Ministry of Electronics and
Information Technology
Government of India

Report by the Committee of Experts on Non-Personal Data Governance Framework

Dated 16 - Dec-2020

1

CONTENTS

INTRODUCTION

1. Brief of the Committee

1.1. The Ministry of Electronics & Information Technology (MeitY) constituted a Committee of Experts to deliberate on a Data Governance Framework. Office Memorandum No. 24(4)/2019-CLES dated 13.09.2019 was issued to create the 8 member committee. Stated goals for the committee were

i. To study various issues relating to non-personal data.

ii. To make specific suggestions for consideration of the Central Government on regulation of non-personal data.

1.2. The list of the Committee members is provided in **Appendix 1**.

2. Methodology

2.1. Consultations with stakeholders

i. As part of the deliberations the Committee met with representatives from various sectors of business (Indian and global companies) to get their views - health, e-Commerce, Internet, not for profit organizations / think-tanks, technology service providers, etc.

ii. Several experts too presented their ideas / views and discussed with the Committee over meetings / video conference calls / mails.

2.2. In order to understand the current status of this topic across the world, the Committee did a literature review on this topic, and the relevant reports are referred across this document.

2.3. Public consultation process

i. The Committee released a draft version of the report on 12 July 2020 and obtained over 1500 feedback from the public / organizations.

ii. The Committee reviewed the feedback and has proposed this next version of the report.

- Clarified the definition of non-personal data.
- Examined the legal basis for asserting the rights of India and its communities over non-personal data.

- Expanded on the idea of High-value Datasets, a data trustee that manages them, and differentiated the roles of a data custodian and a data processor.
- Provided data sharing recommendations in the context of Public Good purpose. It recognised that data sharing for Business Purpose (i.e. data sharing between two or more for-profit private organizations.) already exists and made no recommendations towards this.
- Retained the ideas pertaining to consent for anonymized data, a Data Business and a Non-Personal Data Authority.

4

COMMITTEE DISCUSSIONS AND RECOMMENDATIONS

3. Context setting – A case for regulating data

3.1. The world is awash with data. Planet scale adoption of the Internet, smartphones, and cloud driven apps, followed by increasing use of AI-systems are the main reasons why we are generating and consuming data at a scorching pace.

3.2. Data creates economic value and wealth, apart from social and public value. Data is increasingly taking the centre-stage in core-technological businesses, all economic sectors around the world and in addressing various social and public administration issues.

3.3. Given the increasing importance and value generation capacity of the data economy, governments around the world realise the need to enable and regulate all aspects of data.

i. World over, Governments have proposed Open Data initiatives and regulations related to personal data (such as Personal Data Protection Bill 2019 (PDP Bill)[1] in India or General Data Protection Regulation (GDPR) in European Union[2]).

ii. Since no specific regulation existed over non-personal data, traditionally such data have been governed by laws pertaining to Intellectual Property Rights (including copyright and trade secrets) or by other access and control rights (including contract law).

3.4. The Committee is proposing a single national-level regulation in India to establish rights over non-personal data collected and created in India. The Committee agreed on these guiding principles in the development of the regulation.

i. Sovereignty: India has rights over data of India, its people and organisations.

ii. Benefit India: Benefits of data must accrue to India and its people.

iii. Benefits the world: Innovation, new models and algorithms for the world.

iv. Privacy: Misuse, reidentification and harms must be prevented.

v. Simplicity: The regulations should be simple, digital and unambiguous.

vi. Innovation and entrepreneurship: The data should be freely available for innovation and entrepreneurship in India.

3.5. The goals of the policy / regulation include:

i. To create an enforcing framework that

- Establishes rights of India and its communities over its non-personal data.

[1] http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf
[2] General Data Protection Regulation (GDPR) Compliance Guidelines – https://gdpr.eu/

- Addresses privacy, re-identification of anonymized personal data, and prevent misuse of and harms from data.

ii. To create an enabling framework that
   - Ensures unlocking economic benefit from non-personal data for India and its people.
   - Creates a data sharing framework.
   - Provides certainty of regulations.

iii. The Committee believes that with such a regulation, India could become the first country to put in place a simple, comprehensive framework for non-personal data.

3.6. The Committee believes that the policy / regulation will lead to the following benefits:

i. Realizing economic value from use of non-personal data. To generate economic benefits for citizens and communities in India and unlock the potential of social / public / economic value of data.

ii. The benefits accruing from processing non-personal data should accrue not only to the organizations that collect such data, but also to India and the community that typically produces the data that is being captured.

iii. Creating incentives for innovation and new products / services and startups in India.

iv. Addressing privacy concerns, including from re-identification of anonymised personal data, preventing collective harms arising from processing of non-personal data.

Refer to **Appendix 2** for more information that the Committee considered on "Data – Trends and Socio-Economic Impact".

4. Definition of Non-Personal Data

The Committee provides the following definition for non-personal data.

4.1. Definition of non-personal data

i. Non-Personal Data – When the data is not 'Personal Data' (as defined under the PDP Bill), or the data is without any Personally Identifiable Information (PII), it is considered Non-Personal Data.

ii. A general definition of Non-Personal Data according to the data's origins[3] can be:
   - Firstly, data that never related to an identified or identifiable natural person, such as data on weather conditions, data from sensors installed on industrial machines, data from public infrastructures, and so on.
   - Secondly, data which were initially personal data, but were later made anonymous. Data which are aggregated and to which certain data-transformation techniques are applied, to the extent that individual-specific events are no longer identifiable, can be qualified as anonymous data.

iii. The Committee considered various aspects of non-personal data. (Refer to **Appendix 3**). Some illustrative examples of non-personal data are provided below – of data collected by public and private entities, in public and private domains, using public and private data-collecting mechanisms. The jurisprudence on the definitions of private and public is evolving.

3 European Commission, "Guidance on the Regulation on a framework for the free flow of Non-Personal Data in the European Union", 2019, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0250&from=EN

Examples of Non-Personal Data collected by public entities

Illustrative

Data Collected by a Public Entity	Data collected where / about whom (subject)	
	Public Domain (About a community, in public spaces etc.)	**Private Domain (About an individual or a company, in private spaces etc.)**
Data Collecting Mechanisms (devices, instruments, sensors etc.)		
Public	Govt department collecting pollution data of rivers or air	Govt hospital collecting health data of a patient (Anonymised)
	Govt agency collecting data on road conditions, traffic	Public utility collecting consumption data of consumers of electricity, water in a household (Anonymised)
	Govt department collecting feedback on a public portal from citizens (Anonymised)	Agriculture department collecting data on crop acreage among private farmers in a district through analysis of remote-sensing pictures from ISRO
Private	Police department collecting video footage about a public gathering from private news channels	Government assessors visiting companies to collect compliance data
	Education department collecting student enrolment data from private schools to determine Gross Enrolment Ratio for the state (Anonymised)	Tax authorities collecting details of an individual's private bank accounts with deposits above a certain limit (Anonymised)
	The municipality collecting data on the state of the city roads by seeking photos clicked by citizens using their phones	

Examples of Non-Personal Data collected by private entities

Illustrative

Data Collected by a Private Entity	Data collected where / about whom (subject)	
	Public Domain (About a community, in public spaces etc.)	**Private Domain** (About an individual or a company, in private spaces etc.)
Data Collecting Mechanisms (devices, instruments, sensors etc.) — **Public**	A private company that is managing the water utility operations of a city collects data using Government owned sensors on water-levels in tanks	A private mobile operator determining which subscriber has stepped outside a Government defined quarantine zone defined through virtual coordinates
Private	A private autonomous car company collecting road condition data through sensors on its vehicle	A private hospital collecting health data of a patient (Anonymised)
	A private company collecting air pollution data information from their sensors installed on public lamp-posts	A private company collecting information on its employees, vendors, partners, product orders, customers (Anonymised)
	A private company's satellite capturing remote sensing information of India forest coverage	A private car-company collecting data about the condition of the car through sensors on its vehicle

5. Interface between regulation for non-personal data (NPD) and PDP Bill

The Committee has, in the course of its deliberations, identified a few recommendations which pertain to the regulation of both personal data and non-personal data. Hence the Committee suggests that appropriate changes be made in the PDP Bill to incorporate the relevant recommendations.

5.1. The Committee evaluated whether there are any overlaps between the regulations proposed for personal data and on-personal data.

i. The Personal Data Protection Bill, 2019 (PDP Bill) is intended to regulate personal data. It defines personal data as that which is capable of identifying a person. If any data that is personally identifiable is converted to a form that would render it incapable of identifying an individual, it would no longer be personal data and would therefore no longer fall within the remit of the PDP Bill.

ii. This concept is captured within the PDP Bill at Section 2(B) which states that the provisions of the Bill would not apply to any personal data that has been anonymized.

- Anonymization has been defined under the PDP Bill to be the irreversible process of transforming or converting personal data to a form in which a data principal cannot be identified, which meets the standards of irreversibility specified by the Data Protection Authority (DPA). The Committee has collated, for reference, some of the basic anonymization techniques in **Appendix 4**.

iii. Any personal data that has been subjected to this process and consequently anonymized, would become non-personal data that automatically falls outside the purview of the PDP Bill.

iv. The non-personal data regime applies to all data that is not personal data under the PDP Bill or which does not have any personally identifiable information. Since this definition expressly excludes all data that could potentially have been covered by the PDP Bill there is no overlap between the data that is sought to be regulated by the two regimes.

v. Mixed datasets that typically have inextricably linked personal and non-personal data will be governed by the PDP Bill.

5.2. The Committee evaluated what will happen in case there is re-identification from non-personal data.

i. Non-personal data would continue to be regulated by the non-personal data framework for so long as it remains non-personal data. However, if the individuals whose data constitute the anonymized dataset are re-identified in any manner, either (a) as a result of a subsequent failure of the anonymization technology, or (b) by virtue of the association of the anonymized dataset with

other anonymized datasets that together result in re-identification or (c) through any other means of conscious re-identification undertaken by the part of the data fiduciary, such data would no longer be characterised as anonymized data to which the provisions of the PDP Bill will not apply. The dataset will be deemed to have been re-identified and once again fall within the purview of the PDP Bill.

ii. The determination as to whether the PDP framework or the NPD framework applies to a specific kind of data would be determined by the identifiability of that data.
   - All personally identifiable data (including anonymized data that has subsequently been re-identified) will be governed by the PDP Bill.
   - All anonymized data that at the time of evaluation has not been re-identified will be governed by the NPD framework.

5.3. In this regard, it would be appropriate to amend the provisions of the PDP Bill to ensure that it does not regulate non-personal data. At present the provisions of Section 91(2) and Section 93(x) attempt to establish within the PDP Bill a regulatory framework within which even non-personal data could be regulated under the provisions of the PDP Bill.

i. In order to ensure that the two frameworks are mutually exclusive yet work harmoniously with each other it would be advisable to delete these sections from the PDP Bill and ensure that they are appropriately covered under the NPD framework.

ii. If that is done then the words “other than the anonymized data referred to in section 91” in Section 2(B)) could also be deleted as infructuous.

5.4. Consent for Anonymized Data

i. It is clear from industry feedback to the Committee and from its own research that large collections of anonymized data can be de-anonymized, especially when using multiple non-personal data sets. This risk is considered by this Committee to be a valid one. Hence the individual (data principal) needs more protection.

ii. Under the PDP Bill, consent is necessary for the collection and processing of personal data. Since the conditions of ‘specific’ and ‘capable of being withdrawn’, as specified in PDP Bill Chapter II, 11 (2), do not apply to non-personal data, we cannot assume that consent provided for personal data applies automatically to non-personal data.

iii. Therefore, the Committee recommends that data collectors at the time of collecting personal data should provide a notice and offer the data principal the option to opt out of data anonymization.
   - This is a disclosure requirement for data collectors.
   - It provides a notice to data principals indicating that their personal data may be anonymized and used for other purposes.

- An opt-out mechanism from data anonymization should also be provided to the data principal.
- Opt-outs are effective on a prospective basis. Also, if consent has been provided and the data has not yet been anonymised, then the revocation of consent could be given effect to.

The Committee has proposed another new concept called Data Business, which is discussed in the next section.

12

6. Defining a Data Business

The Committee proposes a new classification of business called 'Data Business' which collects and manages both personal and non-personal data.

Refer to **Appendix 5** for emerging global frameworks in this domain that the Committee considered.

6.1. Create a new classification of business called Data Business

i. A Data Business is any organization (Government or private organization) that collects, processes, stores, or otherwise manages data.

ii. A Data Business can be a data custodian or data processor (Defined in Sections 7.4, 7.5)

iii. A Data Business is a horizontal classification and not an independent industry sector. Existing businesses in various sectors that collect data will get categorized as a Data Business.
   - For example, companies in banking / finance, telecom, Internet-enabled services, transportation, consumer goods, travel, universities, private research labs, non-government organisations etc. may be classified as 'Data Businesses' based on a certain threshold of data collected / processed that will be defined by the regulatory authority.

iv. A Data Business collects and manages both personal and non-personal data.
   - The concept of a Data Business goes beyond just non-personal data. Organizations collect and process both personal and non-personal data and leverage them for various purposes including provision of services and economic purposes.

v. A Data Business will share meta-data and the underlying data under appropriate regulations.
   - Meta-data refers to data that provides information about other data.
   - The meta-data that will be shared will be the names of the data-fields collected by the Data Business.
   - For example, the meta-data that a hospital collects about a patient may include the following – {"patient name", "age", "weight", "Symptoms"}.
   - The framework for data sharing is detailed in Section 8.

6.2. A Data Business above a certain data threshold is required to register in India.

i. Threshold parameters like the following may be considered - gross revenue, number of consumers/households/devices handled, % of revenues from consumer information.[4]

ii. The thresholds suggested in the PDP Bill for Significant Data Fiduciary should be harmonised with data thresholds suggested for non-personal data.

[4] Derived from the California Digital Privacy Law

iii. Below the threshold, registration as a Data Business should be voluntary.

iv. Registration should be a one-time activity that is a disclosure-based, rather than license-based, compliance.

v. As part of the initial registration, a Data Business should provide the following information:

- A business ID, digital platform/business name(s), associated brand names, rough data traffic and cumulative data collected in terms of number of users, records and data, etc.
- The nature of data services provided – like data collection, aggregation, processing, uses, selling etc.
- Locations where data is stored and processed.
- This is similar to disclosures required by companies in the pharma or food products industry.

6.3. The information is stored in a meta-data directory managed by the Non-Personal Data Authority (NPDA – Defined in Section 7.6).

i. The meta-data about data being collected, stored and processed by the Data Business is stored digitally in meta-data directories in India. Open access is provided within India to these meta-data directories.

ii. It is suggested the NPDA define appropriate time period(s) for Data Businesses to submit meta-data.

iii. There should be a harmonisation of data-related directories and disclosures required for personal data and non-personal data, so that businesses have to supply the same information only once.

6.4. A view into the meta-data will give information about the Data Business and will enable further innovations using the underlying data.

i. Organizations registered in India will have open access to the meta-data repository. They can query this repository but not download the meta-data.

ii. By analysing the meta-data, data trustees (Defined in Section 7.10) may identify opportunities for combining data from multiple Data Businesses for community benefit.

iii. Subsequently, data trustees can make requests for relevant sub-sets of data available through High-value Datasets (Defined in Section 7.9).

- For example, automobile companies may collect data about roads through various sensors. A data trustee will know that this data is available based on the meta-data provided by automobile companies. The data trustee can request for access for this data (through an appropriate High-value Dataset) and can combine this data with public traffic data to create a data set on road safety. A startup can then analyse this data in the High-value Dataset to recommend safe and least bumpy routes.

- For example, a government funded research lab may collect and publish data on air pollution across different locations in the city. The traffic department and a real time navigation app may publish road traffic data. A data trustee for citizens in the city can request for both pollution data and traffic data to identifying safe and least polluted routes. A startup can then analyse this data in the High-value Dataset to recommend least polluted routes.

7. Establishing rights over non-personal data

The Committee considered mechanisms to establish rights over non-personal data collected and created in India. The Committee considered some key questions in this regard:

- What are the rights over the non-personal data?
- Who will exercise the rights of the non-personal data?
- What are the key roles required to facilitate the non-personal data ecosystem?

Non-personal Data Roles – Community

7.1. The rights over non-personal data include

i. Right to derive economic and other value and maximising data's benefits for the community and

ii. Right to eliminating or minimizing harms from the data to the community.

7.2. Who will exercise these rights over non-personal data?

i. In case of personal data, the rights are exercised by the data principal. However, in case of non-personal data, once the personal data is anonymised or in case the data pertains to things other than a person (such as machine, natural phenomenon, etc.), there is no data principal associated.

ii. The Committee recognises that, in the absence of a data principal for non-personal data, a community can exercise these rights over non-personal data.

- The Committee defines a community as any group of people that are bound by common interests and purposes, and involved in social and/or economic interactions. It could be a geographic community, a community by life, livelihood, economic interactions or other social interests and objectives, and/or an entirely virtual community.
- The benefits accruing from the processing of non-personal data, should accrue not only to the organizations that collect such data, but also equally to India and the community that typically produces the data that is being captured. Data being non-rivalrous, the value of data may be consumed by several organizations and communities, without degrading its value to the relevant community.
- The community (through a non-profit organization - Section 8 company, Society, Trust) should be able to raise a complaint with a regulatory authority about harms emerging from sharing non-personal data about their community.

7.3. In order to provide institutionalized mechanisms for the community to exercise these rights, the Committee recommends the creation of

i. Defined roles such as data custodian and data processor;

ii. High-value Datasets (HVDs);

iii. A new role, data trustee, to exercise the rights of the community over non-personal data collected in these HVDs;

iv. NPD Authority, to govern the rules and regulations on non-personal data.

Non-personal Data Roles – Data custodian, data processor

7.4. The Committee defines a data custodian as follows:

i. The data custodian is an entity that undertakes the collection, storage, processing, use, etc. of data. Typically, it is the data custodian that has a relationship with the consumer from whom data is collected.

ii. The data custodian may either be a Government or a Private organization.

iii. The data custodian has an obligation / responsibility to share appropriate NPD when data requests are made for defined data sharing purposes. (Refer to Section 8.5 for guidelines / safeguards on non-personal data sharing; Sections 8.1 – 8.3 for data sharing purposes.)

iv. Data custodians have a responsibility towards responsible data stewardship and a 'duty of care' to the concerned community in relation to handling non-personal data related to it.
   - The data custodian has a responsibility to ensure that no harms to persons / groups of persons occur by re-identification of non-personal data.
   - Appropriate care should be taken using technology, standards driven approach, governance framework.
   - Hence, the data custodian should employ the best anonymization standards, and follow protocols and means for safe data sharing.
   - The Committee distinguishes between ‘active misuse of NPD’ and ‘Accidental misuse of NPD’. It encourages innovations leveraging NPD and recognises that there may be accidental harms created. The data custodian must have mechanisms to swiftly remedy it.

v. The data custodian will interact with data trustees and NPD Authority (NPDA).

7.5. The Committee defines a data processor:

i. A data processor means a company that processes Non-Personal Data on behalf of a data custodian.

ii. data processors include enterprise software, Software-as-a-Service providers, cloud service providers, Global Capability Centres (GCCs), IT and ITeS companies, who process data on behalf of their client / data custodian.

iii. The data processor will not be considered a data custodian under the Non-Personal Data Governance Framework for the data belonging to the data custodian which it is processing. It will not be expected to share such Non-Personal Data.

iv. The data processor will be considered a data custodian for data that it collects, stores, processes, uses, etc. as part of its business operation (and not the data of the data custodian).

Non-personal Data Roles – High-value Datasets (HVD) and data trustee

7.6. An HVD is a dataset that is beneficial to the community at large and shared as a public good, subject to certain guidelines pertaining to the management of an HVD and data sharing as defined in Sections 7.7, 7.8 and 8.7.

i. Useful for policy making and improving public service and citizen engagement
ii. Helps create new and high-quality jobs
iii. Helps create new businesses – startups and SMEs
iv. Helps in research and education
v. Helps in creating new innovations, newer value-added services / applications
vi. Helps in achieving a wide range of social and economic objectives including
vii. Poverty alleviation
viii. Financial inclusion
ix. Agriculture development
x. Skill-development
xi. Healthcare
xii. Urban planning
xiii. Environmental planning
xiv. Energy
xv. Diversity and Inclusion
xvi. And others

7.7. The Committee has defined a data trustee as an organization, either a Government organization or a non-profit Private organization (Section 8 company / Society / Trust), that is responsible for the creation, maintenance, data-sharing of High-value Datasets in India.

i. A data trustee can be organically created by the coming together of some community members and they can decide to host an HVD. A process for creating HVDs and criteria for becoming a data trustee is provided in Section 7.8.
ii. Data trustees have a responsibility towards responsible data stewardship and a 'duty of care' to the concerned community in relation to handling non-personal data related to it.
    - A data trustee has obligations to ensure that HVDs are used only in the interests of the community.
    - A data trustee has a responsibility to ensure that no harms to persons / groups of persons occur by their re-identification of non-personal data.

- A data trustee is obligated to establish grievance redressal mechanisms so that the community can raise grievances

iii. A data trustee is a Data Business.

iv. The following are certain guidelines pertaining to the data trustee, its management of an HVD and data sharing.
   - For every HVD, there will be one data trustee.
   - A given data trustee may be responsible for more than one HVDs.
   - A data trustee will maintain the HVD in a data infrastructure, which corresponds to technical-material elements like actual databases, APIs, organisational systems, etc.
   - A data trustee will request Data Custodians for required data.
   - Data Requesters can request the data trustee for access to datasets contained in the High-value Dataset maintained by that data trustee. These requests for access may come from Public or Private organizations registered in India. These requests cannot come from individual persons.
   - The data trustee of the HVD may levy a nominal charge (towards data infrastructure, data processing) to the Data Requesters.

7.8. The Committee proposes a process for creation of HVDs.

i. An HVD is a dataset that is a public-good and benefits the community at large.

ii. In consultation with the NPDA, a Government or non-profit private organization (like an industry body, community body) in its role as a data trustee may request for a creation of an HVD.

iii. The NPDA will set detailed guidelines to determine appropriateness of the chosen HVD and data trustee (in terms of dataset, objectives, size, actors involved etc.)
   - Objective and impact of the HVD – is it in public interest? Will it be a public good?
   - Is a valid data trustee proposing the HVD – is it a valid Government or a non-profit private organization?
   - Has the data trustee, proposing the HVD, secured an expression of interest from a minimum number of community entities to be part of the HVD initiative?
   - Does the data trustee have the capacity and capability (in terms of people, technical capability of the people etc.) to handle the HVD?
   - A public consultation / organic process is managed to map the contours of the HVD.

iv. Jurisprudence over this process will evolve over time.

7.9. Some examples of data trustees include:

i. The Ministry of Health and Family Welfare, Government of India can be the data trustee for HVD with data on diabetes among Indian citizens.

ii. The State Government of Manipur can be a data trustee for HVD on data on Meitei language.

iii. A non-profit citizens group registered in Whitefield locality in Bangalore can be the data trustee for HVD on data on solid waste management in Whitefield.

iv. A public university in Hyderabad that is collecting data on the state of roads in Hyderabad as part of research project can be a data trustee of HVD on road conditions in Telangana.

v. The Directorate of Urban Land Transport may become a data trustee of HVD of traffic data with data inputs from multiple ride-sharing platforms, city police department.

vi. An industry body (like NASSCOM) may become a data trustee for a Skills Registry HVD of all IT professionals in India which would benefit IT and ITeS / BPO industry / startups.

vii. A Farmer Producer Organization in a particular district in India may become a data trustee of an HVD for agriculture related data (soil data, rainfall data etc.) in that district that may benefit the farmers there.

Non-personal Data Roles – Non-Personal Data Authority (NPDA)

7.10. The Committee proposes creation of separate Non-Personal Data Authority (NPDA).

i. NPDA must be created with industry participation and should be harmonised with other bodies like the Personal Data Protection Authority (DPA), CCI, etc.

ii. NPDA's enabling function
   - Ensure unlocking economic benefit from non-personal data for India and its people / communities.
   - Create a data sharing framework.
   - Manages the meta-data directory of Data Businesses in India.

iii. NPDA's enforcing function
   - Establish rights over Indian non-personal data in a digital world.
   - Address privacy, re-identification of anonymized personal data, prevent misuse of data.
   - In case of data sharing for High-value Datasets, the NPDA will adjudicate only when a data custodian refuses to share data with the data trustee.

7.11. The Committee considered several questions with respect to the creation of the NPDA.

i. Can the sharing of non-personal data be self-regulated by business and other stakeholders?

ii. Can various sectoral regulators address issues that are related to non-personal data?

iii. Can the Data Protection Authority (DPA), proposed in the PDP Bill, address non-personal data too in coordination with the Competition Commission of India (CCI) and other sectoral regulators?

iv. Can a department within the Government coordinate the roles of various regulators such as the DPA, the CCI, and other sector regulators to regulate non-personal data?

7.12. Ultimately, the Committee felt that the best option is to create a separate NPDA.

i. This is a new and emerging area of regulation. The regulatory authority will need specialized knowledge (of data governance, technology, latest research and innovation in the space of non-personal data, etc.) and will have to keep pace with the rapidly evolving technological landscape.

ii. The nature of tasks and focus required of this authority are quite different from those of existing ones.
   - Unlike the DPA which is focussed on prevention of personal harm, this authority will focus on unlocking value in non-personal data for India.
   - Unlike CCI, this authority will be a proactive actor providing early and continued support for Indian digital industry and startups, and ensuring that necessary data is available for the community. This authority must evaluate the nature of data sharing requests to avoid unfair or spurious requests which don't serve social, public or economic purposes.
   - Unlike sector regulators, this authority will have the expertise and a cross-cutting view and role for ensuring data sharing (which requirement often crosses sectoral boundaries), and sectoral regulators can build additional data regulations etc. if required, over those developed by this authority in a horizontal fashion.
   - This authority should work in consultation with the DPA, CCI and other sector regulators, as appropriate, so that issues around data sharing, competition, re-identification or collective privacy are harmoniously dealt with.

iii. The roles of the proposed Personal Data Authority (from PDP Bill 2019), the Competition Commission of India (under the Competition Act, 2002), and the proposed Non-Personal Data Authority, should be harmonised.

7.13. The central government has the authority to pass laws on non-personal data, similar to how it has created one regulation for all of India with the IT Act, PDP Bill.

i. The regulations proposed for non-personal data can be enforced effectively and at a national scale only if they are incorporated as part of a new national law. The Committee strongly recommends that the proposed Non-Personal Data

Governance Framework becomes the basis of a new legislation for regulating non-personal data.

8. Data Sharing

Data sharing refers to the provision of controlled access to non-personal data for defined purposes and with appropriate safeguards in place.

Data Sharing Purpose

Why, and under what conditions, should data be requested and shared? The Committee has identified three purposes for Non-Personal Data Sharing.

8.1. Sovereign Purpose

i. Data may be requested for purposes of national security, legal purposes, etc. Some non-exhaustive examples of these are:
   - Data requested for mapping security vulnerabilities and challenges, including people's security, physical infrastructure security and cyber security.
   - Data required for crime mapping, devising anticipation and preventive measures, and for investigations and law enforcement.
   - Data required for pandemic mapping, prediction and prevention, and also subsequent interventions.

ii. Already regulations exist in India which address sharing of data for Sovereign purpose. This framework only reiterates the need for such data sharing and does not propose anything new or additional.

iii. Data requests for sovereign purpose will be made only by public / Government entity.

iv. Data requests may be made to public or private data custodians.

v. The granularity of data requested is determined on a case to case basis by a Government / public entity under the sovereign purpose. Data requested will be typically a combination of personal and non-personal data.

vi. NPDA will not adjudicate validity of data requests in case of data requests under sovereign purpose.

8.2. Public Good Purpose

i. Under this purpose, a HVD is created which is a dataset that is a public-good and benefits the society at large. Sections 7.7 to 7.10 provide details on HVDs and how a data trustee manages them.

ii. Data may be requested for community uses / benefits or public goods, research and innovation, for policy development, better delivery of public-services, etc.
   - Certain data held by private sector when combined with public-sector data or otherwise may be useful for policy making, improving public service, devising public programs, infrastructures, etc. and, in general, supporting a wide range of societal objectives including science, healthcare, urban planning etc.

iii. India should identify HVD domains like health, geospatial and/or transportation data.
   - For example, the India Urban Data Exchange (IUDX) platform is being developed by the Indian Institute of Science and the Smart City Mission in Ministry of Housing and Urban Affairs.[5]
   - Progressively identify other priority domains for harnessing the economic and community benefits from leveraging non-personal data. For example, agriculture, education, skills development, MSMEs support, logistics etc.

iv. Utilize HVD for research purposes
   - For example, non-personal data can also be used by Indian researchers and government agencies for creating public goods and services like Indian language translation etc. which can then be leveraged by both public and private organisations.

v. Request for access to the High-value Dataset can be from any organization registered in India. But not from an individual person.

8.3. Business Purpose

i. Under this purpose, the non-personal data is shared between two or more for-profit private entities.

ii. Data sharing for a Business Purpose already exists.

iii. Given that such data sharing already exists between two private entities, the Committee does not make any recommendations on this.

8.4. The scope of data sharing with respect to 1) NPD Requester – NPD Purpose and 2) NPD Requester – NPD Collector is shown pictorially below.

8.5. Guidelines / Safeguards on non-personal data sharing for HVDs

i. Non-personal data sharing is proposed only for specific purposes.

ii. Such data sharing should benefit greater public good.

iii. The data requests should not be generic or broad-based and instead be specific and targeted at the purpose defined.

iv. The Committee recommends that data trustees share HVD for public good purpose with public and private organizations.

v. The table on granularity of HVDs (defined in Section 8.9, 8.10) bring in what the private and public data collectors share.

vi. For sharing data for HVDs, certain reasonable charges may be paid to the data custodian towards processing of data (anonymization, aggregation, data sharing

[5] IUDX - Forum for the India Urban Data Exchange – https://forum.iudx.org.in/

etc. – but not towards data collection, which may be considered as part of their business operations.)

vii. Where the data is not part of an HVD in India, requests for such data directly to public or private data custodians are not in scope of the Committee's recommendations.

viii. Outside of a Public Good purpose, private entity to private entity mandatory data sharing is not considered in scope of the Committee's recommendations.

8.6. Such data sharing should not harm both the data collector and data source. Kinds of non-personal data that will not be included for sharing:

i. When data sharing would involve access to private companies' trade secrets or other proprietary information regarding their employees / internal processes and productivity data.

ii. When data sharing is likely to violate privacy of individuals, groups, or communities.

25

NPD Requester – NPD Purpose matrix

Data Sharing Purpose	Data Requester: Public Entity	Data Requester: Private Entity	Data Requester: Person
Sovereign Purpose	Already exists Not part of Committee's recommendations	No	No
Public Good Purpose	Yes	Yes	No
Business Purpose	No	Already exists Not part of Committee's recommendations	No

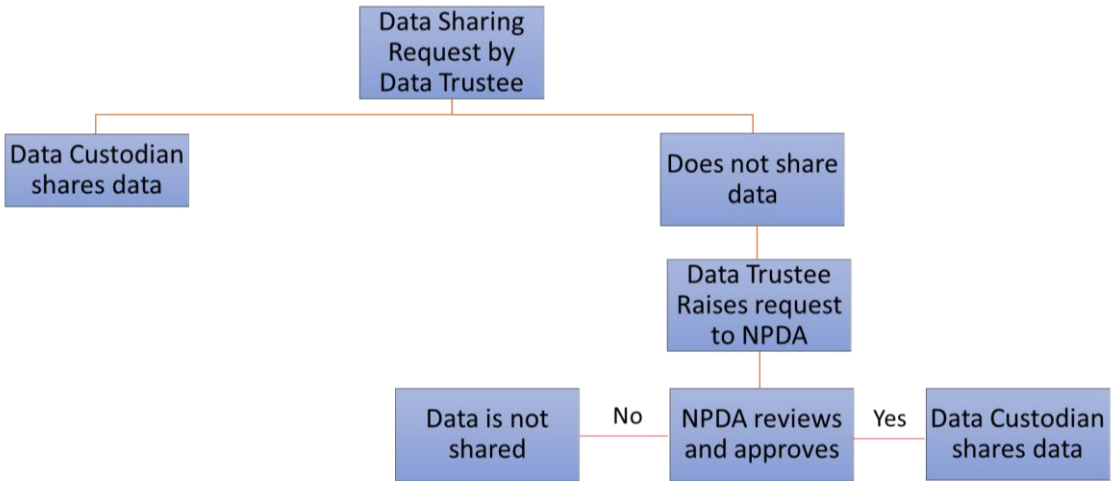
NPD Requester – NPD Collector Matrix

Data Collector \ Data Requester	Public Entity	Private Entity	Person
Data Trustee of HVDs	Yes	Yes	No
Public Entity	Already exists Not part of Committee's recommendations	Already exists Not part of Committee's recommendations	Already exists Not part of Committee's recommendations
Private Entity	Already exists Not part of Committee's recommendations	Already exists Not part of Committee's recommendations	Already exists Not part of Committee's recommendations

Creating and Sharing HVD – Granularity and Process

The Committee suggests the granularity of non-personal data that is to be collected for creating a HVD, and lays out a process for sharing HVDs.

8.7. Creating HVDs



8.8. Data trustees should request for data from all major data custodians in the corresponding data-category to create HVDs. There should be a non-discriminatory access to data from the ecosystem.

8.9. Definition of granularity of the data shared for High-value Datasets

i. Raw / factual / transactional data level (this is the base level of data provided or observed – for example a purchase order by a person (anonymised), a taxi trip detail of a traveller, census information of a citizen (anonymised), weather data for a given day, etc.)

ii. Aggregate data level (this is an aggregated view of the data (like mean, median, mode of the data sets), across several data points without revealing the base level data – for example, aggregated view of purchase orders by all customers (anonymised) in a day in a city, aggregated details of daily taxi trips of all travellers (anonymised) in a locality etc.)[6]

iii. Inferred data level (this is an inferred or derived view of data, where insights are developed by combining different data points typically involving trade secrets, algorithms, computational techniques, advanced analytics etc.)

8.10. The Committee recommends the granularity of the data for creating High-value Datasets, as shown in the table below.

[6] Derived from https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en

Granularity of High-value data	Collected by Private Entity	Collected by Public Entity
Raw / factual / transactional data level	Complete datasets- No Specific Subset of data fields- Yes	Complete datasets- No Specific Subset of data fields- Yes
Aggregate data level	Yes	Yes
Inferred data level	No	Yes (except in cases of national security)

Let us consider a few examples.

8.11. An HVD for Transportation data is set-up by a Non-Profit agency playing the role of a data trustee for this dataset.

i. It may work with the Directorate of Urban Land Transport, ride-hailing / sharing platforms, city police department, to collate traffic data in order to develop a city traffic solution.

ii. The data request to all major and significant ride-hailing platforms will be specific and for the purpose
   - It will be for a subset of the data fields collected by these data custodians.
   - The data sought may be factual / raw data related to start-time of a ride, end-time of ride and location, in order to under city traffic situation.
   - The frequency of data transfer will be determined by the purpose and specific application / use-case.

8.12. An HVD for Cancer data is set-up by the Ministry of Health and Family Welfare with the department playing the role of a data trustee for this dataset.

i. It will work with government cancer hospitals, government cancer research institutions and public-funded universities to build this dataset. The data trustee believes that radiation data from cell towers would be a useful input into this dataset.

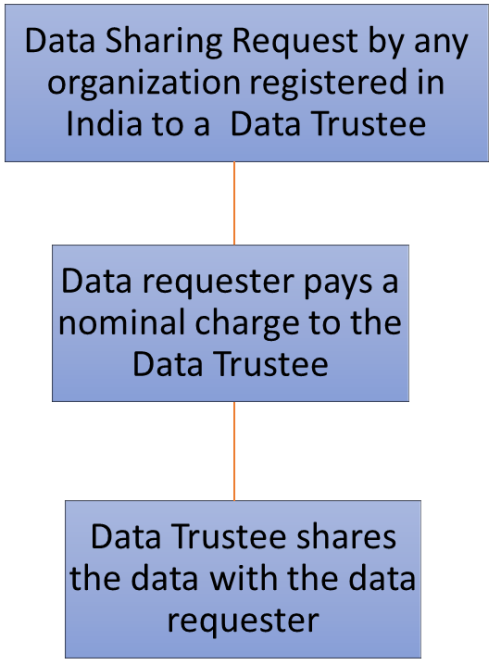
ii. The data request to all major cell tower companies will be specific and for the purpose
   - It will be for a subset of the data fields collected by these data custodians.
   - The data sought may be factual / raw data related to strength / intensity of radiation and location of cell towers.
   - The frequency of data transfer will be determined by the purpose and specific application / use-case.

8.13. An HVD for Obesity data is set-up by the Ministry of Health and Family Welfare with the department playing the role of a data trustee for this dataset.

i. It will work with government and private hospitals, corporates (departments that look into their employee health and welfare), food companies to build this dataset.
ii. The data trustee may make a data request accordingly.
   - Body Mass Index of employees calculated through Employee Wellness programs.
   - The sale of food items (above a certain calorific value) on an e-Commerce platform and food manufacturing companies.
   - The data requested will be for a subset of the data fields collected by these Data Custodians.
   - The frequency of data transfer will be determined by the purpose and specific application / use-case.

8.14. The Committee recommends a process for data-sharing request for using a High-value Dataset.

i. data trustee collects Non-Personal Data that constitutes high-value dataset from public and private data custodians.
ii. A data request may be made by any organization registered in India to the corresponding data trustee.
iii. Data requests by persons to High-value Datasets are not allowed.
iv. The data trustee of the High-value Datasets may levy a nominal charge (towards data infrastructure, data processing) to the data requesters.
v. This report recommends a technology architecture for High-value Datasets.



30

8.15. The Committee proposes mechanisms which can provide certain checks & balances to the HVD creation and sharing process.

i. Location – The non-personal data comes from multiple facets of people's lives and are prone to deanonymization and if exposed would constitute a critical loss of privacy. Hence, non-personal data derived from personal data shall inherit the sensitivity of the underlying personal data for storage requirements as specified in the PDP Bill.
   - For example, non-personal data about health of people (even though it may be anonymised and aggregated), will inherit for purpose of storage requirement the sensitivity of the underlying data (on health) which is classified as Sensitive Personal Data as per Clause 3 (36) of the PDP Bill.
   - For example, non-personal data collected about say, broadband subscription in a city (when aggregated and anonymised) will inherit for purpose of storage requirement the sensitivity of the underlying data (on broadband subscription) which is classified as general Personal Data.

ii. Tools – Testing and probing tools are continuously run on the data in secure clouds and reports generated, auto-submitted by cloud providers and registered organisations to check compliance.

iii. Liability – Organisations are to be indemnified against any vulnerability found as long as they swiftly remedy it and adopt a standards-driven approach (like annual light-weight, self-reported, self-audited digital compliance reports).

iv. Academia-Industry Innovation Advisory Body – The NPDA shall establish an innovation advisory body consisting of highly accomplished experts from academia, Government, industry and society to develop / enhance / innovate on aspects like data sharing, data governance, technical standards like interoperability, privacy-protection, and data stewardship.

8.16. Refer to **Appendix 6** for the background information that the Committee considered on data creation and sharing rules and regulations, mechanisms and approaches in other countries.

9. Analysis of the NPD Framework – Legal and Economic

9.1. The Committee did a legal analysis of proposed NPD Governance Framework from the perspective of Property Law, Copyright Law, Trade Secrets Law, IT Act 2000, Competition law and the Indian Constitution.

9.2. Indian property law has not recognised a property right (akin to ownership of land or goods) over data and there are no statutory protections in this regard. However, certain rights – in the nature of proprietary rights – have conventionally been derived from two sources - copyright and trade secrets law.

9.3. Copyright Law:

i. Copyright law does not protect ideas or raw data. Section 2(o) of The Copyright Act, 1957 (Copyright Act) does, however, protect compilations of data that form original databases as literary works and allows the authors of such databased to exercise ownership in the form of copyright over them. The Copyright Act does not grant sui generis protection to databases and requires an organisation to have exercised some non-trivial skill and creativity in compiling and organising the database in order to claim copyright protection. The nature of data that comprises this database is irrelevant to the determination of whether or not a database is protected under copyright law and it is the originality in the compilation and organisation of the database that determines its copyrightability.[7]

ii. Therefore, even if the database is composed of entirely raw public non-personal data, but this data has been compiled or organised by the exercise of some skill or creativity, copyright would vest in such a database. On the other hand, where the grant of copyright over a compilation of data would in effect amount to conferring a property right over the underlying data, in such circumstances the database would not be copyrightable. An example of this sort of a database would be where there is only one way to express the compilation of data.[8]

iii. As a result, while it is plausible to argue that any purely raw public non-personal data may not be covered under the Copyright Act, if the entity collecting this data deploys skill and creativity in the very act of compilation, then the resultant database that it possesses may be protected.

iv. As per the Committee recommendations, data sharing may be mandated only for designated high value data-sets, where the fields for data to be shared are also pre-determined (which are expected to be a subset of the fields in the original database), and are relatively straight-forward. If the extraction is done per given pre-set fields, such extraction would not violate the database design copyright.

[7] *Tech Plus Media Private Ltd*. v *Jyoti Janda and Ors*. 2014 SCC OnLine Del 1819 (Para 23).
[8] *Emergent Genetics India Pvt. Ltd.* v *Shailendra Shivam* 2011 (47) PTC 494 as interpreted in *Tech Plus.*

9.4. Trade Secret Law

i. India presently does not have a statute that codifies trade secret protection. Accordingly, this form of intellectual property protection is derived from two sources – contractual obligations, and equity. The former arises when there is a contract between two parties requiring them to keep confidential certain information that one party has acquired through some amount of investment and skill and, which has been provided to the other party.

ii. While it is hard to lay down a bright line test as to what information constitutes a trade secret under principles of equity, it is clear that any information that, by its nature and context, may be expected to be confidential would be considered to be its trade secret. Despite there being precedent to suggest that trade secret protection may also extend to the raw data that forms a part of the alleged trade secret[9], the majority of case law suggests that ordinarily and freely available raw data would not be granted trade secret protection[10].

iii. The entity claiming trade secret protection has to demonstrate that the information was inherently of such a nature so as to be protected as a trade secret. Further, the majority of case law in this regard has been in the context of claims between parties where there already existed a relationship of confidence (such as employment, business partnership etc) and there is no consistent precedent recognising trade secret protection over data in the form of property rights enforceable against third parties.

iv. Consequently, while it might be difficult to categorise raw public non-personal data as falling within the realm of trade secret protection, if the act of compiling or processing any non-personal data leads to an inherently non-public and secret compilation of data, then such a compilation of data would be entitled to trade secret protection. However, as mentioned above, this protection is unlikely to cover a proprietary right over this data to prevent the eminent domain of this data.

9.5. Data under The Information Technology Act

i. The Information Technology Act, 2000 (IT Act) defines data in Section 2(o) as "a representation of information, knowledge, facts, concepts or instructions which are being prepared or have been prepared in a formalised manner, and is intended to be processed, is being processed or has been processed in a computer system or computer network, and may be in any form (including computer printouts magnetic or optical storage media, punched cards, punched tapes) or stored internally in the memory of the computer."

[9] *Burlington Home Shopping Pvt. Ltd* v *Rajnish Chibber and Anr.* 61 (1996) DLT 6
[10] *Diljeet Titus, Advocate* v *Alfred A. Adebare* 2006 (32) PTC 609 (Delhi); *American Express Bank Ltd.* v *Ms. Priya Puri* 2006 (110) FLR 1061

ii. However, there are no provisions in the IT Act that create property rights in data. While the privacy protections for sensitive personal data set out in Section 43A are the closest when it comes to establishing proprietary rights over data, it is likely that this provision will be repealed once the Personal Data Protection Bill, 2019 (PDP Bill) comes into force.

9.6. Data under the Competition Act

i. While the competition regime in India has taken note of the impact of data on competition regulation – particularly in terms of network effects and market power, the Competition Act, 2002 is primarily oriented towards the regulation of anti-competitive effects of this data, as opposed to creating any rights in the data itself. The regulations set out in the context of the competition regime would also therefore not be relevant to this exercise.

9.7. The Indian Constitution

i. The Indian Constitution provides significant covers for a data sharing law. Its Article 39 (b) and (c) provide as Directive Principles that the State shall, in particular, direct its policy towards securing, (b) that the ownership and control of the material resources[11] of the community are so distributed as best to subserve the common good; and (c) that the operation of the economic system does not result in the concentration of wealth and means of production to the common detriment. Constitution's Article 31 C shields from constitutional challenges legislations that give effect to the intent of 39 (a) and (b). It can be argued that a data sharing law is in pursuance of objectives of Articles 39 (a) and (b) and is thus protected by 31 C from constitutional challenges (related to Articles 14 or 19). Courts have averred, however, that an adequate nexus must be shown between a legal measure giving effect to 39 (a) and (b) and the intent and wording of these sections.

9.8. The Committee analysed the basis for establishing community rights over non-personal data. Refer to **Appendix 7** for background information that the Committee considered.

i. A review of current and evolving governance frameworks for community resources – traditional knowledge, natural resources, a community's genetic resources and so on, from both global and national levels, bring up five key principles that seem to underpin them. (1) a community's right over resources associated collectively with it, (2) consent of the community for use of such resources, (3) benefit sharing with the community, (4) transparency in recording

11 The manner in which courts have interpreted the term 'material resources' in 39 (b) to include "every thing of value or use in a material world" (State of Karnataka v Ranganath Reddy AIR 1978 SC 215), it can be considered to include informational or data resources, since data is certainly of immense value in the contemporary material world. Precedents for such interpretation also exists in international law.

community resources to prevent misuse and enable easy access of the legitimate kind, and (5) community's participation in governance of community resources.

ii. The Committee considered these principles in formulating its recommendations. For instance,
   - Principle (1) – enables the community to establish its rights over NPD.
   - Principle (3) – enables a community to seek sharing of NPD collected about it by various parties (through data trustee creating an HVD). At the same time, such community rights may not interfere with a data collector's own access to and use of the concerned data resources.
   - Principle (4) – guides the concept of High-value Dataset.
   - Principle (5) – underpins community's right to govern its NPD through appropriate data trustees.

9.9. The Committee did an economic analysis of the proposed NPD governance framework. This conceptual analysis is based on Ronald Coase's work which is used to determine institutional models used in public policy. The key principles considered for analysis are transaction costs and externalities.

i. Transaction costs. The feedback received from organizations representing different actors in the ecosystem emphasised that the governance framework should not unduly increase the transaction costs, especially compliance related, IP related, etc.
   - The proposed model takes this feedback into account and ensures that the transaction costs are reduced.
   - For example, registration for a Data Business is a one-time activity and this is a disclosure-based rather than license-based compliance.
   - Data creation and sharing is recommended only in the context of HVDs.
   - Data sharing is compliant with IPR laws.

ii. Externalities – As the HVD creation and sharing process evolves, NPDA should consider actions to address certain negative externalities that may emerge in the future:
   - Some data trustees can become powerful and can try to appropriate many more HVDs under their "control". Using this "control" they can aim for regulatory capture (influence policy), influence over data custodians etc.
   - Large and powerful Indian registered organizations with non-Indian control can become disproportionate users of HVDs.
   - Distinguish between 'active misuse of NPD' and 'accidental misuse of NPD' so that innovations in the use of non-personal data are not stifled.

10. Technology Architecture

The Committee considered some technology related guiding principles that can be used for creating and functioning of shared data directories / data bases, and for digitally implementing the rules and regulations related to HVD creation and sharing.

10.1. The guiding principles for such a technology architecture include:

i. Mechanisms for accessing data – A number of different mechanisms exist for accessing data including downloads, Application Programming Interfaces (APIs), and data sandboxes.
   - All sharable non-personal data and datasets created or maintained by government agencies, companies, startups, universities, research labs, non-government organisations, etc. should have a REST (Representational State Transfer) API for accessing the data.

ii. Distributed for data security – data storage in a distributed format so that there is no single point of leakage; sharing to be undertaken using APIs only, such that all requests can be tracked and logged; all requests for data must be operated after registering with the company for data access etc. Even when data is stored in a distributed or federated form, as appropriate, there could be coordinated management of them like would be required for data infrastructures for important non-personal data in different sectors.

iii. Creating a standardized data sharing approach – should be able to take-in any form of data and produce output that is standardized and usable to all stakeholders.

iv. Prevent de-anonymization – Best of breed Differential Privacy algorithms may be used to create anonymized data. Mechanisms must be put in place to ensure that re-identification of anonymized data does not occur.

10.2. The Committee has encapsulated these technical guiding principles into an illustrative three-tiered system architecture spanning legal safeguards, technology and compliance. **Refer to Appendix 8**. There may also be other appropriate ways to technically implement the recommendations of this Committee.

Appendix 1: List of Committee Members

Members of the Committee

i)	Shri Kris Gopalakrishnan, Co-Founder Infosys	Chairman
ii)	Additional Secretary / Joint Secretary, DPIIT	Member
iii)	Ms. Debjani Ghosh, President NASSCOM	Member
iv)	Dr. Neeta Verma, DG, National Informatics Centre	Member
v)	Shri Lalitesh Katragadda, Founder Indihood	Member
vi)	Prof. Ponnurangam Kumaraguru, IIIT Delhi	Member
vii)	Shri. Parminder Jeet Singh, IT for Change	Member
viii)	Additional Secretary, MeitY	Member Convenor
ix)	Krishnan Narayanan, N. Dayasindhu, itihaasa Research and Digital	Report Preparation
x)	Rahul Matthan, Partner Trilegal	Report Preparation

37

Appendix 2: Data – Trends and Socio-Economic Impact

Data availability and value generation from data

1. The world is awash with data. Planet scale adoption of the Internet, smartphones, and cloud driven apps, followed by increasing use of AI-systems are the main reasons why we are generating and consuming data at a scorching pace.
    i. There are over 3 billion smartphone users in the world12. Instagram had over 277,000 stories posted, Google had over 4.4 million searches and Uber had over 9,700 rides every minute of the day in 201913.
    ii. Estimates suggest that the world will generate about 90 zettabytes (approximately a billion terabytes) of data in this year (2020) and the next, more than all the data produced since the advent of computers[14]. By 2025, worldwide data is expected to grow to 175 zettabytes, with much of the data residing in the cloud[15].
    iii. AI techniques like Machine Learning (ML) and deep learning require large data sets to provide accurate prediction models. A public database used to build deep learning models like Imagenet has more than 14 million hand-annotated images[16].

2. Digital transformations are happening all around the world. A proliferation of big data, analytics and AI has led to the creation of many new data intensive services and the transformation of existing services into data intensive services.
    i. It is estimated that the global AI-derived business value in 2020 is likely to be about USD 2.65 trillion[17]. Between 2018 and 2019, organizations that have deployed AI grew from 4% to 14%[18].
    ii. Abundant availability of data is a primary driver for AI. We are witnessing increased traction for AI solutions in India. This AI powered economic growth in India has not only created new services but has also improved the quality of

12https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/, accessed on 15/03/2020
13https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/#1dd7255b3d66
14 https://www.economist.com/special-report/2020/02/20/a-deluge-of-data-is-giving-rise-to-a-new-economy
15https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html
16https://www.newscientist.com/article/2127131-new-computer-vision-challenge-wants-to-teach-robots-to-see-in-3d/
17https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/
18https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/

existing services. NASSCOM forecasts that India's analytics revenue in 2025 will be around USD 16 billion USD, about 32% of the global market[19].

iii. The demand for AI has in turn created a demand for AI talent in India. According to NASSCOM, the total demand for AI and big data, analytics talent in India is likely to grow from around 510,000 in 2018 to about 800,000 in 2021[20].

3. Traditionally there was value in selling processed data. Today, the typical process of value creation from data is as follows:
   i. Data collection
   ii. Cleansing and curating raw / factual data
   iii. Populating databases in standardized formats
   iv. Doing data mining and data analysis using various tools and techniques
   v. Using curated data to train AI/ML systems
   vi. Converting information into insights that help in prediction and decision making for revenue / profit generation as well as for social and public interest activities.

4. There are three ways in which organizations realize the value of their data – 1) Direct monetization, 2) Internal investments, and 3) Mergers and acquisitions. There are a number of approaches developed to measuring the value of data and this is an evolving field[21, 22, 23, 24, 25, 26].

5. Frameworks are being developed to better understand the uses and benefits of data and its value[27] including 1) Treating data as an asset 2) Activity or usage value of data 3) Future value of data and 4) Prudent value of data.
   i. Data is treated as an asset and monetized directly by trading it or building a service on top of the data.

---

19 https://community.nasscom.in/wp-content/uploads/attachment/nasscom-indian-analytics-data-to-decisions-june-2016-sec.pdf
20 https://www.nasscom.in/knowledge-center/publications/talent-demand-supply-report-ai-big-data-analytics
21 Chiehyeon Lim et al., "From data to value: A nine-factor framework for data-based value creation in information-intensive services", International Journal of Information Management, Volume 39, April 2018, Pages 121-135
22 Michael Chui et al., "Notes from the AI frontier: Applications and value of deep learning", McKinsey Global Institute, April 2018
23 Asha Saxena, "What is Data Value and Should it be Viewed as a Corporate Asset?", Dataversity, March2019
24 John Akred and Anjali Samani, "Your Data Is Worth More Than You Think" MITSloan Management Review, January 2018
25 Hanna KozlowskaJuly, "How much is your data worth?", Quartz, July 2019
26 AmirataGhorbani and James Y. Zou, "What is your data worth? Equitable Valuation of Data", https://arxiv.org/pdf/1904.02868.pdf
27 John Akred and Anjali Samani, Your Data Is Worth More Than You Think, MIT Sloan Management Review, January 2018

ii. Data's value is based on the number of users and frequency of data access. Unlike a physical asset, the more a data is used, the more valuable it is likely to become.

iii. Data is treated as an intangible asset whose value may be discoverable at a future date, say during a Mergers & Acquisition activity.

iv. The prudent value approach values data sets based on the extent to which they could advance key business initiatives that support a company's overall business strategy.

Imbalance in data and digital industry

6. Some examples of the data based businesses include – social media, search, map-based services, online retail, ride-hailing platforms, digital healthcare, credit rating, etc.

   i. User data and user generated content are collected and analysed often with AI to make better decisions for businesses and organizations. Our society experiences such data-enabled services in the form of platforms like Google Maps, Uber, Amazon, etc.

   ii. It is reported that Google and Facebook together control about 60% of the Internet advertising market in the USA[28]. It is also estimated that Amazon had a 37% share of the online ecommerce market in the USA in 2019[29]. This is reflected in the very large market capitalization of these corporations.

7. For a few companies that dominate the digital and Data Business, the network effects lead to outsized benefits and creates a certain imbalance in the data/digital industry.

   i. So far, a few startups from the 1990s and 2000s have gone on to become USD 1 trillion market capitalisation multinational corporations. One of the primary drivers of value of these companies is their ability to collect and analyse data of users which often leads to network effects that help them grow and become very dominant actors in the economy. These companies have also been in the forefront of adopting AI to analyse this data.

   ii. In the list of the worlds' 70 largest platforms with respect to market capitalisation – America has 73%, China has 18% and Europe has 4% of the platforms[30].

   iii. For example, the United States allowed patenting of human genes isolated in labs. And at one point, it is estimated that about 20% of human genes were

28 https://www.reuters.com/article/us-alphabet-facebook-advertising/google-facebook-have-tight-grip-on-growing-u-s-online-ad-market-report-idUSKCN1T61IV
29 https://www.bloomberg.com/news/articles/2019-06-13/emarketer-cuts-estimate-of-amazon-s-u-s-online-market-share
30 https://www.economist.com/business/2020/02/20/the-eu-wants-to-set-the-rules-for-the-world-of-technology

patented. This helped some companies become a monopoly in diagnostic testing services based on genes they had isolated and hence garner disproportionate profits. In 2013, a landmark judgement of the US Supreme Court ruled that naturally occurring DNA segment is a product of nature and not patent eligible merely because it has been isolated[31].

iv. In a data economy, companies with the largest data pools have outsized, unbeatable techno-economic advantages. For example, studies[32] have shown that increasing a speech corpus size by 5 times reduces word-error-rate (i.e. errors in speech to text translation) by 10% or more, while cutting cost by significantly reducing the need for manual rating. Such a 10% reduction in error-rate used to take a generation of research earlier. But now, access to exponentially increasing data set sizes, large R&D budgets and unprecedented computing power are making it possible in much shorter time periods.

v. A combination of a "first mover advantage" for these large data-driven platforms and businesses, with the sizable network effect and enormous data that they have collected over the years, has left many new entrants and start-ups being squeezed and faced with significant entry barriers. This may be the right time to set out rules to regulate the data ecosystem (which includes data collection, analysis, sharing, distribution of gains, destruction etc.) to provide certainty for existing businesses and provide incentives for new business creation, as well as to release enormous untapped social and public value from data.

vi. India is second most populous country in the world[33]. India also has the second highest number of smartphone users in the world[34]. Given this, and the current levels of Internet penetration in India, India can arguably be projected as being one of the top consumer markets, and by extension data markets in the world in the foreseeable future. Allowing the possibility of data monopolies, in a large consumer market such as India, could lead to the creation of imbalances in bargaining power vis-à-vis few companies with access to large data sets accumulated in a largely unregulated environment, on one side, and Indian citizens, Indian businesses including startups, MSMEs and even the Government, on the other. Therefore, the Government's role is to catalyse the Data Businesses in a manner that maximizes overall welfare.

vii. At the same time, the requirement for providing certainty and incentives for new business creation cannot be understated. It is because of robust IP rights, various data related privileges, that a lot of data-driven innovation has occurred.

[31] Supreme Court of the United States, Association of Molecular Pathology et. Al. vs. Myriad Genetics Inc. et. al. https://www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf
32 https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43230.pdf
33https://www.worldometers.info/world-population/india-population/
34https://www.statista.com/statistics/748053/worldwide-top-countries-smartphone-users/

Therefore, while ensuring that markets function properly, sufficient and adequate incentives for new business creation must therefore be safeguarded.

viii. Lastly, potential harms could arise in terms of privacy violations arising from re-identification of anonymized data, or from the derivation of personally identifiable insights from non-personal data. Adequate measures would have to be developed in order to ensure that any data sharing framework does not dilute the protections afforded by the Personal Data Protection Bill, 2019 (PDP Bill). Accordingly, any eventual regulation will have to mitigate against the risks of privacy harms.

ix. Not only economic, but most key social, political and cultural activities will depend upon data, and suitable access to it. For instance, governments would need wide access to data in all sectors for public policy development and delivery of public services. While public agencies produce a lot of data, much of the required data will be collected by and be in the hands of private companies. Besides data philanthropy, some systematic mechanisms need to be developed to tap the social and public value of data.

The Case for Regulating Data

8. Data creates economic value and wealth, apart from enormous social and public value. Data therefore is increasingly taking the centre-stage in core-technological businesses, all economic sectors around the world and in addressing various social and public administration issues. It is in this context, that the Committee has sought to set out the case for regulation of data. As a starting point therefore, one needs to understand the nature of data as an economic good, as also its social and public value. In this regard, data can be viewed through two lenses[35] – economic and informational.

    i. Data as an economic resource has huge externalities: From an economic lens, data is non-rivalrous, yet excludable, and its use could have both positive and negative externalities.

    ii. Data offers intrusive information about its subject: From an informational lens, one needs to recognise and understand the subject, content and use of data, and understand how any content and use of data, could give rise to harms. For instance, sensitive or personal data could lead to privacy harms. Even non-personal data, including anonymized Personal Data, could provide collective insights that could open the way for collective harms (exploitative or discriminatory harms) against communities.

    iii. Collective information / data is needed for social and public interest use. An instance of a collective harm is when such data is closed for public use and leads to welfare losses.

[35] Bennett Institute for Public Policy and Open Data Institute, "The Value of Data – Policy Implications", 2020

- Collective Privacy is an emerging concept[36,37,38] that will need to be examined and defined in detail in the future.

iv. Market transactions and market forces on their own will not bring about the maximum social and economic benefits from data for the society. Appropriate institutional and regulatory structures are essential for a thriving data economy and a well-functioning data society.

[36] https://link.springer.com/chapter/10.1007/978-3-319-46608-8_8
[37] https://link.springer.com/article/10.1007/s13347-019-00351-0
[38] https://www.springer.com/gp/book/9783319466064

Appendix 3: Examples of Non-Personal Data

1. Data may be categorised in many ways:
   - **i)** Arising from the subject of data (e.g. personal data)
   - **ii)** In relation to its purpose (e.g. AI training data, e-Commerce data)
   - **iii)** The sector to which it belongs (e.g. health data)
   - **iv)** The source of data (e.g. soil data)
   - **v)** The level of processing (raw / factual versus derived data)
   - **vi)** The collector of data (e.g. public / Government or private data)
   - **vii)** The extent of involvement of stakeholders in the creation of data (provided, observed, derived, or inferred)

2. A mixed dataset, which represent a majority of datasets used in the data economy, consists of both personal and non-personal data.
   i. In the European Union context, the non-personal data Regulation applies to the non-personal data of mixed datasets; if the non-personal data part and the personal data parts are 'inextricably linked', General Data Protection Regulation apply to the whole mixed dataset.

3. Categorisation of data based on its creation[39] – A categorisation of data can help assess the extent to which different stakeholders are involved in the creation of data, including cases where users (consumers and businesses) interact with a data product (good or service) such as an e-government service, a social networking service, etc.
   i. One approach includes four categories of data: i) provided (applications registrations, survey responses, social network postings etc.); ii) observed (cookies on a website, data from sensors etc.); iii) derived (computational scores, classification based on common attributes etc.); and iv) inferred data (scores developed using statistical, advanced analytical techniques, or AI/ML).
   i. Such a categorization helps in framing regulation & policy. For example, in the European Union, the right to data portability under the GDPR would include 'provided' as well as 'observed' data. It would however exclude data 'derived' (& 'inferred') from additional processing – data that are often considered proprietary.

4. Anonymized Data
   i. Anonymization allows data to be shared, whilst preserving privacy. The process of anonymising data requires that identifiers (both direct identifiers like names

39 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019, https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en

and indirect identifiers like age or occupation) are changed in some way such as being removed, substituted, distorted, generalised or aggregated[40].

ii. However, new research[41] shows that current methods for anonymizing data still leave individuals at risk of being re-identified. So, policymakers should be careful about what constitutes anonymized data. Also, the technical specifications and architecture should ensure that the chances of re-identifying anonymized data are minimised significantly.

iii. The Committee has collated some of the basic anonymization techniques in this report in Appendix 3: Primer on Anonymity.

5. AI Training Data

i. During the development of an AI system, three different sets of data are required to train, fine-tune and test the machine learning models. They include the training dataset, the validation dataset, and the testing dataset. The training data set will include input data and expected results and is used to train a machine learning algorithm. These are typically mixed data sets.

ii. Training data for autonomous vehicles in India would include data on Indian roads and vehicles. Training data on fashion purchases in India would include data on purchases of clothes and the buyers on an e-Commerce platform.

iii. In the case of Generative Adversarial Networks (GANs), two AI engines compete against each other to produce data for reinforced learning for the underlying AI system. This may be considered an example of a derived non-personal data.

6. E-Commerce Data

i. e-Commerce data relate to customers' orders, needs, preferences, interests, shopping patterns, feedback, customer satisfaction level, delivery times etc. It also includes insights related to products on the store, competitors' data, and technical data as well. These are typically mixed data sets.

ii. Typical e-Commerce Data attributes[42]

- Customer demographics like age, gender, location
- Product Discovery KPI - the factors that help understand how and when customers find the product
- Onsite traffic metrics - the factors that reveal the amount and time of traffic to a web store
- Email / social media engagement

[40] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation.aspx
[41] https://www.sciencedaily.com/releases/2019/07/190723110523.htm
42 https://datarade.ai/data-categories/ecommerce-data/guide

- Conversion attributes - conversion of visitors into customers on a particular e-Commerce store

7. Government or Public Data
   i. The Open Government Data (OGD) Platform of India or data.gov.in is a platform supporting the open data initiative of Government of India. It provides access to datasets and documents published by ministries / departments of the Government of India. India may build on its OGD initiative and expand on its national data strategy.
   ii. Some countries have started to specify a new class of data at a national level – data of public interest or high-value dataset, like geospatial and/or transportation data. The Governments are combining both public and private sector data as well as personal and non-personal data to create such data of public interest[43].
      - Australia has classified its Geocoded National Address File (G-NAF) as one of its most high-valued data sets.
      - In Germany, the government has established the research initiative mFUND, to support the development of data-based business models for smart mobility (Mobility 4.0). A central aspect for the programme is the provision of mobility and geo-data (e.g. transport and traffic data, hydrological data, climate and weather data). For this purpose, data access and sharing are promoted according to open data principles and technically supported by the creation of a central, open data access point for mobility-related data (mCLOUD). This initiative is funded by the German Federal Ministry of Transport and Digital Infrastructure with EUR 150 million to be invested between 2016 and 2020.
      - National governments have started to specify a new class of data at a national level – data of public interest or high-value dataset, like geospatial and/or transportation data[44].
      - The European Commission is proposing to create nine common European data spaces - industrial (manufacturing), green deal, mobility, health, financial, energy, agriculture, public administration, and skills[45].
      - The European Commission has identified six data types that appear to have the most value: geospatial, earth observation and environmental, meteorological, statistics, company data, and transport data[46].

43 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019, https://www.oecd-ilibrary.org/science-and-technology/enhancing-access-to-and-sharing-of-data_276aaca8-en
44 OECD, "Enhancing Access to and Sharing of Data : Reconciling Risks and Benefits for Data Re-use across Societies", 2019,
45 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf

8. Private Data
    i. Private non-personal data is data collected by private players from and about things, processes, etc that are entirely private to them, or owned by them, or those aspects of 'derived data' which arise from private effort
        - It includes inferred or derived data / insights involving application of algorithms, propriety knowledge.
        - The example of two AI engines competing against each other to produce derived data for reinforced learning for the underlying AI system is an example of private non-personal data.
        - It may also include a global dataset that contains information about non-residents collected in foreign jurisdictions (other than India).

46 Open Knowledge Foundation, 'What data counts in Europe? Towards a public debate on Europe's high value data and the PSI Directive', 2019

Appendix 4: Primer on Anonymity

A primer on anonymization techniques is provided here. Some of these techniques are academic pursuits and some of them are methods already used in industry tools.

1. K-anonymity[47]
   i. A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appear in the release[48]. This is one of the most popular and old techniques for structured data.

2. L-diversity[49]
   i. The l-diversity model is an extension of the k-anonymity model which reduces the granularity of data representation using techniques including generalization and suppression such that any given record maps onto at least k-1 other records in the data. The l-diversity model handles some of the weaknesses in the k-anonymity model where protected identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalized or suppressed, especially when the sensitive values within a group exhibit homogeneity. The l-diversity model adds the promotion of intra-group diversity for sensitive values in the anonymization mechanism[50].

3. T-closeness[51]
   i. An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness[52].

4. Diffix (High-Utility Database Anonymization)[53]
   i. Diffix acts as an SQL proxy between the analyst and an unmodified live database. Diffix adds a minimal amount of noise to answers—Gaussian with a standard deviation of only two for counting queries—and places no limit on the number of

[47] http://dataprivacylab.org/dataprivacy/projects/kanonymity/kanonymity.pdf
[48] https://en.wikipedia.org/wiki/K-anonymity
[49] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and MuthuramakrishnanVenkitasubramaniam. 2007. L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1, 1, Article 3 (March 2007). DOI=http://dx.doi.org/10.1145/1217299.1217302https://personal.utdallas.edu/~muratk/courses/privacy08f_files/ldiversity.pdf
[50] https://en.wikipedia.org/wiki/L-diversity
[51] N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 106-115. doi: 10.1109/ICDE.2007.367856 https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf
[52] https://en.wikipedia.org/wiki/T-closeness
[53] https://aircloak.com/wp-content/uploads/Diffix-High-Utility-Database-Anonymization.pdf

queries an analyst may make. Diffix works with any type of data and configuration is simple and data-independent: the administrator does not need to consider the identifiability or sensitivity of the data itself.

5. ARX[54]
   i. ARX is another tool to anonymize data. ARX is divided into four perspectives, which model different aspects of the anonymization process. As is shown below, these perspectives support 1) configuring privacy models, utility measures and transformation methods, 2) exploring the solution space, 3) analysing data utility and 4) analysing privacy risks. ARX is built on many research publications, they have a team to maintain the code, bug fixes, etc.

6. Amnesia[55]
   i. Amnesia is a flexible data anonymization tool that transforms relational and transactional databases to dataset where formal privacy guaranties hold.
   ii. Amnesia is a data anonymization tool, that allows to remove identifying information from data. Amnesia not only removes direct identifiers like names, SSNs etc but also transforms secondary identifiers like birth date and zip code so that individuals cannot be identified in the data. Amnesia supports k-anonymity and km-anonymity.
   iii. km-anonymity requires that each combination of up to m quasi identifiers must appear at least k times in the published data. The intuition behind km-anonymity is that there is little privacy gain from protecting against adversaries who already know most of the terms of one record, and significant information loss in the effort to do so.
   iv. There is an online GUI based system of Amnesia[56].

7. μ-ARGUS & τ-ARGUS[57]
   i. μ-ARGUS to be used to protect microdata and τ-ARGUS to be used to protect tabular data.
   ii. These tools are available in both Windows and other platforms[58].

8. Anonimatron[59]
   i. There are also publicly available open source projects on anonymization, including GDPR compliant testing. Some of the features of Anonimatron are:

[54] https://arx.deidentifier.org/
[55] https://amnesia.openaire.eu/
[56] https://amnesia.openaire.eu/amnesia/
[57] http://neon.vb.cbs.nl/casc/mu.htm
[58] https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_7_PPdeWolf.pdf
[59] https://realrolfje.github.io/anonimatron/

- Anonymize data in databases and files.
- Generates fake email addresses, fake Roman names, and UUID's out of the box.
- Easy to configure, automatically generates example config file.
- Anonymized data is consistent between runs. No need to re-write your tests to handle random data.
- Extendable, easily implement and add your own anonymization handlers
- 100% Java 1.8, multi-platform, runs on Windows, Mac OSX, Linux derivatives.
- Multi database, uses SQL92 standards and supports Oracle, PostgreSQL and MySQL out of the box. Anonimatron will autodetect the following JDBC drivers: DB2, MsSQL, Cloudscape, Pointbase, Firebird, IDS, Informix, Enhydra, Interbase, Hypersonic, jTurbo, SQLServer and Sybase.

9. Differential Privacy

i. Goal is to perform aggregative analysis (statistics about the data) without compromising the privacy of an individual data point[60]. Differential privacy offers strong and robust guarantees that facilitate modular design and analysis of differentially private mechanisms due to its composability, robustness to post-processing, and graceful degradation in the presence of correlated data[61]. This method is prominently used in technological implementations now, etc. Apple uses differential privacy in its iPhone.

[60] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Third conference on Theory of Cryptography (TCC'06), Shai Halevi and Tal Rabin (Eds.). Springer-Verlag, Berlin, Heidelberg, 265–284.https://link.springer.com/chapter/10.1007%2F11681878_14

[61] https://en.wikipedia.org/wiki/Differential_privacy

Appendix 5: Emerging Global Frameworks related to Data Business

In the data economy, the proliferation of big data, analytics and AI has led to the creation of information intensive services where information interactions exert the greatest effect on value creation. Thus, a new category of business, 'Data Business', may be envisaged that collects / manages / or otherwise manages data, and meets certain threshold criteria.

i. One study[62] developed a nine-factor framework for data-based value creation in information-intensive services. The factors include (1) data source, (2) data collection, (3) data, (4) data analysis, (5) information on the data source, (6) information delivery, (7) customer (information user), (8) value in information use, and (9) provider network.

Globally, such a concept of defining a new category of 'Data Business' is only emerging. Here are a few examples of related global taxonomies.

1. Bureau of Economic Analysis (BEA), USA definition of Digital Economy[63] – BEA in a 2018 working paper includes the following categories under Digital Economy:
   i. Digital-enabling infrastructure needed for a computer network to exist and operate – computer hardware, software, telecommunications equipment and services, structures like data centres, IoT, and support services
   ii. e-Commerce – digital transactions that take place using that system – Business-to-business (B2B) e-commerce, Business-to-consumer (B2C) e-commerce, Peer-to-peer (P2P) e-commerce
   iii. Digital media – the content that digital economy users create and access

2. OECD classification of data-enabled services[64] – In a 2018 paper on recording and measuring data, OECD categorizes data-enabled services as follows:
   i. Providing services for free or at very low prices to gather data of users which are subsequently used to detect behavioural patterns to provide other producers with targeted advertising services (like Google Ads, Facebook, etc.), or to offer other services (e.g. using information from payment systems etc.)
   ii. Using data generated as part of the primary production process, to improve the efficiency of the internal operations and/or to detect behavioural pattern to

62 Chiehyeon Lim et al., "From data to value: A nine-factor framework for data-based value creation in information-intensive services", International Journal of Information Management, Volume 39, April 2018, Pages 121-135
63 https://www.bea.gov/sites/default/files/papers/defining-and-measuring-the-digital-economy.pdf
64http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=SDD/CSSP/WPNA(2018)5&docLanguage=En

support own sales. (like Amazon using dynamically generated recommendations, Walmart using analytics to optimise supply chain and pricing models.)

iii. Creation of new types of services by using and analysing big data.

iv. Provision of data-related services by collecting data from a vast number of different, mostly free, available data sources, normalising formats and providing access, with revenues from subscription or usage fees.

v. Data facilitators, providers of data tools such as providing storage media, servers and workstations, data collection, analysis and visualisation software, database management software, encryption technology and software, data protection technology, etc.

vi. Creation of freely available information or knowledge by communities of people, providing their contributions for free. (like Wikipedia, ResearchGate)

3. A framework[65] for establishing the 'data-drivenness' of a market:

   i. Market definition (user centric) – an index of data-drivenness applied at the industry level would indicate, for instance, industry A has a high degree of data-drivenness and therefore mandatory data sharing is warranted, whereas industry B is only mildly data-driven such that there should be no mandatory data sharing.

   ii. Study the demand side of the market: what drives users' consumption utility?

   iii. Study the supply side of the market: what drives objective measures of product quality?

65 Jens Prüfer, Friedrich-Ebert-Stiftung, "Competition Policy and Data Sharing on Data-driven Markets", 2020, library.fes.de/pdf-files/fes/15999.pdf

Appendix 6: A Snapshot of some Global Rules and Regulations around Data Sharing

To facilitate data sharing, rules and regulations need to be established. These rules and regulations may address aspects like data regulator, user registration, data disclosure requirements, audit requirements, data usage context and others.

1. Different countries are adopting different strategies and experimenting with regulations to govern data.
    i. The European Commission has published a slew of communications on 'A European strategy for data[66]', and 'Shaping Europe's digital future[67]' and a white-paper on 'On Artificial Intelligence - A European approach to excellence and trust[68]'.
    ii. The last G20 meeting launched the 'Osaka Track', a proposed plurilateral agreement on digital trade, that provided global rules for "data governance" based on "free flow of data with trust". India and a few other developing countries have refused to sign up to the Osaka Track[69].
    iii. In Germany, the Federal Ministry for Economic Affairs announced a federated data infrastructure called "Gaia-x", a legal-cum-software layer to implement granular national data policy, that would allow firms to move data and computing workloads between rival clouds more easily.
    iv. Some Western countries may soon discuss a "Data Freedom Act" which would create a new regulated entity for that purpose[70]. Stronger privacy laws are only a first step, but not enough. 'Financial interests' (economic value of the data pertaining to an individual / community) and 'Control interests' (purposes for which the data of the individual / community may be used) exist beyond privacy interests.[71]
    v. Several jurisdictions such as the EU and US have already initiated investigations into the impact of virtual data monopolies on competition in the market. For example, recognizing these very imbalances, the German Competition Law was amended in 2019, empowering the German Bundeskartellamt with wider powers of monitoring and enforcing competition rules in the Digital Economy. These include amendments, that bring into the ambit of the German Competition Law, non-price offerings (such as search engines). In particular, the German law now clarifies that transactions where no monetary consideration is paid also

66 https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf
67 https://ec.europa.eu/info/sites/info/files/communication-shaping-europes-digital-future-feb2020_en_3.pdf
68 https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
69 https://www.economist.com/special-report/2020/02/20/governments-are-erecting-borders-for-data
70 https://www.economist.com/special-report/2020/02/20/who-will-benefit-most-from-the-data-economy
71 https://issuu.com/radicalxchange/docs/data_legislation_paper_--_20191031

constitute a market and can fall within the scope of competition law. Moreover, aspects that are critical for the market power of platforms and networks (such as network effects and access to data) have been introduced into the law as new criteria for market power.

vi. Legislations in other countries too have allowed access to data for safeguarding national security[72,73].

2. Other countries have put in place systems and mechanisms for data sharing. An example is the Japan's Certification System for data-sharing platforms that support companies that want to share their data (on energy, industrial machine and logistics to solve social problems like accident prevention, energy management etc.).
   i. This system includes a data request system, i.e. a system that allows data-sharing companies to request data that have been provided to relevant ministries and agencies.
   ii. The government also provides support through tax incentives and administrative guidance. It can also revoke accreditation in some cases.

3. Another example is that of the Government of Victoria in Australia, which has put in place a Data Exchange Framework for Government and third party data exchange[74]. The data exchange model consists of the following steps:
   i. Manage data requests, assess readiness and authority to exchange – ensuring the exchange (or transfer) happens in a secure, transparent and compliant manner and sufficiently describing the data and its quality to enable the data recipient to assess fitness for their intended purpose.
   ii. Apply business rules to ensure reliable, consistent and sustainable data exchange and decision making
   iii. Identity mechanisms and tools – which tools and templates to use will support streamlined, safe and authorised data exchange
   iv. Exchange data

4. Finland's (2018) Act on Transport Services through deregulation gives more room to develop innovative, digitally enabled services. It obliges all service providers to open certain essential data to all as well as ticketing and payment APIs for single

72 Paul F Scott, "National Security, Data Protection, and Data Sharing after the Data Protection Act 2018", University of Glasgow
73 Louis de Koker et al., "Big Data Technology and National Security", Data to Decisions Cooperative Research Centre, 2018
74 https://www.vic.gov.au/sites/default/files/2019-07/Data-Exchange-Framework_0.pdf

trip/ticket to third parties. The Act makes it possible to examine transport as a whole, i.e. as one service[75].

Data Sharing Mechanisms and Frameworks

5. Types of data sharing
   i. Government data sharing (G2B and G2C): Sharing of public information by the Government for the purposes for re-use by organisations (including companies and startups) and individuals alike.
      - Regulatory examples: Directive (EU) 2019/1024, on open data and the re-use of public sector information
   ii. Private / Industrial data sharing (B2B): Sharing of industrial data between organisations involved in the same commercial or non-commercial point of the value chain.
      - Examples: International Data Spaces (IDS) Association, Industrial Internet Consortium(IIC), Data Market Austria, Ocean Protocol and the IOTA Foundation[76,77,78,79]
   iii. Open data sharing: Sharing of industrial data inside or outside of a value network, Government public information and the data of willing participants shared in their individual or collective capacity through sharing mechanisms/instrument.
      - With respect to industrial data, when such data is legally open, it means that the data is published under an open license and that the conditions for re-use are limited to attribution. Second, the data is technically open, which means that the file is machine readable and non-proprietary.
      - Regulatory examples: Australian Data Sharing and Release bill, 2018

6. Data sharing mechanisms
   i. Government data sharing
      - Data sharing framework: Building on the framework created by National Data Sharing & Accessibility Policy (NDSAP)[80], the default practice should be proactive release of data upon request generated through the Open Data Portal.
   ii. Private / Industrial data sharing (B2B)

75 https://www.oecd-ilibrary.org/sites/276aaca8-en/1/2/5/index.html?itemId=/content/publication/276aaca8-en&_csp_=a1e9fa54d39998ecc1d83f19b8b0fc34&itemIGO=oecd&itemContentType=book
76 https://www.internationaldataspaces.org/
77 https://www.iiconsortium.org/
78 https://datamarket.at/
79 https://oceanprotocol.com/
80 https://data.gov.in/sites/default/files/NDSAP.pdf

- Data monetisation: unilateral approach under which companies make additional revenues from the data they share with other companies. Data can also be monetised through the provision of services.
- Data marketplaces[81]: trusted intermediaries that bring data suppliers and data users together to exchange data in a secure online platform. These businesses make revenue from the data transactions occurring in the platform.
- Industrial data platforms: collaborative and strategic approach to exchange data among a restricted group of companies and/or startups. They voluntarily join these closed, secure and exclusive environments to foster the development of new products/services and/or to improve their internal efficiency. Data may be shared for free, but fees may also be considered.

7. The Government may have to play a role in incentivising and orchestrating data partnerships, either by acting as independent trusted third parties or by engaging with the private sector in Public-Private-Partnership (PPP) mode. This is achieved through appropriate rules and regulations.
   i. For example, the European Commission is examining data sharing between the private and public sector in order to guide policy making and improve public services.
      - Manufacturers of IoT [Internet of Things] objects usually determine access to the non-personal and automatically generated data from IoT objects, which have been triggered by the data-users.

8. In Europe we can see examples like the Finnish Health and Social Data Permit Authority[82], French Health Data Hub[83], European Open Science Cloud[84] that allows Europe's 1.7 million researchers and 70 million science and technology professionals a virtual environment to store, share and re-use the large volumes of information generated by the big data revolution.

9. There exist frameworks[85] that examine the opportunities of enhancing access to and sharing of data. They highlight the factors that need to be considered including data typologies, key data-access mechanisms and the main types of actors and their roles.

81 http://www.bdva.eu/sites/default/files/BDVA%20DataSharingSpace%20PositionPaper_April2019_V1.pdf
82 https://www.findata.fi/en/
83 https://www.health-data-hub.fr/
84 EOSC Strategic Implementation Roadmap 2018-2020, https://ec.europa.eu/research/openscience/pdf/eosc_strategic_implementation_roadmap_short.pdf#view=fit&pagemode=none
85 https://www.oecd-ilibrary.org/sites/b4d546a9-en/index.html?itemId=/content/component/b4d546a9-en&mimeType=text/html

Appendix 7 – Frameworks for Community Data Rights

1. When existing frameworks for allocating rights to various resources are found inadequate or inappropriate, as appears to be the case with data, sui generis (meaning; fully new, or 'one of its own kind') frameworks have been devised worldwide.
   i. Requirement for a sui generis framework may arise due to (1) the very nature of the resource involved, (2) the extant social and business processes around the resource, and/or, most importantly, (3) social impacts of a resource and society's requirements in this regard.

2. This has been undertaken, for instance, regarding traditional knowledge, plant varieties, genetic resources of flora/fauna, and in providing special database rights in the EU going beyond established IP principles. Some of these sui generis frameworks provide individual rights (plant varieties, EU database directive) while others establish collective or community rights to various resources (traditional knowledge, genetic resources of flora/fauna).
   i. The international Convention on Biological Diversity (CBD) takes up governance of genetic resources of flora and fauna associated with a nation and its communities. Closer to traditional knowledge in being collectively associated with specific communities, genetic resources are even more clearly not subject to IP frameworks. Occurring naturally, it is impossible to consider them intellectual products or 'creations of mind' – whether individual or collective. Nagoya Protocol of the CBD employs sui generis formulations to institute national and community rights over relevant genetic resources. Such rights, on one hand, evolve over traditional knowledge framework, the involved flora and fauna being closely associated with a nation/community's cultural practices. On the other hand, such rights connect to a nation's ownership over its natural resources, and its right to manage them for its people's best benefit.[86]
   ii. Giving effect to the framework of global CBD, India's Biological Diversity Act provides communities rights to benefit from, and participate in governance of, a community's genetic resources, including through Biodiversity Management Committees at local levels.
   iii. India's Mine and Minerals Act has set up District Mineral Foundations as non-profit trusts for ensuring 'benefit sharing' and participation in governance for communities associated with mineral resources. India's Forest Act provides

86 Common Article 1 of the International Covenant on Civil and Political Rights and the Covenant on Economic, Social and Cultural Rights.

community forest rights to relevant communities for accessing and using various forest resources,87 and also of participation in governance of forest resources.

iv. India's Traditional Knowledge Digital Library is a database containing 34 million pages of formatted information on some 2,260,000 medicinal formulations in multiple languages.[88] It serves to pre-empt IP claims over and prevent misuse of India's Traditional Knowledge. India's Biological Diversity Act also provides for People's Bio Diversity Registers, that document knowledge of local biological resources, for a similar purpose. Such open documentation of community resources has a defensive function to prevent wrongful appropriation and misuse. They also have a positive purpose to enable recognition of local resources, and legitimate access to them, with prior informed consent and due benefit sharing.

3. With communities having community rights over their genetic data, it would be logical to argue that they also have economic rights to their other kinds of collective health data, be it of an anatomical, physiological or behavioural nature. This flows naturally into – what really is a continuum of community's common data – consideration of a community's economic rights to data about its collective social, cultural and economic behaviour. Most of the data collected by digital platforms is in fact data about a community's social and economic behaviour, which happens to be the most valuable of data. An examination of the evolving thinking and frameworks on governance of community resources worldwide brings forth the question: Why should a nation and community not 'own' – in the sense of having primary economic rights to – its common or community social and economic data? Developing countries,[89] including India,[90] have claimed that their data is akin to their natural resources, and they should own it. To be more accurate, one can call collective data as 'social resources' instead of, or a subset of, 'natural resources'. Both are a nation's or community's collective resources as arising from their natural and/or social spaces, and should be governed as such. New frameworks of data governance aim to not reassert individual control over the terms of one's own datafication or to maximize personal gain, but instead to develop the institutional responses necessary to represent the relevant population-level interests at stake in data production[91].

---

87 These rights are of usufruct kind, meaning the right to use and enjoy the benefits and fruits of forest resources without depleting the overall resource system. This looks quite akin to managing an intangible resource commons.

88 https://www.wipo.int/wipo_magazine/en/2011/03/article_0002.html#:~:text=India's%20TKDL%20is%20a%20unique,protecting%20the%20country's%20traditional%20knowledge.&text=1%20Prior%20art%20constitutes%20all,claim%20of%20novelty%20and%20inventiveness.

89 Submission to the WTO by South Africa on the behalf of the African Group.

90 India's draft e-commerce policy.

[91] Viljoen, Salome, Democratic Data: A Relational Theory For Data Governance (November 11, 2020). Available at SSRN: https://ssrn.com/abstract=3727562 or http://dx.doi.org/10.2139/ssrn.3727562

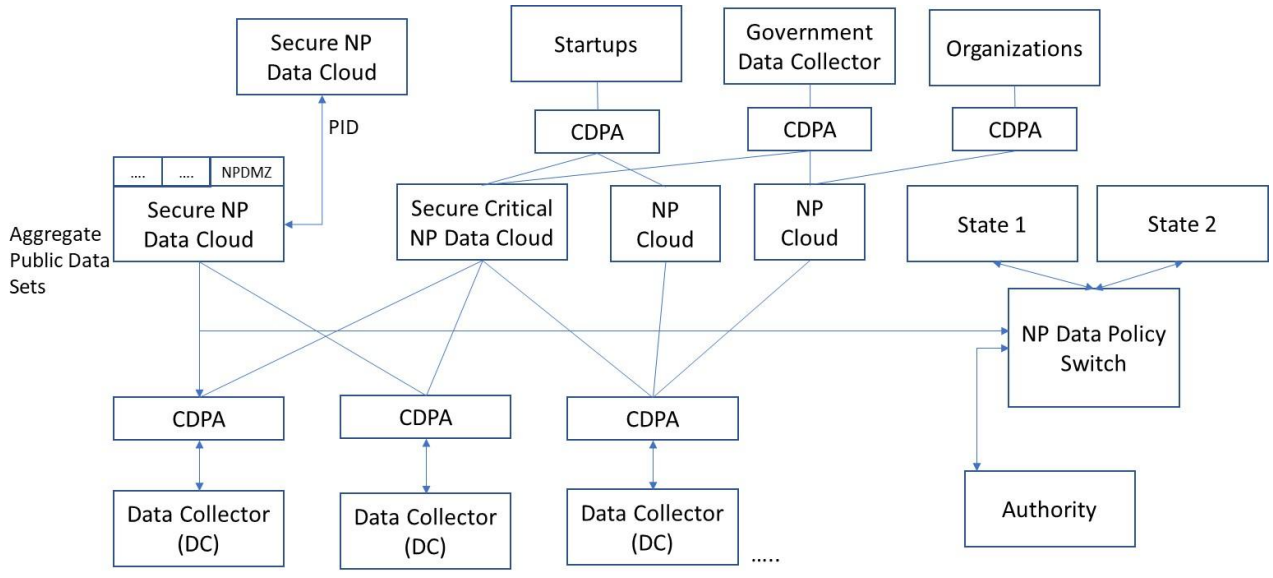
4. Existing community resource frameworks offer the following five guiding principles; (1) a community's right over resources associated collectively with it, (2) prior informed consent of the community for use of such resources, (3) benefit sharing with the community, (4) transparency is recording community resources to prevent misuse and enable easy access of the legitimate kind, and (5) community's participation in governance of community resources, including through non-profit trusts.

5. Application of eminent domain alone as the basis for a data sharing law, without framing appropriate background principles and norms, is also problematic in a digital economy context where data has considerable global aspects. It will finally be required that some kinds of data sharing principles get recognized as global norms. For this reason, too, it is useful to develop adequate theoretical, jurisprudential and legal bases – that also involves existing and evolving global principles, frameworks and precedents – as part of a sui generis framework for data sharing. Coming out of India, which has taken a lead in this respect, such a framework can in time help shape relevant international norms, agreements and practices for data sharing.

59

Appendix 8: Illustrative Technology Architecture for Data Sharing

The Committee presents an illustrative three-tiered system architecture spanning legal safeguards, technology and compliance to enable data sharing.



NPDMZ – Non-personal Data Demilitarized Zone
PID – Processed Insights Derivatives
CDPA – Certified Data Processor Algorithm

Figure: Architecture of different stakeholders, data and control flow.

1. A technology architecture that enables data sharing.
   i. Data Business / Data trustees may implement this architecture when they are faced with a data request.
   ii. Best of breed Differential Privacy algorithms [Refer to 'Appendix 4: Primer on Anonymity' in this report for different algorithms] are used to create anonymized data to best effort by the data custodians and in compliance with rules set by the regulator.
      - These best of breed Differential Privacy algorithms should be jointly evolved by Indian academia and industry, continuously improved using a combination of global open source improvements and with funding to Indian research organisations.
      - These algorithms along with their open-source implementations are made available to Indian organizations along with minimum recommendations for each major type of data.

- These recommendations may be cemented and continuously evolved by leading technical experts using an open standards-based IETF process, perhaps making these global standards as well through IEEE and WWW.

iii. The data sets so anonymized are then submitted or when real-time, streamed into "Secure non-personal data Clouds".

2. Policy Switch

i. Each data trustee may want to exercise its authority to govern data deemed in their respective domains. However, the best innovation happens in the boundaries and interconnections between datasets - traditionally separated by such governance functions. This can significantly reduce economic value realisation and stifle innovation if each data trustee creates a separate window of clearance and rules for using data under their regulation. A new approach is suggested to address this aspect – of a digital non-personal data Policy Switch ("Policy Switch") as defined below.

ii. Using the Policy Switch, even though regulations can emerge from various institutions and regulatory bodies, the encoding, rationalisation (to ensure no contradiction), implementation and clearance/ compliance enforcement may be with a single authority - who is subject to the regulatory guidelines issued by various data trustees.

- And since handling data subject to multiple regulatory bodies can get complex exponentially, a way to efficiently and rapidly realise economic benefit and large scale public good of non-personal data without sacrificing regulatory granularity or diluting individual authorities is to bring these together digitally.

iii. The central idea of the Policy Switch is a single digital clearing house for regulatory management of non-personal data. The Policy Switch is defined by a set of APIs and a Policy Markup Language spanning all aspects of managing non-personal data publicly and privately. The Policy Markup Language encodes all interactions and transactions relevant to non-personal data spanning:

- Policies: e.g. access rules, anonymization standards, aggregation standards, business rules, security standards
- Adjudication workflows: e.g. verification, exception adjudication, certification
- Compliance: e.g. registration, compliance submissions, that are applicable to non-personal data such that non-personal data custodians, both public and private, only have to interface with and comply with the Policy Switch digitally, no matter the types or sources of data with which they are engaged.

- To reduce the burden on various governance authorities, the Non-Personal Data Authority will create a base set of minimum set of policies, workflows and compliance rules with which all non-personal data must comply.
- In addition, it is recommended that the Non-Personal Data Authority manage a stream of academic research and grand challenges to create reference policies, evolve the markup language and reusable tools to simplify the management of non-personal data by regulators, data custodians and data processors.
- A further suggestion is to design this policy markup language to be evolutionary. For example, rules, often stated as principles and guidelines, rarely spell out every corner case. A well-implemented policy switch will continuously capture corner cases that emerge via built-in adjudication workflows and after verification, update the marked-up policies so that corner cases are captured in definitions as whitelists or blacklists; and as conditional exceptions in the rule hierarchy.