T20 Policy Brief

G20 BRASIL 2024

T20 Brasil 2024 Let's rethink the world

Task Force 05

INCLUSIVE DIGITAL TRANSFORMATION

Towards a Framework of Data Protection for Open Data

Paula Luvini, Researcher, Fundar (Argentina)

Apoorv Anand, Data Lead, CivicDataLab (India)

Mariana Kunst, Coordinator, Fundar (Argentina)

Sai Krishna Dammalapati, Senior Data Engineer, CivicDataLab (India)

TF05

Abstract

Governments are one of the world's largest generators and users of data. As such, they usually carry open data initiatives to release data for public use. The G20 has played a pivotal role in endorsing open data principles, particularly addressing data privacy and personal protection since the Hangzhou's Leader's Communique in 2016 and reinforcing it in later meetings (Branford-White, 2017; Godhwani et al., 2023).

Data openness is a crucial movement that empowers civil citizens and organizations, allowing them to make conscious and informed choices. Nevertheless, open data initiatives sometimes face challenges while handling personal information and privacy. Creating protocols and benchmarks on database catalogues can address these situations. In addition, it is essential to include good practices for data anonymization to avoid re-identifying individuals on the basis of their personal data.

This policy brief is built upon two cases where open data portals and data privacy are at a crossroads: the protection of children's privacy in the criminal justice system and using sensitive worker information. These examples are intended to provide a broad and global perspective on the data protection issues since they reflect the experience and work of two different think tanks from different regions.

The G20 can encourage developing and adopting agreed protocols and guidelines to balance data privacy and open data. This policy brief outlines key considerations that governments should consider to implement safe and responsible data protection policies, taking the aforementioned examples to discuss the application of this framework. Data sharing facilitates evidence-based policymaking and fosters transparency, stressing the importance of creating a framework for data protection in the public sector.

Diagnosis of the Issue

Every time citizens fill out government forms, open a company, or pay their taxes, they share information with an agency that is probably storing it. Governments are, in fact, one of the largest data generators in the world. They also consume data: to make informed decisions, enhance service delivery, promote innovation and design more efficient and effective public policies. To work with data it is necessary to ensure information is stored safely since it might contain sensitive citizenship information.

Governments usually carry out open data initiatives to release data for public use. Data openness is a movement that empowers citizens and organizations, allowing them to make conscious and informed choices. Furthermore, enhancing data sharing between government agencies is desirable to make public policy decisions. Nevertheless, such initiatives face challenges while handling personal information and privacy.

The G20 has played a pivotal role in endorsing open data principles, particularly addressing data privacy and personal protection since the Leader's Communique in 2016 and reinforcing it later (Branford-White, 2017; Godhwani et al., 2023). Moreover, the UN has stressed the importance of having quality data to implement and monitor progress on the Sustainable Development Goals (SDGs) and the need for a "new open data management framework" to foster innovation (UN, 2017; UN, 2022).

- **Case Study 1: Protection of children's privacy in the criminal justice system**

This section addresses the challenges and necessities of balancing privacy and transparency in judicial proceedings, particularly in cases involving sexual crimes and children. Increasingly, in India, orders and judgments pertaining to cases of sexual crimes

are not being made available on the e-Courts[1] portal. The non-availability of relevant information affects the right to information of the parties in a case, making them fully dependent on their counsels, and increasing their vulnerability to corrupt and exploitative practices. It also hampers bona fide research, review and social audits that are necessary for good governance. Keeping in mind the fact that information concerning a case is confidential and any information on a public platform revealing the identity of the victim/survivor can be detrimental to their rehabilitation and well-being and would be a violation of their rights, there is a need to identify a way to achieve the twin goals of privacy and confidentiality of victims and witnesses and judicial data transparency, access and accountability.

A particularly sensitive case relates to situations involving children since they must be especially protected from the damaging effects of engaging in the criminal justice system. The protection of children from all forms of violence is recognized in the UN Convention on the Rights of the Child and in SDG 16. Hence, it is mandatory to protect them from the persecution and stigmatization they might suffer from their involvement in these cases. Some previous work has dealt with this issue, (Elder et al. 2020), where recommendations and practices in protecting children's confidentiality were given.

- **Case Study 2: Usage of sensitive worker information**

This section will discuss the initiatives by the Ministry of Culture of Argentina to enhance data-driven governance and transparency in the cultural sector. The Ministry of Culture of Argentina created the Federal Registry of Culture (FRC) in 2021, during the COVID-19 pandemic. On the registry's website, people working in the cultural sector

[1] https://ecourts.gov.in/ecourts_home/

could enroll and create a profile to apply to programs of the Ministry. The registry is also processed to create management tools for the government.

Additionally, the registry allows to characterize the population targeted by the Ministry's policy. This diverse population includes some profiles as set designers, craftsmen, musicians, illuminators and teachers of artistic disciplines, among others. It also covers a wide spectrum of recipients in terms of geographical location, employment status, income, age and gender identity. It is also the first administrative registry of the Ministry to include a question about belonging to indigenous communities (Directorate for Management Planning and Monitoring, 2023). The registry also integrates data from other government sources to characterize the population targeted by cultural policies. It is necessary to have this granularity as part of the Ministry's objectives is to implement public policies that promote the visibility of cultural diversity.

Regarding the data availability, the "Culture in Data" website[2] was created to publish the databases. Data confidentiality is ensured by restricting access to sensitive information exclusively to authorized personnel from the Ministry and through an anonymization process so the citizens can use this data without compromising the identity and privacy of those included in the registry.

[2] https://www.argentina.gob.ar/cultura/cultura-en-datos

Recommendations

Promoting an inclusive world in an increasingly digitalized environment is essential. In order to address issues such as poverty, inequality or climate change, governments and international organizations need to have tools to diagnose and measure the impact of their actions. For that reason, this policy brief encourages the countries of the G20 to produce, share and use data in innovative ways to tackle the challenges of sustainable development.

One opportunity governments have to include new data sources is by exploiting administrative registers to extract valuable information from them, like the mentioned example of the Federal Registry of Culture. The registry was a precious case as a part of it was a displaced and marginalized population that was not included in other official administrative registers. The digitalization and use of data from the judicial system is also essential, not only to hold this system accountable but also to analyze how sentences are handed down and to monitor its functioning.

However, governments need to consider the consequences of dealing with sensitive data and find ways to protect the integrity of personal information. Nowadays, collecting personal and user information is quite usual, beyond the specific information that each database provides. Identity preservation is a right of citizenship and an obligation of the person or agency that safeguards the information. The informed consent must clarify how data will be used and whether it will be retained or deleted after its use: citizens must have control over how their data is used. Both primary uses (for which the data was collected) and secondary uses (subsequent uses for other purposes, or uses of data collected by other entities or for different objectives) should be considered.

In this sense, states should have accessible and clear legislation relating to how privacy interacts with sensitive cases. To take an example, the judicial system should determine

whether or not access is to be granted regarding case records through the relevant legislation. A process to access court records substantiated in legislation would be a significant step towards promoting judicial accountability and demonstrates the willingness of the court system to commit to the concept of open justice.

Once the legal processes are considered and it is decided to open and share data with personal information, it is necessary to carry out an anonymization process to preserve the identity and privacy of the individual. Certainly, not every information exchange carries the same risk: making a database public, as in the case of the registry is not the same as sharing it between secretariats of the same ministry. Hence, before sharing information we should check two important things.

First of all, **the content of the database should be checked.** Not all data is sensitive or personal and should be treated similarly. All countries and government agencies should have their own data catalogue that declares how sensitive the information in each database is. After creating the catalogues, governments should create protocols based on them so they know where they need to be careful. Secondly, **potential users of the database should be checked.** There are different degrees of data openness: inside the organization, to external people, and to the public. There should be a protocol for which data can be shared with other members of the organization. For example, the registry might be shared among members of the Ministry of Culture of Argentina if they work in an authorized area and can be partially anonymized. But in the case of opening data to the public, there should be more assiduous work in the anonymization.

After both those points are considered, some decisions should be made to share data. According to the outlined protocols, if the data is shared among members of the organization or other teams that have signed some agreements of non-disclosure of the information, then deleting variables such as names, email addresses, cell phone numbers

and addresses is enough. If data is going to be open to a wider audience, we can follow the steps below -based on (Yankelevich, Daniel 2021) and (Luvini, P. 2022)- to handle sensitive data carefully:

1. **Identify data:** Analyze all the variables in the database and consider whether they can be cross-referenced with other external data sources that allow for the re-identification of individuals. If any of these variables are sensitive to cross-referencing, then they should be masked or encrypted. If some variables are partially masked, there is a greater chance that the citizens will be identified. For instance, the Federal Registry of Culture collected glocalization data that if it was left, individuals would be very easily identified in places with low population density.

2. **Identify risks:** Data encryption should be prioritized and will depend on the risk and usage scenarios. Suppose the risk of citizens being re-identified is high and will violate their identity or have implications for their lives. In that case, this must be prevented by modifying and anonymizing the database. For instance, if children are involved in some part of a judicial case their identity should be preserved at least until they reach adulthood and can decide whether or not to keep their anonymity. This empowering feature allows victims of sexual crimes for example to take autonomy over their own lives.

3. **Identify solutions:** Some techniques can help avoid personal data identification. First, we should remove unnecessary columns with sensitive information, and the remaining ones that might be dangerous can be hashed and encrypted. In some cases, it will also be necessary to group data and give up some granularity. Some measures taken in children protection cases were to employ name suppression techniques, such as using initials or pseudonyms, and some redaction mechanisms to remove or erase from a record before it is shared.

4. **Identify attacks and problems:** Considering that attacks can evolve and change through time, it is important to check periodically for new risks. If there exists a potential database to be crossed with the one we are sharing that would allow some re-identification, we should consider it while anonymizing and encrypting.

Scenario of Outcomes

Opening and sharing data will always have a tight relationship with protecting sensitive information. In this policy brief, we have identified some steps and protocols that the governments of the G20 can follow to open data. This framework is quite general, leaving some space for countries to be specific in their benchmark and standards. In this section, we will list the advantages and disadvantages that the proposed framework of open data has.

First, we would like to highlight that **opening and sharing data is important to make public policy decisions based on evidence and to enhance transparency.** As we mentioned, opening data is not a goal on its own, it is also about what that data can be used for. Sharing data safely among agencies and teams of the government can bring a lot of benefits like making innovative analyses to provide public policy-makers with information and tools.

For instance, the Federal Registry of Culture was originally a website where people could enroll in the Ministry's programs but the information it gathered was transformed into a tool to segment groups of people who needed different assistance from the state (Avenburg, A. et al. 2022). To do so, the Ministry's office in charge of the database had to share it with a different organization, thus anonymizing data to secure the identity of those people from the database. By doing this, they were able to do a cluster analysis of

the database and identify different groups and profiles based on the existing data. To do this kind of segmentation is crucial to fight poverty and inequality, because not every displaced group has the same needs or demands, hence the policy we can do to help them will vary.

Children's protection in the justice system is another great example of how we can pursue inclusive development by using data. The 2020 SDG report pointed out that "Sexual violence, one of the most disturbing violations of children's rights, is widely underreported." due to the lack of comparable data across countries (UN, 2020). Recommendations such as those stated in this policy brief and in (Elder et al. 2020) pave the way for progress in this goal. First and foremost, we need to understand and diagnose the issues we face in order to find solutions to them.

On the other hand, **anonymizing and protecting personal data is a complex task.** There is more than one example of intended anonymized databases that were later reidentified. In some cases, the steps mentioned above were not strictly followed and in other cases, a new dataset was published that allowed cross-checking the information and identifying the persons or organizations in the database. A case in point was from (Narayanan and Shmatikov, 2008), who demonstrated that a database of anonymous movie ratings of Netflix users could be reidentified, uncovering the users' political preferences and other sensitive information. This case was a huge scandal that led the company to a court trial and to shut down all the competitions it had in place using its customer's supposedly anonymous information.

Conclusion

Overall, this policy brief aimed to set a framework to enhance data protection. As we previously mentioned, fostering such kind of policies that protect personal data is a means to an end, not an end in itself. To increase the interoperability among government agencies or to share more datasets with the public needs to be a greater goal than just increasing the open data portals. Information availability is fundamental to improving public policy decision-making and scientific investigations. However, sharing data is not free: there is a high cost to be paid to obtain data protection. As pointed out before, citizens have the right to keep their data and identities private, so data protection and anonymization techniques should be implemented. Such solutions are not free of certain risks and attacks that databases can suffer.

Throughout the policy brief, we have stated the benefits and the challenges that arise when a database is shared or opened. Finding the balance between access to personal data for decision-making and privacy is a hard task that governments of the G20 should face. To tackle the challenges of our time it is mandatory to capture new data sources and to use existing information in novel ways.

References

Avenburg, A., Houllé, J., Luvini, P. y Rodrigues Pires, M. “Guía práctica para caracterizar a la población objetivo de una política pública a partir de registros administrativos.” (2022).

Directorate for Management Planning and Monitoring from the Ministry of Culture of Argentina. “Informe Registro Federal de Cultura 2021-2022.” (2023) Available in https://www.argentina.gob.ar/sites/default/files/mc_dpysg_informe_registro_federal_de_cultura_2021-2022_0.pdf

Elder, Kane, Maddison Tan, Nicholas Trappett, and Emilia Turnbull. "BALANCING CHILDREN’S CONFIDENTIALITY AND JUDICIAL ACCOUNTABILITY." (2020).

López, S.; Alonso Alemany, L.; Dias, J.M.; Ación, L. y Xhardez, V. (2023). Guía práctica para la protección de datos personales en salud. Buenos Aires: Fundar. Disponible en https://www.fund.ar

Luvini, Paula. “Guía práctica para la protección de datos.” (2022) Buenos Aires: Fundar.

Narayanan, A. y V. Shmatikov. De-anonymizing Social Networks,” 30th IEEE Symposium on Security and Privacy, Oakland, California, pp. 173-187. (2009)

United Nations, “The Sustainable Development Goals Report” (2017)

United Nations, “The Sustainable Development Goals Report” (2020)

United Nations, “The Sustainable Development Goals Report” (2022)

Yankelevich, Daniel. “Anónimos pero no tanto: cómo hacer una gestión de datos eficiente sin poner en riesgo la privacidad.” (2021) Buenos Aires: Fundar.

G20 BRASIL 2024
BUILDING A JUST WORLD AND A SUSTAINABLE PLANET

T20 Brasil 2024
Let's rethink the world

Let's rethink the world

CEBRI Rethink Tank

FUNDAÇÃO ALEXANDRE DE GUSMÃO

ipea
Instituto de Pesquisa Econômica Aplicada