building a more open government data ecosystem. Increasing access to government data will allow the media to better inform the public, civil society to advocate for the marginalized and hold the government accountable, and the private sector to innovate and drive economic growth.

Throughout this section, we use the term government data to refer to data collected intentionally and incidentally through the execution of government programs. This includes survey data, like the Economic Census and the National Sample Survey, as well as incidentally-collected data, such MGNREGS projects completed and roads built under PMGSY. As far as incidentally-collected data, this paper is strictly concerned with geographically-aggregated data such that individuals cannot be identified, but can understand highly local patterns of development. There are naturally significant opportunities in creating a data ecosystem around individual data as well, but the privacy tradeoffs are more significant and demand greater attention. We deal with this question in Section 5.

### *2.1 The Right to Data*

Government data are a collection of information about the people and businesses of India, as well as the actions of the government. These data are generated and possessed, but not owned by government as a distinct entity from the people of India. As the Chief Information Officer of the United States National Oceanic and Atmospheric Administration put it: *"It's our job to get that data out there. The data doesn't belong to us, it belongs to the American people"* (Rogawski et al. , 2016). Most goods in possession of the government, like schools or canisters of cooking gas, belong to the public but must be given to some, because one person's use precludes another's. Not so with data: there is nothing to stop it from being freely shared with all members of society, as their right.

This is not a new concept in India. The Right to Information Act grants all citizens of India the right to petition the government for information that it holds:

> Right to Information Act 2005 mandates timely response to citizen requests for government information. [...] The basic object of the Right to Information Act is to empower the citizens, promote transparency and accountability in the working of the Government, contain corruption, and make our democracy work for the people in real sense.It goes without saying that an informed citizen is better equipped to keep necessary vigil on the instruments of governance and make the government more accountable to the governed. The Act is a big step towards making the citizens informed about the activities of the Government. (Government of India, 2005)

The right to information can take many forms; its implementation in India takes the form of a government commitment to respond to petitions requesting specific pieces of information. This approach makes sense in a 20th century technological paradigm, where there are significant idiosyncratic costs associated with obtaining and disseminating that information.

In the 21st century, however, much of government information is computerized and stored in the form of structured data, which can be costlessly and rapidly queried. In this context, there is no value added by requiring an intermediary to respond to requests from the public; it is technologically feasible for the public to query the government databases directly, if only they are made unrestricted. In short, the computerization of government activity and the internet for the first time enable a right to information that can be provided by default rather than intermediated through slow bureaucratic processes. In a digital world, the right to data is merely the logical conclusion of the right to information.

Technological change also means that information can now be analyzed in aggregate as data. Many insights are only possible when information is standardized and aggregated into datasets. Statistical methods can be used to identify inequalities, flag unreasonably high procurement costs, and test for the impacts of government programs. The practice of the right to information should keep up not only with technological change in information access and delivery, but also in the potential uses of information.

Currently, government data is closed by default: unless a decision is made to share a dataset with the public, it sits on government servers, often inaccessible or poorly-accessible, sometimes even to those who generate it. If a decision is made to share the data with the public, ad hoc design decisions are made about the subset of the variables to release, the level of temporal and geographic aggregation, and the type of delivery mechanism (API, click-through website, etc).

The implication of the right to data is that government data should be available to users by default, with a clear set of dissemination standards. As with other rights, this right to data has limits. Most speech is free but hate speech that promotes harm to other members of society is prohibited by the Indian Penal Code. Likewise most government data should be freely available to all, apart from that which can cause harm by undermining security or violating privacy.

We describe a set of dissemination standards in Section 5, which would ensure that the public has maximal accessibility and benefit from government data, while retaining safeguards to prevent harmful use.

### *2.2 The Myriad and Unpredictable Uses for Government Data*

Open government data has substantial instrumental value — it serves as a key input in efforts to improve governance, inform the public, create better public policy, or create new economic opportunities. Some impacts are more easily quantified, such as the market valuation of technology firms that depend on public data resources, while others are more difficult to evaluate, like the extent to which data-driven transparency initiatives improve governance.

In this subsection, we demonstrate with five case studies how freely accessible data can yield a range of benefits through both public and private sector channels. We first describe how making data available to officials in Pakistan improved the performance of public health clinics. We then summarize two studies that showed how providing data to citizens can improve democratic performance. In the third case study, we use examples from our own research to demonstrate the insights that can be gained from using government data to evaluate government programs, but only if multiple datasets are available and linkable. Finally, two case studies from the US illustrate the massive potential economic impact of high quality open government data in the hands of the private sector: the rapid growth of the real estate technology sector, and the data products of the US National Oceanic and Atmospheric Administration (NOAA).

A key message is that government data is valuable for a wide range of potential uses, which are impossible for public officials in charge of data generation and dissemination to anticipate. This implies that a policy of data restriction by default will prevent a wide range of potential uses; only a policy of open data by default lays the foundation for the innovative use of data for development.

#### Case Study 1: Data to empower public officials

The first domain in which government data can be leveraged for development is by the government itself. If officials can access clean, reliable data in a format that is accessible to

them, they can improve performance through improved monitoring of staff and spending. One example of this comes from Punjab, Pakistan, where the public health system was plagued by low attendance and performance. Callen et al. (2020) conducted a randomized controlled trial of the introduction of a new inspection tool, Monitoring the Monitors, that replaced the existing paper-based system with a smartphone-based app to collect data on rural public health clinics. Critically, this system both generated high quality data and fed it into an online dashboard that flagged in red underperforming facilities, delivering the data to inspectors in an easily accessible format.

Despite the many other challenges faced by the rural health system, this relatively minor informational intervention yielded impressive gains in performance. At baseline, monthly inspections were occurring in only 23% of clinics and doctors were present in only 24% of clinics during operating hours. Inspections more than doubled in the first six months of the intervention, although more than half of the increase was lost by the next survey another six months later. Doctors also increased their attendance in treatment clinics. Senior policymakers appeared to use the data: flagged clinics saw much larger gains in attendance than similar clinics that had slightly better baseline attendance and thus were not flagged. Taken together, the evidence suggests that providing data to policymakers in a format that focuses attention on areas of underperformance can have major effects even in very low performing agencies and in the absence of other reforms. It is worth noting that government did not develop the monitoring dashboards in house; it was only through partnership and data sharing that they were able to obtain actionable information.

**Case Study 2: Data to empower citizens and improve electoral performance**

Elections are understood to improve governance through two related channels. First, voters choose politicians whose innate characteristics will make government work better for the electorate, either because their policy preferences are more aligned with their voters or their high ability will produce good governance. Second, politicians in office may forgo opportunities for corruption if it makes voters more likely to reelect them. Both channels rest on the assumption that voters have information on politician characteristics and performance, and can thus reward good politicians with their votes and punish bad ones. But voters may find it difficult to access the information required to discipline politicians.

Two recent studies in India suggest that this is the case. In the first, Banerjee, Enevoldsen, Pande, and Walton (2020) conducted an experiment generating report cards grading politicians on how well their spending aligned with the surveyed preferences of slum dwellers in their constituencies. They gathered detailed data on the allocation of councillors' local development funds; notably, this information was only accessible to researchers through Right to Information Act filings and thus not easily accessed by voters in advance of the study. Councillors who received performance information changed their

spending patterns, but only when they were told that the report cards would be published in the newspaper, making the information available to their voters.

In a separate experiment, George, Gupta, and Neggers (2019) studied the impact of providing information to voters on the criminality of political candidates in Uttar Pradesh. Since 2004, the Supreme Court has required all candidates for elected office to submit sworn affidavits detailing their personal information, assets, and any pending criminal cases against them. There is evidence that the election of criminal politicians harms local development outcomes (Prakash et al., 2019) and that voters prefer candidates who are not criminals (Banerjee et al., 2014), yet nearly a third of candidates and elected politicians in India face open criminal charges. This information is theoretically available to voters via election commission websites, yet it is locked away in large PDF files that are difficult to find, download, and read. The Association for Democratic Reforms (ADR), a non-governmental organization dedicated to improving the electoral process in India, has converted tens of thousands of these into machine readable data, making possible a large body of research on politicians in India.[7] George, Gupta, and Neggers (2019) used ADR data to send 600,000 voters information on the criminal charges pending against politicians running in their constituencies via both phone calls and text messages. This information caused voters to redirect votes toward cleaner candidates.

Taken together, these two studies suggest that making government data more available to citizens can lead to cleaner elections and improve politician responsiveness to voter preferences while in office.[8] Both experiments used data that was collected by government but was not functionally available to citizens, either because it was locked up on government servers until the filing of an RTI, or because it was released in a format that made it difficult for voters to access. The studies also demonstrate the creative applications of data that diverse users invent when given access. Banerjee, Enevoldsen, Pande, and Walton (2020) partnered with three institutions to conduct their study: Satark Nagrik Sangathan (Society for Citizens Vigilance Initiatives, an NGO) to file the RTI requests and construct the report cards), Dainik Hindustan (Delhi's second largest newspaper) to publish the report cards, and JPAL South Asia (a research organization) to conduct audits of public goods provision and disseminate information to politicians. George, Gupta, and Neggers (2019) relied on ADR data, partnered with three telephone companies to deliver their information to voters, and then used publicly available data from the Election Commission to observe effects on voter behavior. A key benefit of making government data

[7] See, for example, Asher and Novosad (2021) on the impacts of mining on criminal politicians, Prakash et al (2019) on the economic impacts of electing criminal politicians, Vaishnav (2017) on why criminal politicians are so successful in India, and Fisman et al (2013) on the returns to political office.

[8] In a similar spirit but totally different domain, Berlinski et al (2021) find that information on student performance delivered to parents via text messages improved grades and attendance.

more open, is that it enables innovative uses of that data— uses that are difficult to anticipate in advance.

**Case Study 3: Understanding the impacts of major government programs**

The government spends many crores every year on programs whose impacts are unclear and for which there is no built-in evaluation. Yet rich open government data provide researchers with the opportunity to study the impacts of these programs and provide useful evidence for the improvement of future policy. We conducted a series of studies on the impacts of the Pradhan Mantri Gram Sadak Yojna (Prime Minister's Village Road Program), which spends approximately 15,000 crores INR per year (PRS, 2021) and has to date constructed nearly 700,000 km of rural roads to over 200,000 villages (Asher & Novosad, 2020; Adukia, Asher, & Novosad, 2020; Asher, Garg & Novosad, 2020). Our research sought to provide evidence on the impacts of new roads on economic development, educational attainment, and the local environment. We found that the main value of these roads was to connect people to urban areas: while PMGSY roads had small to no effects on business growth, living standards, agricultural practices, and deforestation, they did increase the exit of workers from agriculture via work outside of the village, as well as educational investments when returns to education in nearby urban areas were high.

This body of work relied almost entirely on government data. We merged data from many different government datasets at the village level: program administrative data from the PMGSY management and information system website, demographic data from multiple rounds of the Population Census, employment in businesses from the Economic Census, occupation and assets from the Socioeconomic and Caste Census, estimates of agricultural productivity and deforestation based on data from US government satellites, and educational attainment from the District Information System for Education. Some of these data were easy to obtain and merge, such as the multiple rounds of the Population Census. Others were easy to obtain but took years to link to the rest of the data due to data quality issues like inconsistent location codes and incomplete documentation. The PMGSY program data was technically publicly available at http://omms.nic.in/, yet it was only available as individual pages on specific roads, requiring much time and money to assemble into an analyzable dataset.

The takeaway of this case study is that government data allows for the evaluation of government programs, providing critical evidence to better allocate future resources. This research was only possible because of the Indian government's commitment to making such data available in some form. However, we spent multiple years and significant expenses to obtain, clean, and link data, work that would have been unnecessary had the government taken small steps (described in Section 5) to make these data available and

interoperable. Evidence on the effectiveness of a huge number of government programs is currently lacking because data in possession of government is not for practical purposes being released.

**Case Study 4: Improving the performance of real estate markets**

One area where open data has created tremendous economic value is in the real estate sector. Real estate is inherently costly and illiquid; buyers require detailed information about properties before making a purchase. The last decades have seen an explosion of innovative companies that combine private data from multiple listing services (essentially, aggregated lists of properties managed by multiple brokers) with municipal records of deeds and liens, tax information, and neighborhood characteristics, vastly expanding the information available to buyers and sellers of real estate.

In the US, just two of the most well known players in the property technology (proptech) space, Zillow and Redfin, have a combined market capitalization of $35B. These firms offer data-intensive services such as neighborhood comparisons, housing indices, real estate search, and property valuation (e.g. Zillow's *Zestimate* product). While these firms have since expanded into mortgage lending and real estate investment, among other activities, the core of their offerings and their original purpose centers heavily on the delivery of public data to customers in a streamlined and specific way. Indeed, Zillow originated as a company doing little more than providing customers with complete information about properties they were interested in, most of which was generated by government but not previously combined.

Without easily accessible, high-quality open data, the proptech market would not exist as it does today. Companies such as Zillow leverage a vast array of public data to fulfill their mission: surveys from the Census Bureau, parcel information in county records, economic indicators, imagery of homes, GIS data (e.g. from Census Bureau, USPS, counties, and Open Street Map), and administrative boundaries (neighborhood, ZIP code, city, county/FIPS, metro/CBSA, state). Furthermore, the proptech sector develops and open-sources additional proprietary data that contributes back to the open data ecosystem (e.g. Zillow Research datasets) and has also partnered with government data providers to advance open data standards and better align data production and consumption between the public and private sectors.

In short, an entire self-sustaining open data ecosystem has developed around the lowly public data held in municipal and county records, an ecosystem of companies and data that would not exist if these county offices used a restricted-by-default approach to property and property transaction data. The market value of property technology companies depends entirely on a system of open government data. And yet the public

benefit gained from the existence of this sector is vastly higher— because consumers capture much of the benefit created by these companies. The U.S. real estate sector transacts trillions of dollars per year; if property tech companies built on open government data can make this sector even a bit more efficient, then the economic value-added of that open data measures in the hundreds of billions of dollars.

The Indian real estate market is expected to reach a trillion dollars in size by 2030. The network of open government data on property characteristics and transaction history does not exist in India. Many firms such as housing.com and Terra Economics and Analytics Lab are already trying to make use of government data in this space, but are constrained by limited access and inconsistent standards. Shifting the government owners of administrative datasets on real estate from a default of restriction to a default of open, clean, and interoperable data could unlock hundreds of billions of dollars in economic value.

**Case Study 5: The many uses of remote sensing data**

The value of proptech companies like Zillow depends on a vast array of upstream public data from a range of sources. In this case study, we examine the downstream value derived from the data products of a single federal agency. The US National Oceanic and Atmospheric Administration (NOAA) is a scientific agency focusing on weather and atmospheric conditions, and serves as a major provider of public environmental data via the National Weather Service and a variety of satellite missions. Diverse users depend daily on NOAA products, from weather warnings to climate, ecosystem, and commercial data and modeling activities.

Many private companies have developed products and services that layer on top of NOAA capacities. The Climate Corporation, which was sold for $1.1 billion in 2013 (Tsotsis, A, 2013), provides data-intensive agricultural consulting and insurance services that depend on and extend NOAA data and forecasts. The United States' $8-10 billion financial market in annual weather derivatives is built in part upon NOAA's data (Rogawski et al. , 2016). More generally, nearly the entire US transportation network is dependent on NOAA to some degree, as weather routing for air and marine freight rely extensively on NOAA forecasts to avoid billions of dollars in losses due to weather interruptions (Government of the USA, 2011).

Private products built upon NOAA data span many sectors and applications, including weather forecasting (micro-forecasting, domain-specific modeling), agricultural and fisheries planning and operations, intelligence for commodities trading, financial risk management / insurance / re-insurance, emergency forecasting and response, property management, energy, and transport.

While private players are now emerging in the field of meteorological data production, they are unlikely to displace NOAA activities as (i) private data are often complementary to NOAA data products; (ii) NOAA provides a stable baseline and benchmark of data and modeling capacity that are reliably free to use; and (iii) NOAA is trusted to provide impartial and unbiased data and models that are insulated from political pressure and the profit motive. Historically, the private sector has added value to NOAA data and sold that value-add in the private market. Now, private partnerships are evolving from strictly value-add to co-production; for example, Google and NOAA have tied up to leverage Google's computing resources to make climate information "as accessible to the public as using Google Maps to get driving directions" (Rogawski et al., 2016).

As in case study #4, the government's initial move toward creating an open data ecosystem has created tremendous private value, embedded both in the companies that use these data and the customers who buy their products. These companies have in turn created new open data products which could have further downstream effects.

As these case studies make clear, there is an enormous range of different applications for data that government generates. In the next section, we develop a theory of the optimal role of government in the production, dissemination, and analysis of such data.

## 3. The Role of Government in Building an Open Data Ecosystem

In this section, we discuss the roles that governments can and should play in facilitating a data ecosystem that maximizes benefits to society. Society has many creators and users of data other than the government, including citizens, the media, civil society, and the private sector. We argue that governments have a comparative advantage and essential role in some aspects of the data pipeline, but should take a back seat and work primarily as facilitators of socially beneficial activities in other areas.

We begin by presenting a framework guiding the optimal use of limited government resources. We show that data on citizen activities like that routinely collected by the government has many characteristics in common with classical public goods in economics; there is thus a strong rationale for governments to play a key role as a data creator. However, there are many civil society and private sector actors capable of data analysis, and the analysis and dissemination of insights have fewer positive externalities, so there is less of a role for government to prevent other actors from playing a role in these domains.

[Figure 3.1 goes here]

Throughout this section, we consider a data production and analysis pipeline as depicted in Figure 3.1. In order, data is (i) collected, (ii) cleaned and validated, and (iii) analyzed; finally, (iv) real-world decisions can be made on the basis of that analysis[9]. Each step of the pipeline can be undertaken by the same actor; alternately, data can be transferred between actors at any stage. Government, non-government organizations, the private sector, and citizens, can all engage in any step of the pipeline, provided they can access outputs from the prior step. Data and analysis at any stage can be kept private or can be made open; making data open would allow all actors to use data outputs in downstream stages of the pipeline. We examine how actors would behave in a free market, and the optimal role for government.

### *3.1 Non-Rival and Non-Excludable Goods, and the Rationale for Government Action*

Economists define two categories of goods in whose production there may be a particular rationale for government involvement: non-rival and non-excludable goods.

#### Non-rival goods and the data pipeline

Non-rival goods are goods where one individual's consumption does not prevent another individual from consuming it. Free markets will often produce non-rival goods without intervention, but their prices are likely to be higher than socially optimal prices. For instance, software and recorded music are both non-rival, and both are produced by vibrant private industries.

However, markets in non-rival goods are characterized by the same distortions as other high fixed cost and low marginal cost markets — indeed for non-rival goods, the marginal cost of production is zero. The distortion arises because firms need to charge positive (and thus inefficient) prices to recoup their fixed costs.

Governments who produce non-rival goods will optimally charge lower prices than the private sector, expanding consumer surplus from those goods. This is the rationale behind various government policies, such as public disclosure of patent contents and patent buyouts (Kremer, 1998), both of which recognize that innovative ideas have an optimal price of zero. In a similar vein, the U.S. National Institutes of Health mandate that any research that they fund must be made open access; research findings are non-rival, and thus social value is maximized when the price of viewing those research findings is set to zero. In contrast, a private publisher of research (such as Elsevier) sets a high price for

[9] The data production and analysis pipeline discussed throughout this section was developed based on the framework laid out in Figure 0.1 in the World Development Report: Data for Better Lives (World Bank, 2021).

access to research findings, which is socially suboptimal given the non-rival nature of that research.

Every output of the data production pipeline in Figure 3.1 is non-rival. Raw data, clean data, and information about the world in the form of data analysis are all non-rival— their use by one party does not preclude others from using them. In fact, the more individuals using a given data source, the greater the value to the others using it, as errors are detected and insights discovered. However, private producers of data are likely to charge suboptimally high prices for data access to recoup their costs of production. The end result is that researchers at well-financed universities in wealthy countries often have better access to Indian data than researchers in India.

**Non-excludable goods and the data pipeline**

Non-excludable goods are goods where it is impossible to exclude non-payers from deriving benefits from those goods. For instance, clean outdoor air and national defense are classic non-excludable goods; if the goods are produced, individuals cannot be prevented from benefiting from them, even if they do not want to pay for them.

Non-excludable goods will be under-produced by a private market, because customers who can derive the benefits of the goods for free will not pay for them. Economic theory thus suggests a clear rationale for government participation in goods production: when non-excludable goods have significant social value, they should be produced by the government. Indeed, governments are the primary producers of many famous examples of non-excludable goods, like national defense, clean air and water (produced by government through regulatory actions), large fireworks celebrations, and lighthouses.

The intermediate and final outputs of the data production pipeline are best characterized as partially excludable (Ostrom & Ostrom, 1977). Each output is in principle excludable, but once a data output is in the public domain, it is difficult to prevent it from being shared further. There is nevertheless an active market in the production and sale of data and analytic outputs, especially in the domain of real-time data, where the eventual escape to the open is less important to a producing firm's bottom line.

### *3.2 The Economics of Data Production, Dissemination and Analysis*

Social and economic data is non-rival and only partially excludable. It will therefore be underprovided and overpriced by the free market, justifying government participation in the data pipeline. Government participation in the production of socio-economic data is further justified by the tremendous fixed costs of generating survey data. Sample surveys and especially censuses are extremely expensive; they involve the hiring, training, and

supervision of hundreds of thousands of enumerators. Few private firms are willing to engage in such costly activities in order to produce partially excludable goods.

Figure 3.2 provides a depiction of each sector's participation in each stage of the data collection pipeline as it currently operates. The size of the boxes indicates the size of each sector at each stage of the pipeline. Private firms, media, and civil society all engage in data collection to one degree or another. Government engages in a tremendous amount of passive data collection just through the operations of its programs. Participants in MGNREGS create an automatic stream of data on government servers; the cost of independently tracking payments and public infrastructure constructed under MGNREGS would be huge, but government obtains this data at no cost, as an incidental side effect of providing services. Across the combination of government schemes, there is an incredible multidimensional flow of information.

[Figure 3.2 goes here]

The private sector also collects a large amount of data passively and actively; we focus on government data in this paper for three reasons. First, it is more representative than private sector data, since government interacts in some form with nearly all of its citizens. Second, government data pertains directly to the operations of public programs, which are in the public interest. Third, government data is owned by the public, so the public has a clear claim to access.

As depicted in Figure 3.2, at present only a tiny subset of government data is used by any sector in society. Government largely only releases data that it has used for its own analysis, and it does not have the capacity to clean and analyze the majority of data that it collects. In contrast, the private sector, civil society, and media often collect data with the explicit purpose of guiding decisions, and thus they use a larger share of the data that they collect.

As shown by the arrows in Figure 3.2, there is significant sharing of intermediate outputs in the data pipeline, especially further downstream. Analytic outputs are widely shared between the different actors; private sector actors use government reports as information sources, and vice versa. Private actors also frequently use data created by the government (such as the NSS or ASI), and conduct their own analyses with them.

The economic framework presented above makes it clear why the majority of data created and analyzed by private firms is retained internally: to the extent that information and analytic results can be treated an excludable goods, they will not be shared, and to the extent that they are non-excludable, they will not be produced.

Government data in practice is also largely excluded from use outside government, but there is little economic rationale for this exclusion. Specifically, the vast majority of administrative data collected by government sits on servers and is never analyzed or disseminated, or is disseminated in a form that is unusable as data. There are of course valid reasons to restrict access, such as for public safety or privacy, but much of the data that the government generates is not actually sensitive. Excluding potential users vastly reduces the social value that the data can generate.

There is little reason for government to sit on a vast non-rival and non-excludable good for which the price has already been paid. As we highlighted in Section 2, tremendous social value can be unlocked by freeing that data, in myriad forms that are difficult to predict. However, it is important to release that data early in the data production pipeline. The majority of government data never even makes it to the validation and cleaning stage; treating dissemination as something that only happens after that stage ensures that the majority of government data will never be used.

An alternate data pipeline is presented in Figure 3.3. This figure represents a world where government recognizes the right to information as a right to data, and non-sensitive government data is made open access by default. In this world, civil society, media, and the private sector all benefit from the mass of passively collected administrative data. They can clean and validate government data (as the ADR has done with politician affidavit information, Section 2) and use it for their own analysis. Those analyses can then feed back into the government policy function, allowing governments to make better decisions on the basis of analysis that it is not capable of conducting in house (as in the case of the health worker attendance dashboards described in Section 2).

[Figure 3.3 goes here]

India has a world class tech sector, a large and sophisticated research community, a free press, and an active non-government sector ready to contribute to India's development through the use of data. Private firms stand ready to invest in new data-intensive business activities as soon as data become available, creating jobs and often using that data to increase the efficiency of markets. Other applications will hold government accountable or generate evidence that can lead to improved policies. Government undoubtedly wants to maintain its own analytic capabilities, but the non-rivalrous feature of data means that in-house analysis will not be hurt by others' using the data, and will likely be supported by having skilled analysts in the private sector and civil society working with the same data.

Maximizing social welfare in a context of non-rival and semi-excludable goods under control of the government dictates that those goods should be made non-excludable as early in the pipeline as possible. In practice, some government investments will need to be made to create usable data sources and APIs, and to aggregate data appropriately to preserve privacy. But government already invests in data portals for many forms of administrative data, though these are often unsuited for disseminating data in aggregate. The marginal cost of putting data in a form that is far more beneficial to the public is low. Section 5 discusses what it actually means to bring the cost to data end users as close to zero as possible. But first in Section 4, we address the issue of data quality.

## 4. Data Quality

A common mantra in computer science is "garbage in, garbage out": any data-based analysis and decisions are only as good as the underlying data themselves. Concerns about the validity of both administrative and survey data produced by governments are widespread (Jerven, 2013). In India, questions have been raised about the quality of core datasets including the Population Census (Gill, 2007), the Economic Census (Unni & Raveendran, 2006), and administrative data from PMGSY (Lehne, Shapiro, & Vanden Eynde, 2019). These concerns focus heavily on the accuracy of public data — whether reported measures correctly reflect reality on the ground, or else have intentional or unintentional errors.

Data quality encompasses much more than whether the values in the data are correct, even if data errors draw the most attention; Table 4.1 highlights one categorization of the key dimensions of data quality, based on the 2021 World Development Report and the UK Government Data Quality Framework (World Bank, 2021; Government of the UK, 2020). Quality can be described more expansively as the extent to which the data meets the objectives of its potential users (Redman, 2008).

[Table 4.1 goes here]

There is no question that the quality of much government data is suboptimal in several of these dimensions. While this paper focuses on the benefits of increased dissemination of government data, these benefits are complementary to improvements in data quality. In this section, we make three key points that pertain to data quality: (1) there is substantial low hanging fruit to improving data quality if only its value is recognized; (2) transparency and openness are likely to improve data quality in the long run, by drawing attention to errors and holding data creators accountable. Most of this section deals with

administrative data, as quality concerns in India's major sample surveys and their remedies have been widely discussed elsewhere.

### *4.1 Low Hanging Fruit for Improving the Quality of Government Data*

The key ingredient to improving data quality is demanding adherence to a quality standard. There is no need to reinvent the wheel— many excellent data quality standards exist, both in and out of India. The U.S. government's General Services Administration issues Data Quality Guidelines to "assure the quality of its information products, including their utility, objectivity, integrity, transparency and reproducibility prior to disseminating information to the public" (Government of the USA, 2019). At the core of the GSA's guidelines is the importance of (1) following best practices in data collection and processing and (2) requiring replicability of the data. The JPAL Handbook on using administrative data describes how to deploy data quality checks when aggregating, coarsening, or removing personally identifying information from the data (Cole et al., 2020). Many standards and frameworks have been written to encourage data quality as part of the Digital India umbrella program and the push towards e-Governance, which contain language echoing many of the priorities of and problems identified in this paper (Government of India, 2020).

What is lacking is implementation. Standards are fragmented across agencies or not implemented at all, and data products delivered to the public do not reflect the aspirations identified in the standards.

[Table 4.2 goes here]

Table 4.2 highlights some low-hanging fruit — ideas that are relatively easy to integrate into current data collection and dissemination practices. Many of these, like standard metadata templates or standardized location schema, simplify the process of collecting and disseminating data and contribute to interoperability across all government data — defined for this context as the ability of datasets to be linked together without loss of information.

These small efforts can yield large rewards. Consider the example of database location schema. With over 600,000 villages, 8,000 towns, and 700 districts in India, data users do not have the capacity to comprehensively correct errors in location names. When the same district is listed as "Kadapa", "Y.S.R.", and "Cuddapah" in different datasets (or in the same datasets), it creates substantial frictions and errors in analysis.

For a second example, consider how health clinics are characterized by the two of the flagship data collection operations of the Indian government: the Economic Census and the

Population Census. The Economic Census characterizes firms according to the National Industrial Classification, and thus classifies health clinics and hospitals under industry codes 86 and 87 (2008 NIC), and records their size (in terms of the number of employees) and public or private ownership. The Population Census records a range of different clinic types (e.g. primary health centre, maternal and child welfare centre, etc.) but does not record ownership. In practice the inconsistency makes it difficult to use these datasets to assess the need for additional health infrastructure; clearer documentation on definition classifications for either of these datasets would make this task much easier. Indeed, many government departments have resorted to creating redundant GIS systems recording data like these (such as NRRDA, which recently released its own inventory of public services in a new data platform).

None of the ideas in Table 4.2 are difficult to implement— but they demand a paradigm shift in the creation of administrative datasets. These datasets normally originate from software designed to track the delivery of government services internally; the data that is created is a side effect rather than a central objective. As a result, the standards in Table 4.2 may not even be on the radar of the ministries generating the data. Designers of data collection platforms need to understand that they are incidentally producing valuable information in the form of data, and these low-hanging fruit can increase that value substantially.

India's large scale sample datasets like the NSS and ASI are released with end users in mind and thus obey many more quality standards. However, they remain imperfect in terms of standardized schema and interoperability, and there is no single metadata standard across the different flagship operations. However, this is an area where standards have improved substantially when compared with old survey rounds.

### *4.2 Quality Concerns, Open Data, and Transparency*

Open access government data is obviously of greater value when the data is higher quality. What is less obvious is that opening access to government data is likely to directly improve the quality of that data as well, through two channels. First, data users will have the ability to identify errors; in the best case, this will lead those errors to be studied and corrected. In the second best case, other users will at least be aware of the errors and able to adjust their analyses for them. Second, transparency creates accountability; if the operators of administrative data-producing systems know that their output will be scrutinized, they will have greater incentive to put high effort into their work and apply some of the quality standards mentioned above.

There is admittedly some risk that data fabrication could increase as data is scrutinized more closely, for instance, to hide the fact that a government program is underperforming. However, this risk is likely to be inflated. First, it is very difficult to fabricate data in a credible fashion— it will be inconsistent with secondary measures of the same real-world values, or it will leave a trail of fabrication that can be detected by data analysts in the public. Fabrication is unlikely to succeed and in fact the incentives for fabrication are likely to fall as the probability of being detected increases.

More importantly, administrative data are already used in the implementation of programs; entries in administrative datasets determine who will get paid under MGNREGS, which firms will receive government contracts, and which households will be eligible for income support. Errors in these data have real consequences for recipients of government programs, and bringing these data to light for the errors to be detected is likely to have substantial social value.

Data quality is not only an input into an open data system, but also a critical outcome. Opening data to use and scrutiny creates a feedback loop that corrects mistakes, improving trust and quality as more users provide more inputs into the system. Data originators in government should not conceal low quality data behind firewalls, but rather open them with the admission and objective that they can be improved.

In their report on open data, the Omidyar Network argues that open data should be considered critical infrastructure (Verhulst & Young, 2016). The first step in doing so is bringing data originators on board with the value of what they are producing (even if the production is incidental), such that they recognize the value of adhering to a quality standard.

## 5. Effective Data Dissemination

This section discusses the "final mile" of data production: taking data that has been collected, and effectively delivering it to policymakers, researchers, journalists, businesses and other potential users. Establishing high quality data production systems requires massive public investments — investments that the Indian government has in large part already been making. Running large-scale surveys, tracking data from government programs, and building the infrastructure to store incoming data is costly and complex. However, once collected, too much government data is hosted in silos across a fragmented ecosystem of websites, locked behind log-ins, hidden in convoluted catalogues, buried in cumbersome dashboards, or displayed in non-machine readable formats. Accessing government data is still costly — either monetarily, or through time and technical capacity.

This need not be the case. Effectively delivering data (with appropriate privacy safeguards) to a wide range of potential users costs very little when compared with the already paid costs of collection and the returns to better dissemination.

Consider the 2013 Economic Census as an example. Between January 2013 and April 2014, 1.17 million enumerators surveyed all 58.5 million establishments in India, covering every state and union territory in the country. The massive effort allowed the government to gather crucial data on businesses and employment that greatly informs decision-makers in the public and private sectors. Commendably, the Ministry of Statistics and Program Implementation (MoSPI) has changed its policies to make Economic Census microdata available for anyone to download— earlier censuses had needed to be purchased. However, the data files are stored in the obscure .nesstar format, which requires specialized technical knowledge to open. An average user or web application cannot access the data inside without considerable time, energy, and technical skill. The location identifiers can be linked to the 2011 population census, but only indirectly and there is no clear documentation for doing so. A huge investment in data collection was made, and the data was even made available for public download— the Economic Census is among the most open data releases of the Indian government. But access to end users remains limited because of insufficient last mile investments. Most other administrative data platforms in India fare far worse on this dimension.

### *5.1 Delivery Principles*

The goal of data dissemination is straightforward: data should be as easily used as possible. In India, dissemination is a key bottleneck between data collection and use. For the widest possible range of users to be able to leverage data at the lowest cost, dissemination must meet three standards: frictionless access, appropriate delivery, and conceptual clarity.

#### Frictionless access

Frictionless access means that users can find and view data of interest with minimal time or monetary cost. Given the near-zero marginal cost of delivering electronic data, the optimal access cost for government data is zero. Restricting access through pricing, approval processes, or location requirements limits the potential applications that could be developed downstream from government data.

Data access is also constrained by search costs. This remains a thorny problem, because similar fields can be found in different datasets with different levels of geographic granularity or population subgroup disaggregation. For example, a user interested in employment data may not be aware that firm-level employment data is reported by the Economic Census while state-level employment rates can be generated from the Periodic

Labor Force Survey. In principle search engines can lower search costs, but they often fail to deliver (see Section 6.2), depending on indexed data with clear metadata standards and consistent documentation, which may not exist for many datasets.

**Appropriate delivery**

Appropriate delivery implies that the right data are served in the right format for the widest possible range of users. There is enormous variety in the format of data that users may request: government decision makers will likely require high-level summary dashboards, researchers need direct access to machine-readable data, and web applications require data to be served via API (Application Program Interface). Ideally, government data are sufficiently standardized such that they be format-agnostic, delivering data in all these formats as needed, providing access to users from a range of technical backgrounds.

APIs have become the standard mechanism for transmitting data across web applications and users. With APIs, organizations can build applications that add analytical or visualization layers on top of government data and have it update in real time, or even develop complex commercial products that depend on public data. Data delivery by API is a universal standard for technical accessibility. Additional formats are useful if they match user interest but should not detract from the primary focus of delivering the raw data in a standardized, accessible manner.

Many government data delivery systems invest considerable time and effort in displaying simplified data through single observation access, dashboards, or visualizations. Accessing a single observation may be appropriate to some users, but is highly inadequate as a primary means of accessing government data; it essentially makes datasets near-impossible to assemble without building wasteful data scraping machinery.

Visualizations are helpful for data communication, but they are not the best way to serve raw data. Visualizations require selections, filters, and explicit choices about which data to display. While a chart or graph only has 2 or 3 axes, datasets have dozens if not hundreds of variables that a user may want to explore. Serving data strictly through dashboard visualizations is effectively restricting access to the vast majority of potential uses of that data. In contrast, allowing users to directly access raw data unleashes applications beyond what any data originator could envision.

**Conceptual clarity**

Conceptual clarity implies that the contents of government data are easy to understand. If a user does not understand a dataset and all the variables it contains, then that dataset is inaccessible and cannot be used effectively. Conceptual clarity is improved when all datasets are accompanied by metadata that describes when, where, and how the data was

collected, and how exactly each variable is defined. Metadata is most effective when it is clear, concise, and presented in an expected format so a user is immediately presented with the most important information. To that end, creating standardized, structured tables with mandatory fields for every metadata file ensures that the information a user will require in order to understand the contents of a dataset is present and reported consistently across data files. A metadata standard in government would vastly increase interoperability between different government datasets.

Finally, transparency around the entire data delivery pipeline is desirable. Government survey data include detailed manuals explaining sampling strategy, questions used by enumerators, protocols for non-respondents and enumerator instruction manuals. This level of documentation should be just as important for administrative data, but is often lacking, at least in part because the data is gathered incidentally and released as an afterthought. But in many cases, these same enumerator manuals exist and are just not published. Documentation for the data preparation and aggregation process may not currently exist for administrative data, but documenting these steps is a best practice which would improve both usability and reduce data errors.

### *5.2 Safeguards*

When discussing data delivery, it is important to consider the reasons why data often are not made available by government. Many of these reasons are not justifiable, such as the fear of exposing program implementation problems to public scrutiny or a lack of vision about how the data could be used by those outside of government. But some of these reasons are valid and should be considered carefully — the most important of these is the concern for privacy. Researchers and businesses are always interested in using the most disaggregated microdata available, as it allows for the richest analysis, but this can risk exposing personally identifiable information (PII) that could be used for harm. The value of insights that can be gleaned from granular data is high, but is always secondary to government's legal and ethical responsibility to protect individuals' rights to privacy, as upheld by the Supreme Court of India in 2017.

There are several well-established techniques to handle privacy concerns. PII can be carefully scrubbed from the data, ensuring that individual records cannot be linked to any identifying information. Data can also be aggregated to higher geographic units, such as shifting from individual records to summaries of neighborhoods, towns, or villages; releasing data aggregated to town and urban neighborhood level poses little risk. If geographic aggregation is not appropriate, data can be pooled across other dimensions or otherwise transformed to mask the identities of individuals or firms — this was recently

done to great effect by Chetty et al. (2020) as they developed data resources out of PII to track the post-COVID economic recovery in the US.

In cases where there is high value to making PII available in government data, secure data centers are a standard solution, allowing permitted researchers selective access to complete data. Proposals for such use need to be solicited and vetted to ensure such data is being used purely for research purposes that serve the public interest. Governments can also elect to allow for the release of complete data including PII after a certain amount of time has elapsed. The U.S. Census Bureau releases all records 72 years after collection.

### *5.3 The SHRUG Open Data Platform*

At Development Data Lab, one of our primary goals has been to make Indian data more accessible. Two key platforms for this work have been the Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG; see Asher et al., 2021) and the DDL COVID-19 India platform, following many of the principles outlined in this paper. The SHRUG currently stitches together 30+ years of socioeconomic data on the universe of individuals and firms in India, with records from censuses, data exhaust from administrative programs, and remotely sensed measures of crop productivity, economic activity and poverty. Geocoded to the village and town, this dataset allows researchers, activists and policymakers to understand the economics, demographics, and public services of every village, town, and legislative constituency over the period 1990–2018. The SHRUG has been downloaded over 10,000 times and is used by all segments of society. The DDL COVID-19 India platform is a series of district-level aggregates put together to provide information for policy-making around responding to COVID, and is described in more detail in Section 7, following similar principles to the the SHRUG.

To maximize *frictionless access*, we freely released SHRUG data under an Open Data Commons Open Database License (ODbL), ensuring that each dataset is catalogued with both high-level and detailed descriptions, and we accompanied all data with extensive codebooks containing information on all platform contents. While limited resources have delayed our ability to develop and maintain APIs, *appropriate delivery* is facilitated by serving bulk microdata downloads in multiple formats (CSV and Stata), and via a mapping platform for easy visualization as an add-on to downloadable data but not a substitute. We target *conceptual clarity* by using a machine-readable metadata standard, ensuring that the same information is represented for each dataset. The codebook further explains every variable, the data collection process, and errors and concerns with the data.

The process of constructing the SHRUG involved ten years of work identifying, collecting, cleaning and linking data across a range of government sources. Much of this

work involved backing out location identifiers which were available to data originators but were not included with the data (for instance when datasets were based on a recent population census but included village names rather than village codes). The requirement to put in this kind of work to obtain usable data is in practice a major barrier to access. We have processed and included data from dozens of government datasets and schemes, but there are hundreds more that we have not had time or funding to integrate. There is no reason that the Indian government cannot deliver its data in a fully interoperable format, eliminating the need for this additional effort.

## 6. Assessing India's Government Data Status Quo

This section evaluates the current state of government data in India. Enormous progress has been made in the computerization of government, and impressive efforts have been made to make data available through a range of both program-specific portals and sites that aggregate data from a variety of sources. Yet despite these gains and the existence of multiple policies committing the government to opening its data, much administrative data continues to be restricted access. Further, the subset of government data available in the public domain is often delivered in a way that prevents widespread use. Nearly a decade ago, the National Data Sharing and Accessibility Policy (NDSAP) of 2012 committed the government to the principles of open access, searchability, machine readability, documentation, interoperability, and quality to all non-personal, non-sensitive data produced using public funds. However, government datasets rarely live up to all of these principles.

Fortunately, given the strong foundation of digitization of government and efforts made to encourage availability and a mindset of public ownership of data, the path to realize this paper's vision is largely restricted to comparatively low-cost issues of last mile delivery. The remainder of this section describes and applauds the Indian government's commendable efforts toward digital data production, and outlines the critical missing investments in delivery that can fully capture the potential returns of open government data.

### *6.1 Strong Fundamentals: Digitization and e-Governance Across Center and State*

The government of India has made extraordinary strides not only in moving from paper to computers, but also in the development of a modern vision of digital service delivery. This vision has evolved over the years, from isolated computerization efforts and localized digitization initiatives in the late 20th century to the expansive whole-of-government Digital India flagship program of today. One of the three core vision areas of Digital India is

"governance and services on demand", which has been supported by the National e-Governance Plan (2006) calling for the digitization of governance across multiple domains ranging from agriculture to justice, and its replacement, the e-Kranti program (2015), which strategizes and advocates for the electronic delivery of public services. The language in these foundational documents illustrates a deep recognition of the need for data that is interoperable and integrated, publicly owned, safeguarded, and easily accessed.

Digitizing administrative data is the first essential step to open government data. The government of India has made significant strides toward digitizing data collection; registration of crop pesticides, state-wise details of active taxpayers, water and air quality monitoring, voter registration, motor vehicle registration, and the tracking of cases filed across district courts are some of the many processes that have been computerized. For example, the Ministry of Finance has implemented complex and extensive digitization projects such as setting up identification (TIN) for Income Tax applications, Indian Customs Electronic Data Interchange (ICES), and Automation of Central Excise and Service Tax (ACES). More generally, the Ministry of Electronics and Information Technology (MeitY) is actively pursuing public-private partnerships under Digital India to modernize data collection and governance. The government has also announced that the upcoming Population Census will move away from the traditional paper-based survey to digital data collection, much like the United States' transition to a digital census for the first time in 2020. Further, digital management information systems (MIS) have been set up for a range of national welfare programs and schemes such as NREGA, Pradhan Mantri Fasal Bima Yojna (crop insurance), direct benefit transfers, export promotion schemes, PMJAY-Ayushman Bharat, National Urban Livelihoods Mission, electrification schemes, PMGSY, and so on.

As a result of these significant and commendable ongoing investments, the volume of administrative data generated by all levels of government in India has increased enormously in the past decade. However, much of this valuable administrative data remains locked behind dashboards and user log-ins, is not made available in an appropriate manner, and lacks the necessary documentation and metadata required for use. These remaining barriers mean the potential value of these hard-won digital resources is not being fully realized.

### *6.2 Poor Delivery: Missing Last Mile Investments that Deter Use of Open Data*

As described previously, three standards are required to be met for the greatest value to be delivered to the largest possible set of users: frictionless access, appropriate delivery, and conceptual clarity.

### Frictionless access

**Restricted portals.** The DISHA dashboard, built by the government in partnership with a civil society organization was a promising effort to harmonize disparate government data but still falls short. The platform was intended to harmonize data from 42 national government schemes (such as MGNREGA) in a fully structured interoperable dataset with maximum geographical and temporal disaggregation. The real-time scheme data hosted on the platform can be interrogated at the gram panchayat level, and is supplemented by interactive visualizations. Unfortunately, despite the fact that it does not contain sensitive data, access to the platform is limited to government officials, so the technical success of DISHA is limited, and the potential value among firms, software developers, think-tanks, researchers, and private citizens is unrealized.

**High search costs.** While most government data is locked in restricted access portals, the narrow sliver of data that is published on the public domain is difficult to use because of high search costs and extremely variable documentation standards. The aim of data.gov.in, the flagship national open government data platform for India, was to create a public intent data lake where users can freely access data to explore, test, or power detailed analysis. It is an extensive repository of structured and unstructured datasets. In the absence of high quality search and consistent documentation on the variables contained within each dataset, it is difficult to find relevant data. As one expert we interviewed put it, "one can occasionally come across a very useful dataset on the portal, but this happens mostly by chance."

### Appropriate delivery

**Few APIs.** The government of India has demonstrated an inclination toward API access for non-sensitive publicly available data, but APIs still need to be built across all publicly available datasets at narrow geographical units. Currently, this directive is not consistently met, even for data that are open. The OGD platform offers APIs only for a subset of databases hosted, and the usability and capacity of the APIs is lacking. While the ongoing India Urban Data Exchange initiative prioritizes open APIs and good documentation for every dataset, the platform has limited scope and coverage.

**Excess aggregation.** In many domains, it is impossible to find geographically disaggregated data in the public domain. Most datasets hosted on major public data platforms — such as OGD, National Data Archive, Census digital library — host data at the state and in rare instances district level. Data aggregated to states masks substantial heterogeneity and has limited potential for innovative reuse. Disaggregated data is accessible through select digital portals, such as the MGNREGA public data portal, but this should be the standard for

all government schemes. Platforms such as E-courts allow users to download unit-level (i.e. case level) data, but the data is difficult to process and devoid of any documentation.

**Lack of interoperability.** The ability to link distinct datasets and analyze them together unlocks extraordinary value, but is rarely a feature of India's current publicly available data. Multiple strategy documents suggest that many within the government understand its importance (Government of India, 2018; NSDAP 2012), and an interoperability framework for e-Governance was developed and published by the Ministry of Electronics and Information Technology in 2015 (Government of India, 2015). However, in practice, datasets on the OGD platform are very difficult to combine because of inconsistent units, definitions, and standards (geographic, industry names, and so on). Often, data products generated even within the same department are not interoperable.

**Inadequate safeguards.** In the absence of clear safeguards for privacy, data is neither open nor secure in the Indian context. In the status quo, on the one hand, non-sensitive microdata is arbitrarily held from the public domain. On the other hand, individual level sensitive data are often available on the public domain on a discretionary basis. In some cases, substantial personally identifiable information is accessible on the public domain without any checks or safeguards, such as in electoral rolls and MGNREGA beneficiary details.

### Conceptual clarity

**Insufficient metadata.** Without descriptions and instructions for a dataset at the variable and dataset levels, it is nearly impossible for a user to successfully interpret and deploy the information that has been collected. This includes both higher-level descriptions of the mechanisms and choices applied during data collection as well as specifications of variable construction, type, encodings, and other essential information. The absence of clear metadata for many administrative datasets is likely to lead to analytical errors and misinterpretation. In the Indian administrative data context, metadata is rarely available, and does not follow a consistent standard across data sources.

**Transparency.** It is almost unheard of for administrative datasets to have clear and detailed documentation describing the data collection and aggregation process, and possible sources of errors. As these data are increasingly used for decision making, it is important to demand greater documentation of all of these steps.

A comparison between the national open data platforms of India and the UK makes clear that while the OGD platform should be commended for the quantity of data that it

makes publicly available at zero charge, the quality of both the data and the delivery system can be much improved.

[Box 1 goes here]

The absence of final mile investments in delivery prevents the latent value of the enormous quantities of administrative data generated by the Indian government from being harnessed by researchers, civil society, firms, and other government departments. However, several steps in the right direction are underway to address the delivery deficiencies outlined above. For instance, The National Data Analytics Platform (NDAP) is an initiative under development at NITI Aayog seeking to standardize and centralize access to public data (Government of India, 2020). NDAP aims to harmonize data from across all sectors and ministries in the Indian government and re-host them in a standardized, well-documented manner that will allow free access to users. NDAP will be designed to allow users to access and download data alongside comprehensive and standardized metadata from multiple sectors — health, agriculture, education — in one place, linked together using a common data model. Critically, all data will have consistent geographic identifiers, allowing for the joint analysis of data coming from disparate sources.

Some states have already established effective open data platforms, such as the Government of Telangana open data portal. It is searchable, up to date, well documented, provides API access, and adheres to the principle of maximal disaggregation. The Chief Minister real-time dashboards developed for multiple states are similarly useful, but quality standards vary by state and the backend data is not available for bulk download at the village or town level. Inconsistent quality and the inability to download data in their original format prevent these data from feeding any analysis or application at scale. However, the success of digitized governance at the sub-national (state, urban local body) level through a government partnership with the E-gov Foundation has made impressive progress, which needs to be followed through with a clear commitment to opening the data.

### *6.3 Concrete Steps Towards Realizing the Potential of Open Data in India*

This subsection sketches a few concrete steps that would make major progress towards truly open government data. The core principle is that data should be *open by default* unless there is a justifiable reason for restricted access, and that restricted access can be both safe and much better than no access.

First, all data not referring to individuals should be automatically open at a zero nominal and real cost. This would entail open access APIs at the microdata level, as opposed to geographically aggregated data. Non-sensitive and non-confidential data

collected by any public authority must be hosted on an easy to navigate platform with clear standards for documentation and interoperability.

Second, all data deemed sensitive should still be released in an aggregated form at the minimal geographic level that prevents the potential for harm. For population and asset ownership data collected in the Population Census, this is likely to be the enumeration block, which would allow for systematic access to neighborhood-level data for the first time in Indian history. Other datasets, such as health insurance claims through PMJAY, may only be collected with village/town identifiers and thus should be released at that level. The protection of marginalized communities may argue in favor of releasing religious or caste data at lower resolutions, but the point here is that there is always a level of aggregation that prevents harm and delivers valuable data to potential users.

Finally, where appropriate, personally identifiable data should still be made available to researchers and policymakers at minimal cost and hassle through standardized and secure procedures following global best practices (e.g. anonymization, secure environment, remote access on a controlled server for analysis). There is no need for India to reinvent the wheel on restricting access to sensitive data. For instance, in Japan, under the recent Act on the Protection of Personal Information (APPI) adopted in 2017, an independent agency has been set up to handle two specific kinds of data: personal information (name, date of birth, email address or biometric data) and special-care information (medical history, marital status, race, religious beliefs and criminal records). The system of the UK Data Service Secure Lab elaborated in Box 1 is another example of open by default and restricted use in a controlled environment for personally identifiable and sensitive information. The US Census similarly allows researchers access to sensitive data in a secure environment that prevents the risk of leaking personally identifiable information.

In the next section, we illustrate how the existence of the mandate under a proposed coherent Right to Data could have already saved significant lives and livelihoods in the context of the pandemic in India.

## 7. Application: The COVID-19 Crisis in India

The pandemic is a compelling example of how open data could have literally saved lives and livelihoods in India. Even after two devastating waves of COVID-19, incomplete data on testing and deaths continues to hamper our understanding of the virus and to plan the policy response. In this section, we first call attention to the successes of Indian civil society in transforming fragments of discordant data released by various levels of governments into high-value efforts to inform the COVID-19 response. We then show that the absence of critical open government data has led to a series of missed opportunities to save lives.

Despite the paucity of open access health infrastructure, civil society has filled the information vacuum. Consider district-level COVID-19 infection and death counts, the most

basic information required for responding to the virus. While individual states have been releasing infection and death data through daily bulletins and reports, these daily updates have not been machine-readable and were often released as images by disparate official state and district government accounts on social media. This made them relatively difficult to access as data, until a volunteer-based organization, covid19india, set up a system to automatically aggregate these daily updates and hosted the data for all states in a single open access database — covid19india.org. Crowd-sourcing efforts based on media accounts also created the first public dataset describing COVID-19 cases disaggregated by age and gender. These crowd-sourced projects have been the single most important source of information for citizens, journalists, think tanks and researchers trying to understand the pandemic.

At Development Data Lab, we created an open access portal, posting and linking a wide range of policy-relevant variables at the district level, including demographic data extracted from censuses, public and private hospital capacity data, migration, vaccination counts, price and volume from agricultural markets, among others.[10] We supplemented this with regularly-updated infection and death data from covid19india, an easy step given their data release in the form of an API.

To our knowledge, the site was the only data source directly linking COVID-19 information to external social and economic data. Journalists used data from the platform for investigative analyses of rural-urban divergence in disease spread (Radhakrishnan & Sen, 2021) and vaccination disparities across districts (Radhakrishnan & Sen, 2021) and gender (Deshpande, 2021). Health secretaries of state governments used the platform to plan quarantine infrastructure for returning migrants. Epidemiologists used platform data to develop risk forecasting models.

The parsimonious reports released by the government were transformed into useful data by civil society. Unfortunately, a considerable volume of essential data has been withheld by the government. We highlight three examples.

First, real-time and reliable testing data continues to be the single largest gap in COVID-relevant open data in India. Testing numbers are essential for understanding whether changing case counts in a district reflect changing infections or just changing testing rates. If daily cases for a given district appear to be declining while the number of tests conducted are also being scaled down, a false sense of security is created.

[10] github.com/devdatalab/covid

The Indian Council of Medical Research (ICMR) has an operational portal where all testing centers report daily tests conducted. This portal is accessible to state governments for monitoring, but the data from this portal were never made public, even in aggregate, a decision strongly in tension with the spirit of Right to Information clause 4(2) (Government of India, 2005). The public's option to file an RTI request is of no help when data are needed in real time. Testing data could have been used to design early-warning systems, inform public health campaigns, and to allocate aid and medical resources.

Next, consider the use of serological surveys. India has been at the forefront in the gathering of seropositivity data, with dozens of studies conducted across India through the ICMR. But disaggregated data from these serosurveys were never released. Protection of privacy is not a defense, as district-level rates do not reveal anyone's private information, nor does concern that the data were noisy or subject to error. Releasing these data would have put valuable information in the hands of analysts, inform vaccine prioritization and non-pharmaceutical interventions. In contrast, in Brazil, another leader implementing national serosurveys, serodata was made much more widely available, allowing better tracking of infection rates across regions (Hallal et al., 2020; World Bank, 2021).

Finally, the case of gated GST transaction data is a considerable missed opportunity to leverage open data for an effective pandemic response in India. This was highlighted by Pronab Sen during the TN Srinivisan lecture delivered as a part of the India Policy Forum in 2020 (Sen, 2020). Sen rightly pointed out that the GST database, which is gated (like most administrative datasets), is extremely valuable to track the economic consequences of the pandemic and associated lockdowns, but has not been put to use. It is unprecedentedly granular in terms of geography and economic transactions. The real-time GST dataset is an excellent example of microdata that can be released with open access APIs aggregated at the village, town, or sub-district to protect confidentiality of parties while still adding tremendous value. In the absence of these data, researchers had no choice but to resort to imperfect proxies for economic activity such as nightlights (Beyer et al., 2020), agricultural weekly market data (Lowe et al., 2020), and online retail data (Mahajan & Tomar, 2020) to uncover the impact of COVID-19 on economic activity in India. Removal of the artificial barriers on administrative data such as aggregated GST records could have provided a substantially higher resolution understanding of the economic impact of COVID-19 that could have then informed policies focused on economic recovery.

The missed opportunities from inaccessible administrative data have hampered the response to the COVID-19 crisis in India since March 2020. However, the examples laid out in this section illustrate how safeguarded yet high resolution administrative data should be made open by default as a priority to prevent unnecessary ignorance in future waves of COVID-19 or other crises.

India in COVID-19 has been in a state of crisis; it is understandable that the government did not have capacity to make measured decisions about which data to release. However, this only demonstrates the weakness of a policy of data restriction by default— it implies that essential data is unlikely to be released during crises, when the analytical capacity of the private sector and civil society may be most needed. In contrast, a policy of data access by default, a right to government data as soon as it is created, would have put far more information about COVID-19 in the hands of the public, and would almost certainly have saved lives.

## 8. Conclusion

In this paper, we have argued that government data should be freely accessible to all members of society, both as their right, and because opening government data contributes in myriad ways to development and economic growth. Open government data improves policymaking, contributes to accountability, empowers citizens, and provides valuable inputs to firms throughout the economy. To achieve its potential, data must be high quality and the marginal time and monetary costs of accessing it must be close to zero, so that all potential users can make use of it.

To conclude, we highlight three additional investments that are complementary to opening government data and would have high returns in terms of economic growth and development.

### *8.1 Private Sector Data for the Public Good*

This paper has focused on the value of open public data, but increasingly, the richest data on the Indian economy is in the hands of the private sector. Payments platforms generate data on consumer expenditure, job sites capture information on labor supply and demand, and banks record information about default and household savings. The private sector has particularly rich real-time data, as the operations of firms generate huge amounts of information on the economy. But these data are rarely used for research or the design of public policy.

One reason for this is that there is not a clear unified framework for such contributions to be made. Facilitating the creation of linked public and private sector data would benefit researchers and policymakers alike. Economic researchers would gain access to much richer data on consumer and firm behavior. Policymakers would be able to respond to needs much more rapidly with real time data at their fingertips. Participating

firms contributing data to this effort would signal a strong commitment to being socially responsible citizens.

Privately held data describing spending, business activity, employment, education, and public health have been safely leveraged in the United States by Opportunity Insights at Harvard University to understand the economic impacts of COVID-19, and to inform policy making in the U.S. (Chetty et al., 2020). This data release is exemplary in safely balancing the tradeoff between privacy and precision. While the underlying high-resolution data contains individual information, data is shared with the public at an aggregate and anonymized level, maintaining high geographic granularity but virtually no possibility of identifying individuals.

### *8.2 Investing in Data Literacy Throughout Society*

Data on its own does not improve development outcomes; it is an input, like electricity or education. For data to contribute to India's development, it needs to be used for decision-making. We have argued that a wide range of actors across the public, civil society, and private sectors stand ready to put government data to productive use. But data literacy in India remains low and the capacity of the government in particular to use data effectively is limited.

As importantly, government should have better capacity to make use of data to improve its own functioning. Open and interoperable data is a starting point for making evidence-based policy, but the generation of evidence also requires data analytic skills and the resources (time, computing, etc.) to apply them. Building this capacity can be done in many ways. The Indian Statistical and Economic Services could be expanded to provide a pool that policymakers and administrators could draw on to help answer the questions critical to their programs. Data analytics units could be created in every ministry (center and state) to organize and release administrative data, and to use that data to provide insights and flag problems.

### *8.3 Open Data for Decentralization*

With the passage of the 73$^{rd}$ and 74$^{th}$ Amendments in 1992, India committed itself to improved governance through decentralization of powers to the municipal level (rural panchayats and urban local bodies). Social scientists often write about decentralization as a tradeoff between improved information and incentives on the one hand, and the potential for elite capture and decreased professionalism on the other (Bardhan, 2002). Open data can help to support the implementation of decentralization in India, in particular by providing citizens, gram sabhas, and elected panchayat officials with essential data on the

economic status and performance of government programs in their local regions, and information on how those compare to other regions.

Since many of these local bodies may have limited data literacy, they need something more than raw data. But the government creators of these data do not have the comparative advantage in conducted the market research to understand the information that leaders need, nor to develop appropriate delivery platforms. The government's role is to make the raw data open, at which point advocacy and private sector organizations can build the information provision layers that will make these data useful to their audience of local leaders.

## References


Government of India. 2018. *Strategy for New India @ 75.* New Delhi: NITI Aayog, Government of India.

Government of India. 2020. *National Data and Analytics Platform: Vision Document.* New Delhi: NITI Aayog. Government of India.

Adukia, A., Asher, S., & P. Novosad. 2020. "Educational Investment Responses to Economic Opportunity: Evidence from Indian Road Construction," *American Economic Journal: Applied Economics*, *12*(1), 348-76.

Asher, S., Garg, T., & P. Novosad. 2020. "The Ecological Impact of Transportation Infrastructure," *The Economic Journal*, *130*(629), 1173-1199.

Asher, S., Lunt, T., Matsuura, R., & P. Novosad. 2021. "Development Research at High Geographic Resolution," *World Bank Economic Review, 2021;, lhab003*.

Asher, S., & P. Novosad. 2020. "Rural Roads and Local Economic Development," *American Economic Review*, *110*(3), 797-823.

Banerjee, A., Enevoldsen, N. T., Pande, R., & M. Walton. 2020. "Public Information Is An Incentive for Politicians: Experimental Evidence from Delhi Elections," *Working Paper No. w26925*. National Bureau of Economic Research.

Banerjee, A., Green, D. P., McManus, J., & R. Pande. 2014. "Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India," *Political Communication*, *31*(3), 391-407.

Bardhan, P., 2002. Decentralization of Governance and Development. *Journal of Economic Perspectives*, *16*(4), 185-205.

Berlinski, S., Busso, M., Dinkelman, T., & C. Martínez. 2021. "Reducing Parent-School Information Gaps and Improving Education Outcomes: Evidence from High-Frequency Text Messages," *Working Paper No. w28581*. National Bureau of Economic Research.

Beyer, R., Jain, T., & S. Sinha. 2021. "Lights Out? COVID-19 Containment Policies And Economic Activity," *COVID-19 Containment Policies and Economic Activity (May 29, 2021)*.

Callen, M., Gulzar, S., Hasanain, A., Khan, M. Y., & A. Rezaee. 2020. "Data And Policy Decisions: Experimental Evidence From Pakistan," *Journal of Development Economics*, *146*.

Chetty, R., Friedman, J. N., Hendren, N., & M. Stepner. 2020. "The Economic Impacts Of COVID-19: Evidence From A New Public Database Built Using Private Sector Data," *Working Paper No. w27431.* National Bureau of Economic Research.

Cole, S., Dhaliwal, I., Sautmann, A., & L. Vilhuber. 2020. *Using Administrative Data for Research and Evidence-based Policy.* Cambridge: Abdul Latif Jameel Poverty Action Lab.

Deshpande, A. 2021. "As India gets vaccinated, women are falling behind.", *Indian Express*, June 21, https://indianexpress.com/article/opinion/columns/as-india-gets-vaccinated-women-are-falling-behind-we-must-mind-this-gap-covid-7368044/, accessed July 1, 2021.

Fisman, R., Schulz, F., & V. Vig. 2014. "The Private Returns To Public Office," *Journal of Political Economy*, *122*(4), 806-862.

George, S., Gupta, S. & Y. Neggers. 2018. "Coordinating Voters against Criminal Politicians: Evidence from a Mobile Experiment in India," *Working paper.* Harvard University.

Gill, M. S. 2007. "Politics Of Population Census Data In India," *Economic and Political Weekly*, 241-249.

Government of India. 2015. *Interoperability Framework For E-governance (IFEG)*. New Delhi: Department of Electronics and Information Technology, Government of India.

Government of India. 2020. *Implementation Guidelines Of Open API Policy For E-governance.* New Delhi: Ministry of Electronics and Information Technology, Government of India.

Government of India. 2005. *The Right to Information Act*. New Delhi: Ministry of Law and Justice, Government of India.

Government of the USA. 2011. *Value of a Weather Ready Nation.* Washington DC: National Weather Service and NOAA, Government of the USA.

Government of the UK. 2020. *The Government Data Quality Framework.* London: Government Data Quality Hub, Government of the UK.

Government of the USA. 2019. *Data Quality Guidelines.* Washington DC: U.S. General Services Administration, Government of the USA.

Hallal, P. C., Hartwig, F. P., Horta, B. L., Silveira, M. F., Struchiner, C. J., Vidaletti, L. P., & C. G. Victora. 2020. "Sars-cov-2 Antibody Prevalence In Brazil: Results From Two Successive Nationwide Serological Household Surveys," *The Lancet Global Health*, *8*(11), e1390-e1398.

Jerven, M. 2013. *Poor Numbers*. Ithaca: Cornell University Press.

Kapur, D. 2020. "Why Does The Indian State Both Fail And Succeed?.,"*Journal of Economic Perspectives*, *34*(1), 31-54.

Kremer, M. 1998. "Patent Buyouts: A Mechanism For Encouraging Innovation," *The Quarterly Journal of Economics*, *113*(4), 1137-1167.

Lehne, J., Shapiro, J.N. and Eynde, O.V., 2018. Building connections: Political corruption and road construction in India. *Journal of Development Economics*, *131*, 62-78.

Lowe, M., Nadhanael, G. V., & B. N. Roth. 2020. "India's Food Supply Chain During the Pandemic," *Working Paper No. 21-070.* Harvard Business School.

Mahajan, K., & S. Tomar. 2021. "COVID-19 and Supply Chain Disruption: Evidence from Food Markets in India," *American Journal of Agricultural Economics*, *103*(1), 35-52.

Ostrom, Vincent, & Elinor Ostrom. 1977. "Public Goods and Public Choices," In *Alternatives for Delivering Public Services: Toward Improved Performance*, ed. Emanuel S. Savas, 7–49. Boulder, CO: West-view Press.

Plowden, W. C. 1883. *Report on the Census of British India, taken on the 17th February 1881*. Vol 3. London: Eyre and Spottiswoode.

Prakash, N., Rockmore, M., & Y. Uppal. 2019. "Do Criminally Accused Politicians Affect Economic Outcomes? Evidence From India," *Journal of Development Economics*, *141*, 102370.

PRS. 2021. "Demand for Grants 2020-21 Analysis : Rural Development," https://prsindia.org/budgets/parliament/demand-for-grants-2020-21-analysis-rural-development, accessed July 1, 2021.

Radhakrishnan, V. & S. Sen. 2021. "Data | Kancheepuram's COVID-19 case rate 10 times that of Porbandar, but vaccination rate 14 times lower," *The Hindu,* April 22, https://www.thehindu.com/data/data-kancheepurams-covid-19-case-rate-10-times-more-than-that-of-porbandar-but-vaccination-rate-14-times-lower/article34386346.ece, accessed July 1, 2021.

Radhakrishnan, V. & S. Sen. 2021. "COVID-19 test positivity rate declined faster in urban districts than rural parts in May," *The Hindu,* June 3, https://www.thehindu.com/data/covid-19-test-positivity-rate-declined-faster-in-urban-districts-than-rural-parts-in-may/article34718845.ece, accessed July 1, 2021.

Redman, T. C. 2008. *Data Driven: Profiting From Your Most Important Business Asset*. Harvard Business Press.

Rogawski, C., Verhulst, S., & A. Young. 2016. "NOAA Open Data Portal: Creating a New Industry Through Access to Weather Data," https://odimpact.org/files/case-studies-noaa.pdf, accessed July 1, 2021.

Sen, P. 2020. "The 2nd T. N. Srinivasan Memorial Lecture: Data in Coronavirus Times," https://www.ncaer.org/IPF2020/Papers/The-2nd-TN.Srinivasan-Lecture-Pronab-Sen.pdf, accessed July 1, 2021

Tsotsis, A. 2013. "Monsanto Buys Weather Big Data Company Climate Corporation For Around $1.1 B," *TechCrunch News,* October 2, https://techcrunch.com/2013/10/02/monsanto-acquires-weather-big-data-company-climate-corporation-for-930m/, accessed July 1, 2021.

Unni, J., & G. Raveendran. 2006. “Are the Results of the Economic Census Robust?,” *Economic and Political Weekly*, 3558-3560.

Vaishnav, M. 2017. *When Crime Pays*. New Haven: Yale University Press.

Verhulst, S., & Young, A. 2016. “Open Data Impact When Demand And Supply Meet Key Findings Of The Open Data Impact Case Studies,” *Working Paper.* Available at SSRN: https://ssrn.com/abstract=3141474.

World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington DC: The World Bank.

**TABLE 4.1: Data Quality Dimensions And Explanations**

| ***Dimension*** | ***Explanation*** |
|---|---|
| Granularity | Does the data contain maximally specific geographic and temporal resolution? For example, are dates and years of collection recorded? How granularly are locations identified? |
| Accuracy | Do the values in given fields correctly describe the real-world phenomenon being measured? |
| Completeness | Does the dataset include the data that are required or expected? Does incompleteness in the sample or in missing values introduce bias? |
| Uniqueness | Is the unit of observation (e.g. a village, person, or firm) clearly defined and measured only once, or is there duplication? |
| Consistency | Do collected data agree with each other when they should, in both values and terminology? |
| Timeliness | Are the data up to date? Is there a lag between data collection and publication? |
| Validity | Are data values in the correct format? Are expenditure variables numeric and birth dates in a valid date format? |

**TABLE 4.2: Some Low-hanging Fruit For A Data Quality Standard**

| Quality Standard Element | Description |
|---|---|
| Standardized schema (e.g. location identifiers) | For example: geographic identifiers (village and town names and ID codes) should be systematically based on the most recent population census and reference those codes. Late into census periods, alternate sampling frames (e.g. LGD) should be standardized and used across all ministries. |
| Standardized variable definitions | Unify variable specifications across all producers as appropriate. For example: industrial codes, land cover classification types, binary values for yes/no variables. |
| Metadata standards | Metadata for administrative data should be as detailed as it is for survey data. A standard metadata template can serve as a guideline for both dataset-level fields (e.g. data producer and owner, sampling methodology, spatial and temporal coverage) and variable-level fields (e.g. variable type, encoding, construction notes, questionnaire and enumerator instructions). |
| Routine validation checks | Automated tests that catch common data errors. For example, negative incomes or years of education should raise red flags. if personally identifying information is being stripped from medical records to ensure anonymity, the total population count should remain the same before and after anonymization. Or if household incomes are aggregated to the village level, then district-wise and state-wise total and mean incomes should remain constant before and after aggregation. |

**BOX 1: A comparison of national open data platforms of India and the United Kingdom**

A comparison between the national open data platforms of India and the United Kingdom, widely considered one of the world leaders, illustrates the high return investments that India is not yet making. Both platforms host enormous quantities of open access administrative data at zero monetary cost. Despite hosting open data at comparable scale as the UK, the national open data portal for India falls short of delivering high returns because of delivery issues described below:

**Searchability:** The search functionality for data.gov.in requires users to know the exact name of the dataset, and tables are stored in a flat structure without an ability for the user to track multiple datasets that may be components of the same data collection exercise. On the other hand, the primary feature of the landing page of data.gov.uk are categories of data to guide the user in her exploration of useful datasets. Related datasets are nested and displayed on a single page with technical notes and supplementary information for the user to understand how the components are related.

**Documentation:** data.gov.uk has clear documentation that walks the user through steps to access data via API or publishing a database on the platform. All datasets hosted on the platform are machine readable. Each individual table is supplemented with documents and technical notes describing the data and contact information for further queries. On the other hand, only a subset of datasets hosted on data.gov.in have supporting metadata. There are no structured metadata fields required for describing what is in the data either at the dataset or variable level. Instead, there is usually a link to the ministry that produced the data. It is highly likely that a user will need to seek more information to unlock value using the data.

**Disaggregation:** The two open data platforms have substantially different approaches to disaggregation. While the integrity of most datasets are generally maintained from production to release on data.gov.uk, this is not the case on data.gov.in. The datasets hosted on UK's national open data platform are typically available at the unit of data collection (for instance, a number of datasets hosted under the health category are disaggregated at the hospital level; some datasets are available at a spatial resolution of 10km square). On the other hand, most datasets hosted on the OGD platform for India are disaggregated at the state level and rarely at the district level. The low spatial resolution of data limits usability severely by non-government and government actors.

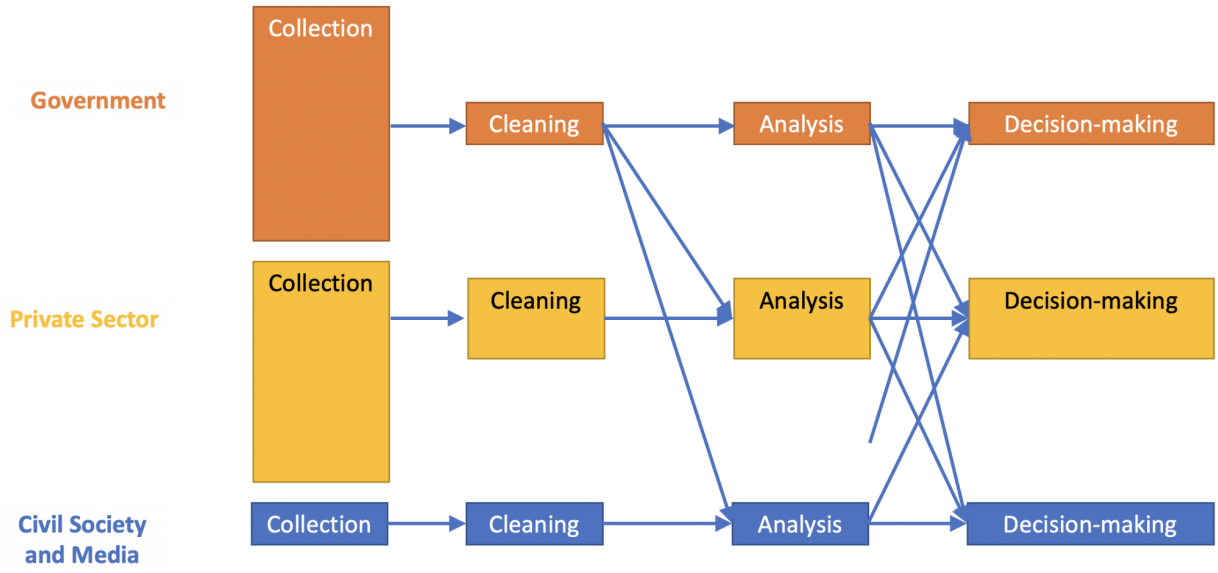
**Open by default with safeguards for privacy:** Finally, the UK approach follows the principle of open by default, and restricted access only when justifiable. The UK Data Service Secure Lab[11] was established to ensure data that is too detailed, sensitive or confidential can still be accessed for analysis but in a controlled environment. Specialized staff apply statistical control techniques to guarantee safe delivery. Data accessed through the Secure Lab cannot be downloaded. Once researchers and their projects are approved, they can analyze the data remotely from their organizational desktop, or by using a Safe Room. In the Indian context, microdata is almost never released in the public domain. The absence of protocols in place to ensure confidentiality when microdata is sensitive leads to a system where the ability to access data hinges on connections with bureaucrats in appropriate government departments. In India, the government is the de facto owner of data whereas in the UK, public intent data belongs to the people both in spirit and practice.

[11] https://www.ukdataservice.ac.uk/use-data/secure-lab.aspx

**FIGURE 3.1 : The data production pipeline**


---

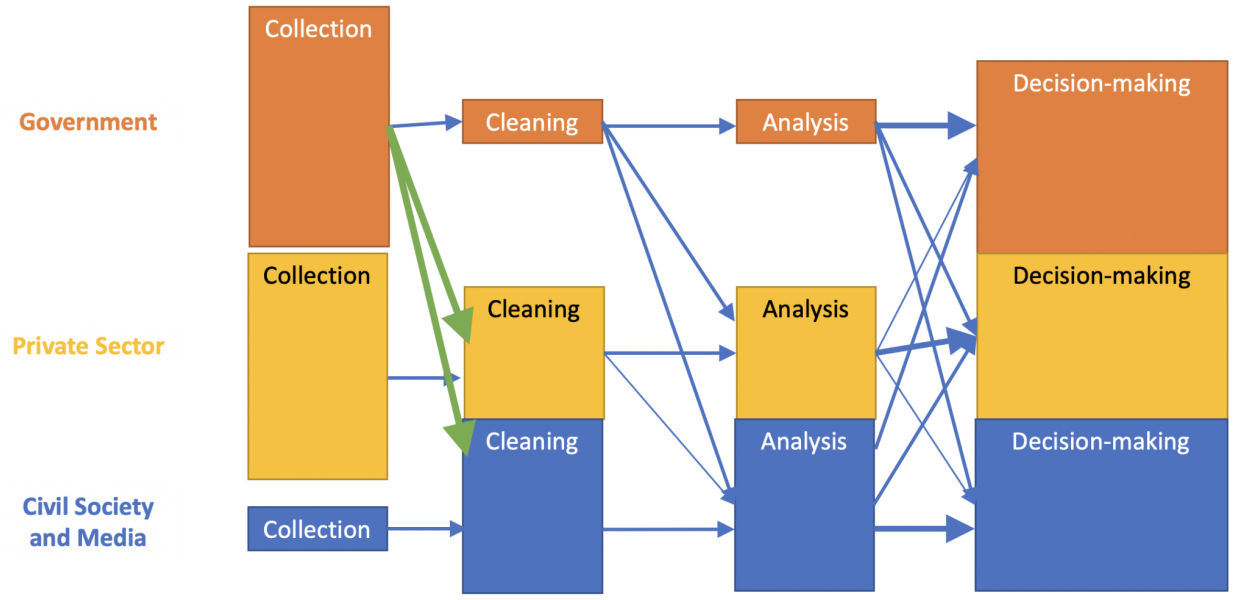
**FIGURE 3.2 : The current role of government, private sector and civil society in data production and analysis**




Currently, government passively collects far more data than it is able to analyze and use due to its limited capacity. Rich analytical findings are buried in that data, and private sector and civil society have the capability of analyzing that data, but they are not able to access it.

---

**FIGURE 3.3 : Enhanced access to data and analysis for decision-making when government data is opened early in the pipeline**




If government disseminates data early in the pipeline, taking care only to document and aggregate it to a level that preserves anonymity, the private sector and civil society can clean and analyze core components of that data, and use it to improve their decision-making. Analyses produced by civil society and the private sector can even be used by government, allowing government to make better decisions on the basis of its own data that it does not have the internal capacity to analyze. Decision-making in all sectors improves substantially when government data are made open early in the data pipeline.