Policy Opinion

Centre for The Digital Future

Enhancing public service delivery through research using high frequency de-identified government data

BY

CENTRE FOR THE DIGITAL FUTURE

NOVEMBER, 2021

This paper is based on a workshop organised by the Centre for the Digital Future (CDF) on October 26, 2021

Introduction

India is a transforming economy, where 'data' is very dynamic and live and digitization of government services is happening at an increasing pace. Utilization of data within domains like within ministries or departments within ministries has been a positive move within governance systems. It is where each ministry combines both the knowledge of what it is available and the authority to direct its usage. Yet, it is believed that at a broader level government by itself is less adept and less enterprising to tap the true potential of large amounts of data available in silos within ministries. Indeed the accuracy, the comprehensiveness, and the completeness of data are revealed only when the actual usage for specific purposes is attempted.

The government collects and creates data for various reasons – to provide public services to citizens like direct transfers into Jan Dhan accounts, or in its capacity as a regulator like the Telecom Regulatory Authority of India (TRAI), or as a commercial service provider in certain key sectors such as banking and finance or energy. Samples of such data, once de-identified and aggregated appropriately can be invaluable for research and better and more-informed public service delivery. An example of such pubic dataset is the public data portal curated on Mahatma Gandhi National Rural Employment Gurantee Act (MGNREGA), a social security programme by GoI.[1] But this is an exception rather than the norm for public data.

During the pandemic it was encouraging to see some private sector entities taking the lead in making such datasets available voluntarily[2]; more can be done to make such datasets available from the government.

The objective of the workshop was to understand how more such aggregate datasets based on government data can be prepared and made accessible to researchers.

The roundtable discussed the following questions,

- What are the various data available with the respective departments that can be aggregated and snapshots provided for research?
- What are the constraints that government departments face in preparing such datasets?
- How can researchers help solve some of these issues?
- What are the constraints that researchers face in accessing and using these data?

[1] https://nregarep2.nic.in/netnrega/dynamic2/dynamicreport_new4.aspx
[2] https://dataforgood.facebook.com/ , https://www.google.com/covid19/mobility/

- What value can research add to these data?
- How can the regulatory and legal environment be made more conducive to the creation and publication of such datasets?

Access to aggregated high-frequency data and its benefits to public policy decision making

It became apparent during the pandemic that research based on high-frequency datasets, aggregated and de-identified, can be an invaluable input to enhance public service delivery. A few examples from our research work at India Development Foundation (IDF) illustrate this.

In the first paper, daily vaccination data from the COWIN app is combined with Census 2011 data to understand how literacy is related to vaccination coverage. It is observed that not only is vaccine coverage strongly related to literacy rates in the district, but the more literate districts are also increasing vaccination coverage faster. The gender coverage in vaccination is also more equitable in more literate districts. This has implications for the design of targeted vaccination strategies to ensure universal coverage. For another paper, daily mobility data from Facebook is used to understand people's behavioural responses to COVID spread by combining it with another real time dataset – that on the number of COVID cases. This research showed how people's responses to the Virus spread can be used to better design containment strategies during pandemics.

These examples show how a more careful analysis of combining two or three datasets and looking at specific research questions may lead to generation of knowledge that is actually useful in designing responses for decision makers.

Research of this nature may not even require that researchers have access to the underlying data. In most cases, simply providing researchers access to appropriate samples of aggregated and de-identified data may be sufficient. And indeed, research can also inform the form and means of the aggregation of data into datasets.

Practical constraints to share and access public datasets

Departments within ministries can share data across departments and outside the government only with prior permissions from respective ministries.

For example, e-way bill data is one of the datasets electronically generated through Goods & Service Tax Network (GSTN). E-way bill is a permit needed for inter-state and intra-state transportation of goods worth more than INR 50,000**.** The dataset contains details of the goods, the consignor, the recipient and the transporter. With respect to this dataset when the information on supplier and

buyer are removed, then the data is de-identifiable where it simply gives what goods are moving from Place-A to Place-B. Within GSTN, there were lots of analysis done using this dataset on what kind of internal and external trade takes place, and the myths about what gets done where were broken down at pin code level. Some related analysis was also provided to the Ministry of Finance (MoF) for the Economic Survey. However, to share the de-identified version more broadly, GSTN needs to get prior approval from the MoF.

Some departments within certain ministries in the government are ready to share non-sensitive data for research and innovation purposes based on queries submitted to them. Researchers could share the query with the concerned departments within the government system that can run the query, and provide only the outputs without necessarily providing the entire dataset. This is a tried and tested potential solution to share government data outside the government. However, getting approval for each instance adds to the administrative burden.

Of course access to metadata may not always be enough for all research uses of data. There are two different approaches in data science: prediction and causation.

The above examples are where researchers are indeed looking for causal links based on a hypothesis. In such cases instead of making entire datasets accessible, it may be enough to make the metadata accessible to all users and then have a query based system as discussed above**.** Publication of metadata alone and making it searchable through non-special/ non-proprietary platforms like Google search, GitHub, etc. will be revolutionary. Thereby, potential users can indeed look at metadata to know what information exists, where it exists, and which variables are useful. This will help the user of the data to frame better queries related to their needs

The second approach is one of prediction. A research group represented in the discussion said that they took nearly two years in building AI-based solutions to address the Tuberculosis (TB) situation in the country. At the end two years, TRACE-TB (Transformative Research and Artificial Intelligence Capacity for Elimination of TB and Responding to Infectious Diseases) a national programme for bringing AI and TB into the Central Tuberculosis Division, Ministry of Health & Family Welfare (CDT, MoHFW) to support NTEP (National Tuberculosis Elimination Programme) was realised. In this case, the research organisation needed the information on all data variables collected by the CDT, MoHFW to frame the right algorithms which now helps to target nearly 20% of patients. Additionally, more than 50% of likely to be lost to follow up patients are also being followed up regularly. This way completing treatment which is an important element of actually saving lives and also avoiding drug resistant TB strains to emerge are being addressed. Here, merely giving a query to CDT, MoHFW would not have taken the digital initiative this effectively and this far.

In this case request to access TB data to build a tool for TB programme from concerned departments within specific ministries was a challenging and time-consuming process. This is because of the lack of clarity around data sharing policies and the lack of mechanisms to share this data in the country.

Indian researchers in the field of social sciences are perceived as theorists (or 'academics') as they cannot technically establish evidence from available near real time or recent data. They are mostly using five year old data.

Technical opportunities from aggregated data or anonymisation to enable public data sharing in the country need to be prioritized. Discussions within and outside the government (including researchers) will help in identifying data deficiencies and gaps that could help in realizing the incredible utility of public data after ensuring Data Privacy, Data Security, and other regulatory concerns are duly met. Then, the country could become the major researcher in the digital transformation of economies.

Need to institutionalise the access to public datasets by design and default

The data economy in India consists of number of data ecosystems and which in turn consist of smaller data ecosystems. Examples of data ecosystems include Aadhar, Unified Payment Interface **(**UPI), Goods & Service Tax Network (GSTN), National Digital Health Mission, India Digital Ecosystem of Agriculture (IDEA), and National Digital Education Architecture (NDEAR). There is a need to establish an overarching data ecosystem surrounding the smaller data ecosystems to bring them together coherently.

<u>How to promote the availability of public datasets to the public for research?</u>

- **One** is the **access by design** i.e. information that can be shared within each data ecosystem can be classified as **never shared, always shared, and shared conditionally**. The entire dataset minus the negative list (never and conditional categories) should be shared with the public. It should be a dataset by dataset approach.

- **Second** is authorization to access to aggregated data for research purposes. To embed such practice within the governance system by design. For example, in the case of Health data, the ministry (MoHFW) has notified under National Digital Health Mission as a policy of GoI on what kind of data can be shared and how the aggregate data can be shared with researchers. Similar things should be done in other sectors including agriculture, education, and transportation.

- **Third** is the **access by default** i.e. when the purpose of research is clearly specified by a user, a particular institutional body within the GoI can share the aggregated data without anybody's permission.

- **Fourth** is 'the obligation' of data users of data to the society. All public data are collected and created using taxpayers money. Users of data need to ensure that the end product of any research should be useful to the society.
-
- **Fifth** is that the national data sharing and accessibility policies should be treated as a subset of the Right to Information Act itself. So there is a proactive obligation on all the public agencies to publish data in machine readable format. For example, a lot of data are made available in PDF or printed forms. Governments including state governments should prospectively i.e. from a specified future date publish all the data in machine readable format even as they still try to do the processing for the older data.

- **Sixth** is the need for a **National Data Registry (NDR).** There are several registers to begin with. Four registers are being deployed at present in the country. Those are the definition register, the schema register, the code-list register and the catalog register. The documentation of metadata for each dataset gives information on what data is residing where, what are the standards followed, and what are the changes in standards are essential to enable integration between datasets. The NDR as a facilitator should enable integrating data in order to move ahead with machine learning and artificial intelligence. For this, Ministry of Electronics and Information Technology (MeitY) has listed eight criteria to be adhered while creating, generating and sharing the database. The criterions include completeness, consistency, interoperability, compatibility, and others.

- **Seventh** is the need for standardization alongside creation of metadata. Standardization is an evolving process i.e. while trying to use the dataset, when more and more questions are asked, where some questions can be answered and some cannot be. Then there is need for change in standards. If standardization is decided prior to using data, the entire innovative processes will come to a halt. Therefore, creating the metadata and conformance to evolving data standards should go hand in hand; otherwise retrofitting standards and metadata at a later point of time will be very difficult exercise.

- **Eight** is when users of an open data should take the responsibility to integrate the different datasets of interest. There could be multiple uses of a public dataset which includes data used for research and innovative purposes, data used for improving the public service delivery, data used for monetization by private sector, data exchange between government and private sector, and others. It is essential for users of data to be clear in their respective objectives and to best use the existing available information.

- **Ninth** is to pre-decide who – which entity within the government - will host the database servers. Should it be within an arm of the government or should it be with the central bank- the Reserve Bank of India (RBI)? Since the RBI has an established data portal to share data about the Indian economy. What kind of data should be made available through the central bank portal, can other non-finance and non-economics departments come together and make it available with the RBI are some of the areas of concern regarding data publication.

Identification of high value public datasets

Most of the questions on data access, sharing and integrating with other datasets will become relatively easier to answer from the researcher's side if we can anticipate issues of interest to the government, and then try to show how their data can help them answer those questions.

It is essential to explain to the government on what can be done with the data, and where some of the derived information will answer some of their unanswered policy questions which researchers can help in answering. This would induce the government to make more of such public data available. For example Arvind Subramanian's Economic Survey Report of January 2018[3] tried to show what can be done with the Goods and Services Taxes (GST) dataset which was barely six months old since its inception. Another example of a public policy question is, 'what is the impact of the land acquisition done in 2013, has it made farmers better off?'

There is a rich amount of data already available on public services. For example, the website called *etaal.gov.in* gives national real time aggregated data on national and state level e-governance projects in the country within last 20 seconds including GSTN, and Passports. Another reference source could be the recent publication from the National Association of Software and Service Companies (NASSCOM) on the various digital initiatives of the Government of India. Researchers need to compile 10 or 15 high value datasets, demonstrate

[3] https://mofapp.nic.in/economicsurvey/economicsurvey/pdf/032-042_Chapter_02_ENGLISH_Vol_01_2017-18.pdf

how these can be used for informing public service delivery and make the case for a broader mechanism for data accessibility.

Some key points from the discussion:

- **Data management** is not fully implemented in the country yet. There are a lot of governmental departments generating data but there is no lifecycle management of the generated data. There is missing process of standardization, and there is no emphasis on publication. Efforts should focus to bring all this aspects together.

- **Data usage** where any data process must lead to publication of the data apart from that negative list and made accessible for all possible users, including researchers.

- To have open policy for data collected using all tax-payers money. Published data can be accessed by any user for any purpose of their interest.

- To implement the **open linking policy**. This mandates a prerequisite to follow the open evolving standards which facilitates integration of different public datasets.

- Government of India should **encourage and involve** relevant research community along with critical users of public datasets to be part of designing the data ecosystem in the country.

- There is **a prior need for researchers to understand** that no one collects data for them. They as one of the users of public datasets must show ingenuity in taking the data as they appear and try to find out as much information as possible.

- To create **an engaging data ecosystem** between government and multiple users of public datasets.

Concluding remarks

"Let the perfect not be the enemy of the good"

- The first step is to identify certain high value datasets, and then illustrate the following points - is the gap in terms of the policy; is the gap in terms of the implementation of an existing policy; is the gap in terms of standardization; is the gap in terms of completeness; is the gap in terms of not enabling certain key data fields to be made available which enable correlation.

- As researchers are advocating that government need not look for the perfect because that is only going to delay the process of data sharing and access. Similarly, researchers need not necessarily look for a very comprehensive list of different datasets and identify all the existing gaps in every dataset. Instead, they can start by identifying a few of the critical datasets and identify the gaps, and make it a sequential process. Thereby, making easing the data management process as a dataset by dataset approach.

- Identifying key steps will help in stimulating an engaging data ecosystem between government and multiple users of data. It will help to push the process forward, and incremental pushes will be most helpful to the whole process.

This document is prepared based on the discussions from the workshop titled "Enhancing public service delivery through research using high frequency de-identified government data" on 26 October, 2021 hosted by Centre for Digital Future at India Development Foundation.

Moderator of the session:
Shubhashis Gangopadhyay, Research Director, India Development Foundation

Introduction by:
Mr R Chandrashekhar, Chairman Centre for Digital Future

Panellists:

- D Manjunath,
  (Professor, IIT Bombay)

- J Satyanarayana,
  (Former Secretary, Meity, GOI and Chief Adviser, C4IR, WEF)

- P. S. Acharya,
  (Head of Scientific Divisions, Ministry of Science and Technology, GoI)

- Prakash Kumar,
  (CEO, Wadhwani Institute of Technology and Policy)

- Raghu Dharmaraju,
  (President, ARTPARK)

- V. Anatha Nageswaran,
  ('Distinguished Visiting Professor' of Economics at Krea University)

- Vivek Aggarwal,
  (Additional Secretary, Ministry of Agriculture, GoI)

About CDF

The Centre for The Digital Future was launched on 30th October, 2019 with a vision to conduct actionable research on the impact of digitisation on the economy and society. The inquiries are analytical, without any pre-determined bias, multi-dimensional and evidence-based, and provide policy and regulatory insights that enable the transition to an optimal digital economy and society.

The Centre has been established and incubated as an entity by the India Development Foundation (IDF), a private non-profit research organisation set up as a Trust in 2003.