#NAMA: Is defining non-personal data possible? Is anonymising it a good idea?

The first issue with defining non-personal data as the negative of personal data is that personal data is not in a crystallised form in the Personal Data Protection Bill, 2018, said a speaker at MediaNama's roundtable discussion on non-personal data. The second issue is that segregation of personal and non-personal data is close to impossible as there are mixed data sets in the age of digital markets, IoT and Big data. Citing the Personal Data Protection Bill, 2018, another participant said that non-personal data is "100% of the pie with the personal data removed from it", as participants deliberated whether or not non-personal data was a viable category to begin with.

(Note: The discussion was held under the Chatham House Rule; quotes have not been attributed to specific people. Quotes are not verbatim and have been edited for clarity and to preserve anonymity. Also note that this discussion took place before the PDP Bill, 2019, was made public.)

What are the problems with defining Non-Personal Data?

All data can be traced back to a granular level: "In the context of GDPR, the moment you put the clause 'relating to humans', everything becomes personal data. From that definition, 100% of data is either directly or indirectly linked to human beings," a speaker said. The issue arises because the Personal Data Protection Bill, 2018, did not define 'personal data' adequately, as per the speaker. To this, another speaker said that there was a clear distinction between 'human data' and 'personal data'. "I don't think we should use the two interchangeably. Human data is anything that is generated by human activities. Personal data is identifiable to individuals," they clarified. "So when you're talking about whether a community or a region is producing a certain particular effect, it's not the same as saying that it is personal data. We can't use it so interchangeably." However, they conceded that "all kinds of data sharing which related to human activity can eventually be brought down to granular level of individual human beings."

We haven't seen non-personal data in practice: "Much as we try to distinguish between personal and non-personal data, we are not going to be able to because we don't know what non-personal data look like in practice at scale. This is similar to the discussions we had in 2017 when we tried to define localisation. … I think it's best to sort of accept that non-personal data exists and just move on because it is a bounded rationality problem," a speaker explained. Despite acknowledging the problems with defining non-personal data, another participant said, "Non-personal data is a very real thing right now because there is a committee that is deliberating [on it] and a lot of companies and organisations have already presented before it."

Source of data unknown: "We don't know where it emanates from, whose

data it is, who is the producer of the data. When I take a cab, am I the one producing data or a cab aggregator?" a speaker pointed out.

Non-personal data as a concept is continuously evolving: The previous speaker continued, "The definition of non-personal data needs to be about limits because arriving at a good definition of non-personal data is going to be tricky and it will constantly be evolving in some ways. So is true for regulation. It needs to be about limits and not about an exact definition."

Can anonymous data always be traced back to the source?

A number of speakers pointed out that **non-personal data may technically not be possible as the process of anonymisation is not absolute**. "A lot of this data would be the anonymised data and the standard for anonymisation in the PDP bill is an irreversible standard which at this point is not an achievable standard per se. So that automatically creates a lot of confusion in terms of which anonymised non-personal data will be outside the ambit of the PDP bill [2018]," a participant said.

One participant argued that **even if it is not a perfect technique, anonymisation is a useful, functional method.** "Re-anonymisation is possible. So anonymisation is not a perfect construct, but it is a utilitarian concept. That's a political decision that we want to make [whether we want to anonymise or not]. Even census data to a large extent can absolutely be re-identified. Does that mean that we do not publish census information any longer? It doesn't, right? Anonymisation can't be seen as a black and white kind of [concept]. It's a very much a utilitarian concept, and we have to place those political boundaries," a speaker said. Another speaker pointed out that there are different grades of anonymisation techniques available, and it's up to us how we utilise them.

But even anonymised data could eventually be traced back to an individual. "Non-personal data is made out of clusters and something makes those clusters. In the electronic scheme of things, it's very easy to find out which points make each cluster. Anybody with access to anonymous data can actually come back to the source. There is an electronic trace," a participant said.

"The government's understanding [is] that certain kinds of data that is linked to individuals — this doesn't relate to community data or non-individual data like air pollution data — but this does relate to behavioural patterns, aggregated data can be anonymised. It is very much capable of being reverse-engineered and re-identified. The risk is there and it's a fairly high risk to take. The assumption that anonymised data — we don't even know what these anonymisation standards are —, it is safe and is no longer a threat to an individual's privacy needs to be challenged."

A participant pointed out that **in certain circumstances, the ability to trace back and de-anonymise data is crucial.** While anonymised medical data can be used keep entire populations safe from diseases, "we should be able

to trace anonymised data back to its origin," the participant said. "Let's say my genetic data is out there and there is an outbreak in the market which marks me out for a disease. I should come to know of it," they said. They suggested that use of data should determine whether or not data should retain "a certain degree of granularity". This would be determined by law, keeping in mind the risk, they clarified.

Can non-personal data be used to profile groups of people?

Even anonymised non-personal data can be used to profile, persecute, and discriminate against groups of people, some speakers pointed out. At times, de-anonymisation does not even require sophisticated algorithms to target people.

One speaker told us about how political parties send SMSes and videos legally in a geo-fenced area with the help of telecom companies. "Telecom companies will map out the towers that address this geo-fence — without telling the political party the numbers — but they will disclose how many people have latched onto those towers in the last six months, and then allow me to send messages there," they explained. This kind of service can be paired with other data points — such as mosques in an area through data scraped from Google Maps, or even from the voter roll — and can be used to target potential voters with a lot of specificity. The same means can be used to persecute people.

> "Formally speaking, we have a secret ballot system. But when a political party contests elections, their workers always know who voted for them and who didn't vote for them. This happens because this data is identifiable, it can be narrowed down and correlated with other data points.
>
> "For example, if there is a booth and there are 30 families under it. If a particular political party does very badly, [the party workers] are going to relate it to other things, such as, those families did not come for the rally, or who asked questions when a person went to do their door-to-door campaign, etc. So when correlated with other data, parties always know who did not vote for them, and that is how the political workers ultimately give or deny access to to the politician once they are elected. We will be really kidding ourselves if we think that it is possible to fully anonymise any kind of data."
> — a participant at the discussion

Non-personal data will not be used in silos: A participant explained how the Delhi Police collects two kinds of information about people for their crime maps — their location and their socio-economic information. The latter is collected on a region basis and subsequently used to profile regions as "no data would be used in a silo; it would be used with other identifiers so it would always brand a region". They pointed out how these methods are used to "red-line" areas and brand areas as "risky" or "not risky".

Linking databases to distribute government sources: "Even [with] air pollution data, you can say which category of people are creating more pollution, which areas are creating more pollution. You can trace it onto a group of people, say 10 people. How can it not be personal data?" a speaker said. Nikhil Pahwa, editor and founder of MediaNama, pointed out the dangers of linking this data with vehicular data and registration numbers to potentially target people. "If you can trace it [air pollution data] down to an individual, directly or indirectly, it becomes personal data," another speaker remarked.

> "After a point, people will start buying properties where there is less air pollution. You can also imagine lot of things happening with property rates and schools. All nice people living in one location and all 'bad' people living in other location. All the government resources will be sent to those nice locations." — a participant at the discussion

Using health data to profile people: "Let's say anonymise health data by region. There is this area that has very high incidence of triglycerides (a type of body fat) in people. So you say cholesterol. What is the source of cholesterol? Higher meat eating. Is that a social marker? Is that a cultural marker that interests somebody? When you start thinking down that line, the possibilities are endless and very, very scary," a speaker said.

Using people's buying habits against them: "Let's just say we look at data of women who are buying vibrators online on Amazon, and then we narrow it down to a geographical area and that area comes around to say for example, the area around the university in the city. Then will the women of the university be targeted as being of a loose character?" another participant said.

Data collection for aggregated purposes: "When demographic information is sought in my area, do I have the option of saying that I don't want things like my religion being used even in an aggregated way? Because aggregated anonymised non-personal data can lead to significant harms. There are examples in Andhra Pradesh and Telangana, where such databases exist and can be analysed very quickly for religion. And you can have gram panchayat-wise percentage of Hindus, Muslims, et cetera. And I think the possible harms there are very, very clear as 1984 and 2002 have shown us."

- Another person said that during a data collection drive for one of the databases in Haryana, the forms mandatorily asked for caste and religion. Even when this data is aggregated and anonymised, it can still be used to profile groups of people by region.

*

Read our coverage of the our discussion on Non-Personal Data in Delhi here. The discussion was held in New Delhi on November 28, 2019, with support from Amazon Web Services, Facebook and FTI Consulting.