Mozilla Corporation
331 East Evelyn Avenue
Mountain View, CA 94041

Mozilla's Submission to the Public Consultation on the 'Report of the Committee of Experts on Non-Personal Data Governance Framework' in India

To

Ministry of Electronics and Information Technology (MeitY),
Government of India,
Electronics Niketan, 6, CGO Complex,
Lodhi Road, New Delhi - 110003.

10 September 2020

We thank the Ministry of Electronics and Information Technology (MeitY) for the opportunity to provide feedback on the Report by the Committee of Experts on Non-Personal Data Governance Framework (hereafter, "the report" or "report"). We welcome the move to have a broad consultation by inviting suggestions from experts, stakeholders, and the general public and hope this approach is followed for future consultations by the Ministry.

Mozilla is a global community working together to build a better internet, with openness at the core of its functioning. As a mission-driven technology company, we are dedicated to promoting innovation and opportunity online. As our Mozilla and the Rebel Alliance report highlights, there are over 22,000 contributors in over 49 projects managed by Mozilla. We are the creators of Firefox, an open source browser and the family of Firefox products, including Firefox Focus and Firefox Lite, as well as Pocket, used by hundreds of millions of internet users globally. Mozilla's commitment to user security and privacy is evident not just in our products but also in our policies and in the open source code of our products.

In this submission below, we offer analysis and recommendations based on our experience and expertise advocating for data protection, competition, security, and privacy all over the world. In particular, we offer our views on:

- the concept of data as a "national resource";
- forced data transfers;
- restrictions on storage and sharing of data;

Mozilla Corporation
331 East Evelyn Avenue
Mountain View, CA 94041

- nebulous definition of community data and its governance; and
- access to data as a means to encourage competition;

The two main priority areas identified in the Report are increasing the competitiveness of domestic industries at the expense of foreign companies and "unlocking the economic value of data" for Indian communities. However, rather than viewing this as a zero-sum game, there is much for India to gain by leveraging the interconnectedness of the global digital economy while respecting the fact that privacy is a fundamental right guaranteed to all, including groups and communities.

However, some of the blunt strategies proposed in this Report could harm Indians, isolate Indian companies from their global counterparts, and cause other countries to retaliate with similar "data nationalisation" measures that would be counterproductive to India' interests. The digital economy is characterized by diverse partnerships and business relationships to deliver the best services to the user generally at the lowest cost. This interconnectedness is especially important for small and growing businesses who depend on outsourcing a range of functions to global companies in order to keep costs sustainable as they scale. Ultimately, a maximalist focus on boosting domestic industry could hurt the very businesses it is meant to serve, while limiting competition, and diminishing the choices of users. These impacts would be magnified if another country were to enact a similar regime on Indian firms.

While many of the strategies proposed in the Report focus on enabling access to data for the Indian government and businesses, ensuring the privacy and security of this data is merely noted as an afterthought in most instances. Given that India currently lacks a comprehensive data protection law (which the report recognises) and India has some of the weakest regulations around government surveillance in the world, such permissive access to Indians' data does not adequately protect Indian users. These legal safeguards are a precondition to any credible data-focussed digital regulation. As we've argued extensively to the MeitY and the Justice Srikrishna Committee, such a law has the opportunity to build on the globally high standard of data protection set by Europe, and position India as a leader in internet regulation. To this end, we urge the government to prioritize the passage of a strong data protection law, accompanied by reform of government surveillance. Only after the implementation of such a law makes the fundamental right of privacy a reality available to all Indians to all should we begin to look into non-personal data.

Mozilla Corporation
331 East Evelyn Avenue
Mountain View, CA 94041

We recommend that the government reconsider the idea of creating a non-personal data governance framework at this time and focus on implementing an effective data protection framework instead. We look forward to continuing to engage with you and other stakeholders in the Government of India on formulating India's framework for regulating non-personal data. If you have any questions about our submission or if we can provide any additional information that would be helpful as you continue your important work, please do not hesitate to contact Mozilla's Policy Advisor, Udbhav Tiwari, at udbhav@mozilla.com

Warm regards,

Udbhav Tiwari
Public Policy Advisor
Mozilla Corporation

Mozilla Corporation
331 East Evelyn Avenue
Mountain View, CA 94041

Table of Contents

1. Treating data as a national resource undermines individual autonomy

The Report, in different stages, states that the data of Indians is a "national resource" and that rights surrounding it should be governed similar to "economic rights over natural resources." We object to this framing at multiple levels as it paves the way for regulations that undermine individual rights and is in stark contrast to the expectations of Indian users. Moreover, the report, in multiple instances, conflates the government's interests with society's interests, which is a dangerous assumption for the world's largest democracy.

This framing, founded on a flawed model of data ownership, undermines the Supreme Court of India's decision in Puttaswamy v Union of India as well as India's commitments under the International Covenant on Civil and Political Rights. The Puttaswamy judgment held in no uncertain terms that the fundamental right to privacy was "an intrinsic part of the right to life and liberty", predicated on the dignity and autonomy of every individual. To replace this fundamental right with a notion of ownership akin to property, vested in the individual but easily divested by state and non-state actors, leaves individual autonomy in a precarious position. How can one have individual autonomy when one's privacy may be violated by virtue of being a member of a community? The goal of data-driven innovation oriented towards societal benefit is a valuable one. However, any community-oriented data models must be predicated on a legal framework that secures the individual's rights to her data, as affirmed by the Constitution.

The report recognises that "many actors may have simultaneous overlapping rights and privileges" over non-personal data. It attempts to create the legal fiction of "beneficial

ownership/interest" to resolve such concerns without once clearly describing what this term means, how it would operate and how conflicts between multiple claims of such interest would be resolved. This lack of clarity would make such a regulatory framework ripe for exploitation due to insufficient digital literacy and insufficient data protection provisions such as effective accountability, independent oversight, and redressal mechanisms.

2. Forced data transfers are not a solution to the concentration of market power

The Report aims to create the "certainty and incentives for innovation and new products / services creation in India" through requirements that these large market entities share data with smaller firms entering the market. Data driven innovation can unlock great benefits to individuals, businesses, and society. In this spirit, Mozilla, recently released "Common Voice", the largest open dataset of labelled human voice data available for use by start-ups, researchers, and the general public. Although many ML algorithms are in the public domain, training data is not: Most of the voice data used by large corporations is not available to the majority of people, expensive to obtain, or simply non-existent for many languages. For example, the most widely used voice assistant in India (Google Assistant) is only available in 9 languages, despite India having over 22 languages recognized by its Constitution. The innovative potential of this technology is widely untapped. With providing open datasets, Mozilla hopes to take away the onerous tasks of collecting and annotating data. It reduces the main barriers to voice-based technologies and creates a potential market for tech innovators and social entrepreneurs, such as weather and crop information for agriculture, health information in local languages, etc.

However, non-personal data can also constitute protected trade secrets and the insights derived from such data may be protected by intellectual property law, both of which would raise significant concerns around the fundamental right to carry out business and India's obligations under international trade law. Turning over this information to the government or private entities without any checks and balances also raises significant privacy concerns. Information about sales location data from e-commerce platforms, for example, can be used to draw dangerous inferences and patterns regarding caste, religion, and sexuality.

The focus on increasing the competitiveness of the digital economy is both necessary and timely, given the global concentration of market power in a handful of firms. In some cases, the Government could promote competition in the digital economy by supporting dominant firms to grant potential competitors access to privately held aggregate data on reasonable market terms. However, we would urge the government to approach this strategically, and

encourage this data sharing through a targeted incentive-driven framework, rather than imposing blanket coercive measures that will alienate global companies and likely raise legal challenges. In particular, on this point, we recommend:

- Any approach to enable data sharing will have to ensure that robust privacy safeguards are in place. A data protection law must be in place which would ensure that all personal data in the current draft Data Protection Bill (2019) is excluded from any data sharing. As noted earlier, any such policies would need to mitigate the risks of inadvertent re-identification of individuals through combining apparently anonymized data points.

- Rather than blanket coercive measures to open up datasets, the government should conduct an exercise to identify those aggregate and anonymised datasets that would be most valuable to new, nascent businesses. As noted in the UK Governments' recent Digital Competition Expert Panel Report, there are already a number of positive examples of such voluntary data sharing. For instance, Uber has chosen to release anonymised and aggregated data under the 'Uber Movement' scheme to inform and improve infrastructure and planning decisions in India and other countries by collaborating with urban planning government agencies.

- Rather than creating another regulator, who will have conflicting responsibilities with a hopefully forthcoming Data Protection Authority (DPA) and the Competition Commission of India (CCI), both these agencies should be reformed and empowered to more effectively carry out their mandates in regards to the issues raised by this Report.

- There are other technical measures beyond data sharing which could facilitate greater competition in addition to traditional legal methods. Specifically, as we have stated earlier, competition policy should encourage designing for interoperability and standards-centric design and implementation. This could include coupling positive incentive "carrots", including potential safe harbours with corresponding "sticks" of heightened merger review standards; and strengthened enforcement of rules and policies against anti-competitive behaviour by firms.

The report recognises that innovation has occurred due to "robust IP rights, various data related privileges" and that "requirement for providing certainty and incentives for new business creation" is paramount for a country like India. Data collection and processing, when

done lawfully under strong data protection regimes, requires significant investment of skill, time and resources. Forced data transfers of such data, as envisaged in this report, are against such new business creation and must be reassessed in order to be privacy respecting, voluntary, and market driven rather than enforced via data nationalisation.

3. **Restrictions on storage, sharing, and cross-border flow of data will harm India and its interests.**

The report proposes a series of broadly worded strategies targeting the data of Indians held outside India. These range from requirements for local storage of data on servers in India, to multiple restrictions on sharing of data with foreign companies and governments, to requirements to provide adequate notice for data collection. In contrast, the report states that the Government of India should have broad access to the data of Indians, requiring that any entity that stores data (either in India or abroad) must comply with any request for non-personal data from Indian authorities, start-ups, companies, trusts or NGOs.

It is unclear what objective the report seeks to achieve with this medley of strategies. At certain points, the report notes that these restrictions might ensure greater privacy and security for the data of Indians. This objective, however, would be better served by a strong data protection law accompanied by meaningful surveillance reform. MeitY's Draft Personal Data Protection Bill makes a strong start. With certain key amendments, such as strengthening the regulation of data processing by government and surveillance agencies, and bolstering the independence of the data protection authority, we believe that the passage of the data protection law could be a major step forward towards protecting the privacy of Indians.

Without these reforms, localizing data in servers in India is more likely to make data more susceptible to overbroad access by law enforcement and surveillance agencies. As the Justice Srikrishna Committee has noted, surveillance reform in India is much overdue and currently there are little to no procedural safeguards around government surveillance. While surveillance may be a legitimate function of the state, these are powerful tools in the hands of government agents and should be subject to accountability, transparency, due process, and meaningful limits. This follows from the Supreme Court's diktat in Puttaswamy v. Union of India where a unanimous verdict held that any state interference with privacy must be subject to tests of legality and strict proportionality. Given that the current legal framework falls short of these standards, a broad localisation mandate is likely to mean even fewer checks on the power of the government to access personal data. For example, would we want official functionaries knowing which housing colonies in a city have a propensity to buy pro-LGBTQIA+

posters, books about a specific caste, or religious objects? This data may not contain any identifying information about a specific individual but can be used by malicious actors with disastrous consequences without sufficient checks and balances.

Moreover, storing a copy of all personal data pertaining to Indians in a handful of locations could create a "honey pot" for malicious actors, thereby increasing the risk of a breach with a profound effect on India's citizenry. In comparison, distributing storage of data across a network of servers globally means that there is no concentrated point of attack or single point of failure. Finally, while large companies could more easily afford the additional costs associated with this data localization mandate, the expense of compliance may prove disproportionately harmful to small businesses and start-ups.

There is also a meaningful distinction to be made between mandatory data localization regimes -- which require data to be stored within the borders of a given country unconditionally -- and some of the limitations on data transfer found in data protection laws, including the GDPR and MEITY's Personal Data Protection Bill. These laws provide for a variety of options, including standard contractual clauses, binding corporate rules, and determination of adequacy for countries or sectors.

4. Re-identification poses a grievous risk to privacy and security

The report also notes the value of "anonymized and aggregated data" towards creating "data trusts". We acknowledge and support the release of more open datasets that might spur innovation in India. However, we would also warn that research has consistently shown that there are serious risks of re-identification even with apparently anonymized datasets. Paul Ohm's seminal paper concluded that "Data can be either useful or perfectly anonymous but never both." A study by Latanya Sweeney found that 87.1% of people in the United States were uniquely identified by their combined five-digit ZIP code, birthdate, and sex. Another study re-identified data subjects based purely on their movie preferences on Netflix. In light of these risks, we would urge the government not to make the blanket assumption that the public or private release of datasets is an acceptable risk.

Consent for Non-Personal Data: Anonymization is a privacy respecting technique and custodians should be permitted to anonymize data without the need for obtaining any additional consent. The report states that "Personal Data that is anonymized should continue to be treated as the Non-Personal Data of the data principal." This is inherently contradictory, as anonymisation should make it legally and practically infeasible to be able to distinguish

between one data principal and the other. Such a move may force data custodians to not anonymise data sufficiently to be able to track such consent, placing such data at an additional privacy and security risk. Alternatively, it would require the re-identification of the principal in an anonymized dataset. Doing either of these things would defeat the whole purpose of the anonymization in the first place. This ambiguity should be clarified to prevent data subjects from being harmed due to unclear legislative drafting.

Even the Report recognises, in the definition of non-personal data section, that "no anonymization technique provides perfect irreversibility. Such harm may be individual or collective as a group or community (whereby harm can happen even without de-anonymisation)." While the report does mention the importance of privacy, it doesn't actually provide any detail into what these protections can or should be. It makes a passing mention of re-identification being criminalised in the draft data protection bill. It does not, however, elucidate how these protections would be applied to or be enforced for non-personal data nor how much needed exceptions for legitimate security research would interface with such a provision. Without these protections, there is nothing to prevent the government or any business (foreign or domestic) from exploiting non-personal data in ways that contravene the autonomy and dignity of the individual in question. Therefore, any regulatory framework for non-personal data needs to be more focused on how such data will be protected meaningfully, rather than focusing on how it should be exploited for national interest.

5. Non-personal data is not an immutable construct

The committee has chosen to define non-personal data by exclusion, where everything that is not personal data in the data protection bill is non-personal data. While seeming reasonable at first glance, such an outlook ignores that vast scholarship that points to the flaws of such an approach that creates a binary between personal and non-personal data. Paul Ohm, as we have mentioned earlier, said that the moniker of 'identifiable information' is a misnomer and needs reframing. He states this based on the premise that modern re-identification methods make the classification irrelevant as such information can be linked back to individuals, often with the help of alternative data sources. Others (such as Schwartz and Solove) ask for a more fluid definition that grants differing standards for protection according to the context of use/collection alongside the possibility of re-identification. As already stated by Indian civil society, the PDP Bill also uses similar standards of 'identifiability' to denote the scope of its regulation, implying that 'personal data' and non-personal data' are not 'fixed or immutable categories.'

Given this context, the report goes on to create three "new categories" of non-personal data. Namely, public non-personal data, community non-personal data, and private non-personal data. We think that this classification, while attempting to be comprehensive, is reductive and removes the nuance about data sets in the real world. The classifications are neither mutually exclusive to each other nor do they provide clarity in how they would operate in practice. For example, an environmental sciences startup that has received partial government funding to map national parks, under the current classification, will be collecting all three kinds of non-personal data at the same time.

We recommend that this classification be completely redone, with a comprehensive public consultation that uses evidence based research to create a new classification. This will go a long way in helping protect user rights while providing the predictability that regulators and businesses will need for effective enforcement.

6. Community data is a nebulous idea and needs clarity

The definition of "community data" is vague and will lead to significant challenges in compliance and implementation. A community is defined 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."

This definition is incredibly wide ranging and is ill suited to a framework that is being designed to enforce rights and protect the interests of its constituents. Under this classification, religious groups; people from the same educational institutions; vulnerable communities based on class, caste, and economic criteria; and people who once lived in a residential locality, are all valid communities with enforceable data rights. They can all have conflicting interests over data that they may have shared with government and private platforms. For example, a housing society that wants to raze neighbouring trees to build a new road and an opposing group of environmental activists from the same society could ask for both aggregate tree cover data from a mapping provider. With overlapping members, the report doesn't provide the criteria by which the mapping provider should choose between them as the representative 'community' to respond to in this case.

Without a guiding legal framework or principles, which is absent in the report, such a model will have a crippling effect on service providers from a compliance perspective. They will be forced to make legally binding decisions on what is a valid community, what is the scope of

data that can or cannot be shared with such communities, and how to resolve disputes between competing claims to represent a community's interest. The scope of exclusion and discrimination, which many communities already suffer from, will only increase with such a model.

The report is also silent on how such communities should govern themselves when it comes to data governance. It vaguely states that the "corresponding government entity or community body" can be the valid data trustee for a community, which is deeply concerning. Extrapolating from examples present in the report, government departments and agencies seem like the de facto choice for the role of the data trustee in many instances. This points to a direction where rather than communities governing themselves and their data, the government or government appointed bodies would govern their interactions with their data. We should be vary of such considerations, which will just entrench state power and open it up to exploitation of interests close to the state. It's also hard to reconcile this governance structure with the notion of autonomy.

The very idea of "community data" should be revisited and alternative means should be explored to achieve the intended outcomes without the current drawbacks of lack of clarity, conflicting interests, and difficulties of implementation.

7. Data trusts need to be explored more rigorously

The report defers most questions on how data subjects and data custodians will interface with each other via data trusts and data trustees. While it talks about the need for a regulatory framework for such entities, it doesn't lay out any detailed criteria for how they would operate in practice.

Mozilla is currently looking into different types of data governance models, such as trusts, as we believe this concept may hold promise. However, there are a range of challenges and complexities associated with the concept that will require careful navigation in order for trusts to meaningfully improve the state of data management and to achieve a truly ethical and trustworthy data ecosystem.

A couple of examples to provide further context around the kinds of approaches we are looking into are the following:

- **Data Trusts:** Similar to a land trust, a data trust is an independent intermediary between two parties: the people creating data ("data subjects") and the companies collecting that data ("data collectors"). The trust would have a fiduciary duty towards its members, and would negotiate data use with
- companies according to terms set by the trust. Different trusts might have different terms, and people would have the freedom to choose the trust that most aligns with their own expectations. Some data trusts already exist: for instance, UK Biobank, a charitable company with trustees, is managing genetic data from half a million people.

- **Data Cooperatives:** Another proposed approach is the data cooperative model. U.S. citizens first organized themselves into trade unions and credit unions as a counterweight to powerful companies and banks like Standard Oil and J.P. Morgan in the mid-19th century. The result was a complete rebalance of power between large and small players in the economy. Similar to a credit union, a data cooperative would have a fiduciary duty to manage and protect access to the personal data of its members. Because it could also run internal analytics about its members, the data co-op would also be in a strong position to negotiate better services for its members. Some data co-ops already exist: The MIT Trust Data Consortium has demonstrated a pilot version of this system.

While such data intermediary models and related data governance approaches have the potential to provide the guardrails within which ethical and trustworthy data solutions, they have potential to aggravate, or create new and complex challenges of their own. Considerable work will need to be done to ensure these intermediaries do not duplicate the systemic problems that already exist today. For instance:

- **Security:** To the extent that data trusts act as custodians of data, this will require impeccable levels of security and data management. There is a major risk that centralizing data in a data trust could broaden the attack surface for abuse and misuse, amplifying the negative impact on the individuals and the collective. Major tech companies today, in part because their business models depend upon users giving them data, are able to dedicate significant resources and expertise to protecting that data from compromise. Yet, even this is not enough in all instances, -- even the largest, most resourced companies have suffered data leaks and breaches -- and thus security must go hand in hand with lean data practices to ensure that risks are not bigger than they need to be. At the end of the day, data collected is data at risk.

- **Legal:** There are a range of legal complexities to consider, from cross-jurisdictional challenges to consent models. On the former, and
  given the global nature of the economy and society, how would trusts legally interact, particularly in the case that some jurisdictions may rely on individual rights, while others on the collective? On the latter, it cannot be taken for granted that the challenges with data literacy and the ability of users to truly understand how their data is being managed wouldn't manifest in a data trust model as well.

- **Trust:** If rights can be assigned to an entity, how can it be ensured they will behave in the best interest of the individual/collective? Put another way, if there are scruples with current data managers, how can it be ensured that the next generation's data managers will behave ethically and in the best interests of the individual? While the intellectual concept of data trusts holds a lot of promise, and is a useful exercise to explore how data can be managed in more ethical and trustworthy ways, the devil is always in the details, which are surprisingly absent in this report.

Conclusion

As we have highlighted in our submission, many aspects of the report merit a ground up reimagining of the entire construct of non-personal data and its regulation. We recommend that the government reconsider the idea of creating a non-personal data governance framework at this time and focus on implementing an effective data protection framework instead. Once India has an effective data protection law, the process for creating a non-personal data regulation should be started afresh with much greater input from civil society and impacted communities.