Open data on the COVID-19 pandemic: anonymisation as a technical solution for transparency, privacy, and data protection

Thiago Guimarães Moraes*, Amanda Nunes Lopes Espiñeira Lemos**, Alexandra Krastins Lopes***, Camille Moura**** and José Renato Laranjeira de Pereira*****

Introduction

The COVID-19 pandemic has boosted personal data mining. Various experiences in fighting the new coronavirus focus strictly on extensive data processing concerning not only health conditions, but also behaviours and relationships, allegedly to try to control the spread of the virus at any cost.

Several approaches have been put forward by different government. In China, the government has been widely using body temperature, geolocation and user behaviour data to monitor the spread of the infection and commitment to quarantine, and scientists in the USA are processing global scale data to find vaccines for the virus.1 Stakeholders in Brazil have been trying to follow similar approaches.2

This vast data processing raises concerns on uses for secondary purposes that citizens might be unaware of,3 such as the use for surveillance activities,4 which may involve intrusive tools implemented by governments.5 For instance, InLoco, a Brazilian company involved with the use of geolocation for contact tracing, allegedly processes only anonymized data pursuant to the Brazilian Data Protection Legislation, ‘Lei Geral de Proteção de Dados’—LGPD.6 Although it claims to implement anonymization techniques in its privacy

Key Points

- This article aims to discuss transparency and privacy on COVID-19 public data. The Brazilian context was chosen as a case study to understand how to implement transparent COVID-19 government databases while respecting data protection.

- The article intends to answer the question: which technical approach could be implemented to publish transparent data on COVID-19 infections while respecting the privacy of individuals?

- Privacy by Design is brought as a key concept to harmonization between these topics. It is possible to conclude that while transparency databases are essential when facing a pandemic outbreak, they must respect the rights to privacy and data protection.

- We present anonymization as a set of four strategies which aims to provide unlinkability to a given dataset: (i) to minimize, (ii) to separate, (iii) to abstract, and (iv) to obfuscate.

- While anonymization is a technical solution that may help to bring balance between open data transparency and privacy, it needs to be complemented by organizational measures, such as access control, privacy policies, and data protection impact assessments.

---

1 ‘Big Data: O Uso Da Informação Na Luta Contra O Coronavírus’ (Summit Saúde Estadão, 2020) <https://summit.saude.estadao.com.br/big-data-o-uso-da-informacao-na-luta-contra-o-coronavirus/> accessed 19 July 2020

2 L. Belli, ‘Os Big Data Do Coronavírus’ (EL PAÍS, 2020) <https://brasil.elpais.com/opiniao/2020-03-09/os-big-data-do-coronavirus.html> accessed 19 July 2020

3 Y. Harari, ‘The World After Coronavirus’ (Fortune, 2020) <https://www.ft.com/content/19d90308-6858-11ea-a3c9-1f6fe6dcca75> accessed 19 July 2020

4 ‘Mobile Location Data and Covid-19: Q&A | Human Rights Watch’ (Human Rights Watch, 2020) <https://www.hrw.org/news/2020/05/13/mobile-location-data-and-covid-19-qa> accessed 19 July 2020

5 Ibid.

6 BRASIL. Law no 13,709/2018. <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lc/l13709.htm> accessed 29 July 2020.
policy, the lack of a proper audit and more detailed information on the methodology for anonymizing data raise some eyebrows.

At the same time, the role of ensuring access to public information while protecting personal data can prove to be rather challenging. In the Brazilian Access to Information Law, ‘Lei de Acesso à Informação’ (LAI), access to personal data is restricted to public authorities, as provided by law, and the data subject. Although this rule aims to provide balance between transparency and privacy, it may tip to one side or the other if not properly implemented, which may lead to one of the following scenarios:

(i) the denial of requests of access to information due to alleged secrecy without further explanation, or (ii) the inadequate exposure of citizen personal data, supposedly to promote transparency, which can be even more sensitive in the context of public health.

These uncertainties raise the question of which technical approaches should be implemented to publish transparent data on COVID-19 infections while respecting the privacy of individuals. In this article, we argue that the answer to this challenge lies with the concept of Privacy by Design (PbD), which may be implemented through different strategies, one of those being anonymization.

Anonymization is a set of techniques by which data is processed in such a way that the data subject is no longer identifiable. Due to reduced possibility of linking information to an individual, anonymised data are not considered personal data, and thus fall outside the scope of data protection legislations such as LGPD.

Applying solid anonymization techniques is a useful means for enhancing data protection, allowing more robust data security, especially when taking into consideration that health information is considered sensitive data.

The increasing availability of data points and the enhancing scalability of data processing, nevertheless, makes it easier to link datasets and infer personal information from ostensibly non-personal data. Moreover, there always remains a residual risk of identifying data subjects from anonymised datasets. Assessing the degree of such risk is paramount for defining whether data has or has not been reasonably anonymized.

However, the LGPD does not make clear what reasonable anonymization means, and article 12 only mentions that if by reasonable efforts, the anonymization can be reversed, the data protection law will apply. Furthermore it states that this reasonability must consider time, cost, available technology, and the exclusive use of ‘one’s own means’. It is not clear how each one of these criteria should be applied (specially the latter) and it is expected that the Brazilian Data Protection Authority (DPA) will provide more clearance on that subject.

This article aims to contribute to the debate regarding privacy on open databases by identifying anonymization techniques that could be implemented to guarantee the right to data protection on Brazilian COVID-19 transparency databases while respecting an open government data approach. First of all, we present the current status of two Brazilian frameworks: the one of data protection and the other of government transparency. In order to find the necessary balance between them, we analyse the concept of PbD. We argue that anonymization is an essential component of a PbD framework.

In the second part of this article, we highlight some anonymization techniques as identified on guidelines and recommendations of some Data Protection Authorities, as well as some provided in a ‘toolkit of mitigations’ by the Berkman Klein Center, in a study on Open Data Privacy. We portray these techniques.
considering their implementation in the context of the development of transparency health databases in a multilayered governmental framework, such as the case of the Brazilian State. In Brazil, there are three layers of government: Federal level, State level and Municipality level, which adds extra complexities to the issue, since multiple COVID-19 transparency databases have been developed during this crisis. We believe that regulators from other regions with multilayered government structures such as the USA and the European Union (EU) might be able to enjoy the recommendations herein given when assessing compliance with data protection regimes of transparent and anonymous databases.

Although it is not in this study’s scope, it is essential to make it clear that a set of associated organizational measures would also be necessary for ensuring transparency and privacy, such as access control measures and privacy policies. While the PbD framework embeds both types of actions, this article will focus solely on technical measures.

The authors from the Berkman Klein Center’s paper states that merely removing personal identifiers from open databases is not enough. That applies especially in a connected world where multiple datasets may be correlated to identify an individual (a procedure known as re-identification). Therefore, they make four recommendations to guarantee the utility of open data while ensuring privacy: (i) conduct risk-benefit analyses to inform the design of open data programs; (ii) consider privacy at each stage of the data lifecycle; (iii) develop operational structures and processes that codify privacy management throughout the open data program, and; (iv) emphasize public engagement and public priorities as essential aspects of open data programs.

Nevertheless, it is never enough to emphasize that, to guarantee adequate data protection, technical and organizational measures must come hand-to-hand.15

Methodology

Considering that anonymization is not a sole measure, but a set of techniques implemented to achieve unlinkability between the data and its subject,16 this article presents some suggestions for developing a privacy-friendly COVID-19 transparent database.

In this research, we have selected Brazil as the case study model for the presented approach. Brazil presents some features that may allow for the recommendations herein given to be applied in other regions. First of all, it has a continental size, meaning that the scalability and heterogeneous distribution of its population plays a role when building COVID-19 databases. Second, it is a Federal-State, meaning that there are several layers of government—municipal, state/district, and federal—that collect and process health data to build these databases, sharing data between each other. Finally, organizations have been publishing transparency rankings of Brazilian COVID-19 databases, such as the one from Transparência Internacional,17 and the Índice de Transparência da Covid-19, from Open Knowledge Brazil,18 which have assisted us to better understand the transparency level of state databases.

In short, this means that the techniques here presented, must take into account the following features prior to their implementation:

1. Different locations may have diverse population’s size and density, meaning that particular attention may have to be given in areas with low size/density to avoid the identification of infected individuals through the crossing of personal data;

2. To build databases of higher governmental layers (eg Federal level in Brazil and USA or Community level in the EU), data sharing will happen between these layers and lower ones.

3. Resources will not be the same among different governmental entities, and some less resourced entities—with regard to both financial and human resources—may need to give extra care to non-technical, organizational measures that help to enforce PbD.

Theoretical framework

First things first: the state of the art of the Brazilian data protection framework

Before discussing how anonymization can assist in allowing data on COVID-19 to be transparent without...
compromising the individual’s expectations of privacy, it is important to understand the current status of the Brazilian data protection framework.

The Brazilian LGPD is the result of a long debate which started in 2010. Before the pandemic, the legislation was scheduled to come into force in August 2020. Nevertheless, due to the increased pressure of some private sector stakeholders, who claimed that the public health crisis would create a heavy burden on their business, the Brazilian government attempted to postpone its entry in force to May 2021, through the Provisional Measure no. 959/2020. In addition, the enforceability of administrative sanctions was postponed to August 2021.

In a dramatic turn of events, on 26 August, the Brazilian Senate dropped Article 4 from Provisional Measure no. 959, which postponed the full entrance into force of the LGPD to May 2021. As a consequence, the Legislative power established the immediate validity of the legislation, which was sanctioned by the Brazilian President on 18 September. However, administrative sanctions are still to enter into force in August 2021.

Simultaneously, Congress approved an emergency law provided for health measures to be adopted under the pandemic and obliged health authorities and public administration to share data regarding COVID-19 cases. Furthermore, two other legislative acts targeted transparency and data processing, threatening both privacy and access to information rights: Provisional Measures (MP) no. 928/2020 and no. 954/2020. The former determined the summary suspension of the deadlines for responding requests to public information, which are ruled by the Information Access Law (Lei de Acesso à Informação), or LAI, under its acronym in Portuguese. This resulted in making the government’s use of data even more opaque.

The latter compelled telecommunications companies to deliver Brazilian’s personal data, such as name, phone number and address, to the Brazilian Institute of Geography and Statistic (in Portuguese, IBGE), a public body that processes data at federal, state and municipal level. Even though the massive data sharing aimed for statistical data on the pandemic, the initiative failed to provide for sufficient safeguards, and thus had its effectiveness suspended by the Brazilian Supreme Court under the Direct Action of Unconstitutionality, ‘Ação Direta de Inconstitucionalidade’ (ADI) no 6387, as it would violate the constitutional right to privacy.

The resulting decision was a landmark, as it recognized the right to informational self-determination as a fundamental right, an important landmark for data protection in Brazil. It stated that fundamental rights could not be suppressed to face the pandemic, and highlighted the transparency principle when mentioning that data collection under the Provisional Measure was not proportional for its intended purposes. Other principles underlined by Justice Rosa Weber as paramount for any processing of personal data, and that were not provided by MP no. 954, were those of purpose limitation, adequacy, necessity, and data minimization.

Despite these advances, Brazil is yet in the process of establishing its DPA, the Autoridade Nacional de Proteção de Dados (ANPD). The ANPD will play a vital role in Brazil as it will be the body responsible for

---

19 P Soprana, ‘Senado adia lei de proteção de dados por causa do coronavírus’ Folha de São Paulo (São Paulo, 3 April 2020) [https://www1.folha.uol.com.br/mercado/2020/04/senado-adia-lei-de- protecao-de-dados-por-causa-do-coronavirus.shtml] accessed 14 July 2020
20 ‘Brazilian president orders LGPD delay to May 2021’ (IAPP, 2020) [https://iapp.org/news/a/brazilian-president-orders-lgpd-delay-to-may-2021/] accessed 27 July 2020
21 F Rebello, ‘Effectiveness of LGPD Sanctions in Brazil Postponed to August 1st, 2021’ [https://www.lexology.com/library/detail.aspx?g=76e2299d-f6e1-4af9-a2df-c960155b16cb] accessed 27 July 2020
22 Provisional Measure no 959/2020, Art 4: ‘Law No. 13,799, of August 14, 2018, becomes effective with the following changes:
‘Art. 65. ……………………………………………………………………………………………………………………………
II - on May 3, 2021, regarding the other articles.’ (Free translation).
See <https://www.congressonational.leg.br/materias/medidas-provisorias/-/mpv/141735> accessed 29 July 2020.
23 Law no 13,979/2020, Art 5 Every person will collaborate with the health authorities in the immediate communication of:
‘I - possible contacts with infectious agents of the coronavirus;
II - circulation in areas considered to be regions of contamination by the coronavirus.’ (Free translation).

---

24 Provisional Measure no 928/2020, <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/119979.htm> accessed 29 July 2020.
25 Provisional Measure no 954/2020, <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/mpi/mpv954.htm> accessed 18 October 2020.
26 ‘STF suspende compartilhamento de dados de usuários de telecomunicações durante a emergência de saúde pública’ (Senado Federal, 2020) [https://www2.senado.gov.br/noticias/materias/2020/03/24/suspende-compartilhamento-de-dados-de-usuarios-de-telecomunicacoes-durante-a-emergencia-de-saude-publica] accessed 19 July 2020.
27 See <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/141619.htm> accessed 27 July 2020.
28 ‘STF suspende compartilhamento de dados de usuários de telecomunicações com IBGE’ (STF, 2020) [http://portal.stf.jus.br/noticias/verNoticiaDetalhe.asp?idNoticia=442902] accessed 27 July 2020.
29 See <http://www1.folha.uol.com.br/mercado/2020/04/senado-adia-lei-de-protecao-de-dados-por-causa-do-coronavirus.shtml> accessed 14 July 2020.
30 Although the ANPD was created by Law no 13,853/2019, which modified the LGPD, the DPA lacked its organizational structure. In August, 26th, 2020, Presidential Decree provided this structure, but its entry into force was subject to the nomination of its Board of Directors. As of this writing, the five Board members have been referred and are undergoing a
providing guidance on data protection issues. In the given context, it would be able to explain how data protection should correlate with government transparency.

Another legal instrument, Presidential Decree no. 10,046/2019, has also been under the spotlight as it provides for practical guidelines regarding the processing of personal data by the public administration. However, the main reasons for the attention it has received are mostly due to its disregard for key data protection principles.

According to the Decree, to determine the level of access permissions for data transfers between different government agencies, the only aspect to be taken into consideration by the ‘data manager’ (a concept from the Decree that does not exist in the LGPD) is the level of confidentiality of the data, and not its sensitivity or the purpose of the transfer. This means that non-confidential sensitive data may have the same treatment as non-confidential ‘general’ personal data. Such reasoning goes against the principles of the LGPD, especially purpose limitation, whereby every data processing shall be carried out according to the legitimate, specific and explicit purposes of which the data subjects are informed. In this sense, the Presidential Decree allows that individuals’ data be shared with no regard to the contexts to which they will be applied, permitting that both sensitive and non-sensitive data are shared without adequate protection.

The constitutionality of the Decree and its application for national security purposes is now under assessment by the Supreme Court, after the leakage of an agreement for sharing more than 70 million individuals’ data held by the National Transit Department (DENATRAN) with the Brazilian Intelligence Agency (ABIN), signed supposedly in accordance with the Presidential Decree no. 10,046/2019, Art 4. The sharing of data between the bodies and entities referred to in Article 1 is categorized into three levels, according to its confidentiality:

I - wide sharing, when dealing with public data that are not subject to any access restriction, the disclosure of which must be public and guaranteed to any interested party, in accordance with the law;

II - restricted sharing, when dealing with data protected by secrecy, under the terms of the legislation, with access granted to all the bodies and entities referred to in Article 1 for the execution of public policies, whose sharing mechanism and rules are simplified and established by the Central Data Governance Committee; and

III - specific sharing, in the case of data protected by secrecy, under the terms of the legislation, with granting access to specific bodies and entities, in the cases and for the purposes provided for by law, whose sharing and rules are defined by the data manager. (Free translation, emphasis added).

The regulation of government transparency in Brazil

Open data is essential to increase the transparency of investments in public resources and their results. In addition to economic impacts, during the COVID-19 pandemic, it also has an assistant role in combating the health crisis through information to citizens. As mentioned above, Brazilian government public information transparency is ruled by LAI, which establishes parameters to active transparency and defines criteria to assure access to government information through passive transparency alongside Decree no. 7.724/2012, which specifies the procedures to request and provide information.

LAI is the result of Brazil’s leadership at the foundation of the Open Government Partnership (OGP), an international coalition, with currently 78 Member States, dedicated to promote and support the implementation of government policies based on the principles of OpenGov, which involves transparency, public participation, innovation, and accountability. OGP demands governments to design and implement biannual Action Plans. These plans are set through a series of activities involving multistakeholder groups composed by public officials, NGOs’ representatives, researchers, journalists, and other interested citizens. Brazil is currently implementing its fourth OpenGov Action Plan.

The creation of transparency platforms was one of the results of the Brazilian Action Plans that increased public availability of government information, like the Open Data Portal. Designed under the principles of Open Data, the website is a repository for government databases that can be downloaded and freely reused.

In the field of public health, digital information management in Brazil is structured by DATASUS, the data processing department of the Ministry of Health. In short, the division is responsible for providing some...
basic management softwares to all public health agencies across the country and compiling their data to organize, analyse and publish national statistics. Despite the importance of DATASUS as a driver of transparency for the publication of data in the health sector, the public body has developed its solutions without a deep concern with the processing of sensitive personal data, such as safeguards that should be taken into account.37

In the midst of the COVID-19 pandemic, it is necessary to understand the key governance and policy issues related to critical decision-making.38 Transparency is essential in building public policies and ensuring legitimacy in decision-making processes39 and has impacts in several areas. Moreover, research findings state that transparency can be an important government ally while facing an outbreak.40

Brazilian statistics about the pandemic are registered in at least three systems that are maintained by the federal government: e-SUS VE,41 which holds information on COVID-19 contamination with mild symptoms; Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP Gripe),42 destined to track infections that accurately tackle the pandemic, Brazilian public are not covered by DATASUS’s softwares. In short, to since they often need to focus on specific themes that technologies of their own to manage health policies it is common for local governments to also use other provided by both public and private laboratories. In addition to the systems provided by the federal government, it is common for local governments to also use other technologies of their own to manage health policies since they often need to focus on specific themes that are not covered by DATASUS’s softwares. In short, to accurately tackle the pandemic, Brazilian public managers must be able to properly compile, clean, analyse and interpret data collected and processed in several different systems and databases, each with its own structure.

Unfortunately, this does not seem to be possible with the Brazilian open databases: e-SUS VE and SIVEP Gripe have several consistency issues. First of all, they use different approaches to compile data: while the former use separate spreadsheets per state,44 the latter relies on one single huge database with microdata for the whole country.45 Furthermore, different states use different features to categorize symptoms.46

The inconsistency between these different approaches raises serious concerns. As soon as an individual COVID-19 case has its symptoms evolving from mild to severe, its data need to be transferred from e-SUS VE to SIVEP Gripe, in a procedure whereby a given entry has to be switched to fit the latter database. As the federal government does not publish a unique identifier to each COVID-19 case, citizens and researchers cannot control the evolution of datasets and thus be sure about the microdata’s integrity.

Another issue is that, under the Brazilian methodology, if a specific data transfer from one database to the other is carried out incorrectly, it may result in duplicate entries of the same case, causing them to appear in both e-SUS VE and SIVEP Gripe simultaneously. This may have been the reason why the Ministry of Health declared in October that some inaccuracies were identified in the databases, resulting in the number of reported deaths being higher than it should be.47

37 C Berro Zapata and others, ‘Comportamiento Informacional Y Políticas Públicas De Información: Consideraciones Teóricas Alrededor Del Caso De DATASUS En Brasil’ (2016) 30 Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información.
38 M Moon, ‘Fighting COVID -19 With Agility, Transparency, and Participation: Wicked Policy Problems and New Governance Challenges’ (2020) 80 Public Administration Review.
39 Y Panov, M Wanke and T Vogel, ‘Transparency Effects On Policy Compliance: Disclosing How Defaults Work Can Enhance Their Effectiveness’ (2018) 3 Behavioural Public Policy <https://www.cambridge.org/core/journals/behavioural-public-policy/article/transparency-effects-on-policy-compliance-disclosing-how-defaults-work-can-enhance-their-effectiveness/099B43EAE7FEB09AEED27AE1D783F812> accessed 19 July 2020.
40 M Jansen and others, ‘Transparency-By-Design As A Foundation For Open Government’ (2017) 11 Transforming Government: People, Process and Policy <https://onlineibrary.wiley.com/doi/full/10.1111/puar.13214?~casa_token=0xkVt0UsAAAAA63Aovk_E-cr-k-MnNa2zCvzGig8a5ugc5gZrFZ1uq-xrljGxmzFRz6v65qg8LzNa2D9L~> accessed 27 July 2020 and KU Menon and KT Goh, ‘Transparency and Trust: Risk Communications and the Singapore Experience in Managing SARS’ (2005) <https://www.emerald.com/insight/content/doi/10.1108/13632540510621614/full/html> accessed 18 October 2020.
41 e-SUS Notifica (2020) <https://notifica.saude.gov.br/onboard/> accessed 28 July 2020.
42 ‘Dados Abertos - Coronavírus - Plataforma Integrada De Vigilância Em Saúde - Ministério Da Saúde’ (Plataforma.saude.gov.br, 2020) <http://plataforma.saude.gov.br/coronavirus/dados-abertos/> accessed 28 July 2020.
43 Gerenciador de Ambiente Laboratorial (GAL),43 which compiles COVID-19 test results provided by both public and private laboratories. In addition to the systems provided by the federal government, it is common for local governments to also use other technologies of their own to manage health policies since they often need to focus on specific themes that are not covered by DATASUS’s softwares. In short, to accurately tackle the pandemic, Brazilian public
44 Some states have multiple volumes of data. See Open Data SUS ‘SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19’ (2020) <https://opendatasus.saude.gov.br/dataset/casos-nacionais/resource/c283a162-d341-4bd5-a813-56f50f070af9?inner_span=xrlgILximxzFRAxv6Sgqc8LzjNxtD9L-~<processandpolicyarticle/transparency-effects-on-policy-compliance-disclosing-how-defaults-work-can-enhance-their-effectiveness/099B43EAE7FEB09AEED27AE1D783F812> accessed 17 October 2020.45 Open Data SUS ‘SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19’ (2020) <https://opendatasus.saude.gov.br/dataset/bd-srag-2020/resource/d89ea107-4a2b-4bd5-88b8-fa1caaa96550> accessed 17 October 2020.
46 For example, while the state of Espirito Santo has ‘headache’ as one of its features, the state of Amapa does not have it, but has ‘fever’ instead. See more on <https://opendatasus.saude.gov.br/dataset/casos-nacionais/resource/65b9d025-51be-4699-b988-5b2595f9505e> and <https://opendatasus.saude.gov.br/dataset/casos-nacionais/resource/c283a162-d341-4bd5-a813-56f50f070af9?inner_span=xrlgILximxzFRAxv6Sgqc8LzjNxtD9L-~<processandpolicyarticle/transparency-effects-on-policy-compliance-disclosing-how-defaults-work-can-enhance-their-effectiveness/099B43EAE7FEB09AEED27AE1D783F812> accessed 17 October 2020.
47 ‘Ministério Detalha Erro Que Incluiu 857 Mortes Em Balanço Da Covid E Reafirma 525 Óbitos No Domingo’ (G1, 2020) <https://g1.globo.com/bemestar/coronavirus/noticia/2020/06/08/ministerio-explica-dados-diferente-de-covid-19-e-diz-que-divulgou-numeros-errados-de-rondonia-e-do-ceara.html> accessed 17 October 2020.
Instead of solving these potential conflicts, the federal government has been reluctant to improve its information management and promote transparency over the pandemic data since the start of the outbreak. As previously mentioned in this article, in March, the government tried to suppress the right to access public information through the Provisional Measure no. 928/2020 that intended to suspend the response deadlines to public information requests as fixed by LAI. Journalists, specialists, and NGOs heavily criticized the act,\(^{48}\) that was subsequently repealed by the Brazilian Supreme Court through another Direct Action of Unconstitutionality, ADI 6351,\(^{49}\) whereby the court considered MP no. 928/2020 harmful to the constitutional principles of publicity and transparency.

One further serious threat to COVID-19 transparency emerged in June, when the Ministry of Health ‘blackened out’ the available data for an alleged ‘maintenance’ that made the COVID-19 portal more opaque with no information on the numbers of infected and deceased—as well as their rates per 100 million of the population.\(^{50}\) Once more, after social pressure, the government backpedalled and updated these statistics as well as e-SUS VE and SIVEP Gripe microdata although the datasets keep having the same issues mentioned above.\(^{51}\)

To assess the quality of data and information published about the pandemic on the Federal, State and some Municipal level government portals, Open Knowledge Brasil (OKBR) has developed a transparency ranking initiative.\(^{52}\) OKBR criteria for setting the transparency score of government bases are: (i) content; (ii) granularity, and; (iii) format. Its focus is on sanitary and epidemiological data. Data collection by OKBR is fortnightly, and the results of the evaluation are updated weekly. One of the features considered by the ranking system is the level of granularity of available data. Despite being very relevant to provide transparency, this element deserves to be highlighted since this criterion is a possible point of conflict with privacy and data protection, as if the information is too granular, individuals may be identified. Although the organization provides support for public managers with their processes of opening COVID-19 data while protecting citizen privacy, some of the best ranked State governments still struggle to advance properly in providing data protection. Examples are the state of Espírito Santo and specially the Federal government, as we discuss further in this article.

According to OKBR’s study,\(^{53}\) since mid-August all Brazilian State governments have developed their own COVID-19 hotsites as their main active transparency policies, presenting dashboards and, in approximately 70 per cent of the cases, the related microdata. The scenario is a bit different in the State capitals,\(^{54}\) where almost 80 per cent present dashboards, but the same percentage does not publish microdata.\(^{55}\) As a consequence of different state capacities, in local governments, availability and quality of dashboards and databases are deeply impacted by the administration’s ability to access and mobilize financial, human and technological resources.

Regarding passive transparency, which is the duty of the public administration to provide information of citizens requests, a fundamental question that ought to be asked is whether LAI could be used to access personal data processed by the COVID-19 transparency databases. This is a critical matter, since the public information requested via LAI may reveal personal data, which in the context of COVID-19 can be even more sensitive, considering that health data is being processed. In that sense, Green et al. remind us that public records laws (such as LAI) often force cities’ hands into releasing information that might implicate individual privacy.\(^{56}\) In an attempt to avoid some privacy issues, LAI provides for a safeguard in its Article 31,\(^{57}\) requiring that the

---

\(^{48}\) PR Netto, ‘Especialistas criticam medida de Bolsonaro contra Lei de Acesso à Informação’ (UOL, 2020) <https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2020/03/24/especialistas-criticam-medida-de-bolsonaro-contra-lei-de-acesso-a-informacao.html> accessed 19 October 2020.

\(^{49}\) A Pompeu, ‘Moraes suspende MP que autorizava limitação da Lei de Acesso à Informação’ (JOTA, 2020) <https://www.jota.info/stf/do-supremo/moraes-suspende-mp QUE-autorizava-limitacao-da-lei-de-acesso-a-informacao-26032020> accessed 19 October 2020.

\(^{50}\) M Rodrigues, ‘Após Reduzir Boletim Diário, Governo Bolsonaro Retira Dados Acumulados Da Covid-19 Do Site’ (G1, 2020) <https://g1.globo.com/politica/noticia/2020/06/06/apos-reduzir-boletim-governo-bolsonaro-retira-dados-acumulados-da-covid-19-de-site-oficial.html> accessed 17 October 2020.

\(^{51}\) OKBR, ‘Após retrocessos, governo federal volta a abrir dados da Covid-19’ (2020) <https://transparenciacaovid19.ok.org.br/files/Transparencia-Covid19_Boletim_11.pdf> accessed 17 October 2020.

\(^{52}\) OKBR (n 18).

\(^{53}\) OKBR, ‘Após 5 meses,100% dos estados alcançam mínimo ‘Bom’ no Índice’ (2020) <https://transparenciacaovid19.ok.org.br/files/ESTADOS_Transparencia-Covid19_Boletim_4_2.0.pdf> accessed 17 October 2020.

\(^{54}\) OKBR, ‘Transparência desigual de capitais e estados compromete combate à Covid-19’ (2020) <https://transparenciacaovid19.ok.org.br/files/CAPITAIS_Transparencia-Covid19_Boletim_5_2.0.pdf> accessed 17 October 2020.

\(^{55}\) Federal States and State capitals have developed their own databases, which have not been adjusted in order to keep their integrity. Therefore, microdata related to a State capital in a federal State database may diverge from the microdata available in the State capital database.

\(^{56}\) Green and others (n 14) 58.

\(^{57}\) Law no 12,527/2011, Art 31. The treatment of personal information must be done in a transparent manner and with respect to intimacy, privacy, honor and image, as well as individual freedoms and guarantees.

§ 1 Personal information, referred to in this article, related to intimacy, privacy, honor and image:
access to any personal data requested shall be restricted to legally authorized public authorities and its data subject. Exceptions to this restriction are only possible if access to third parties has been provided by law or the data subject has consented to it. Furthermore, with the coming into force of the LGPD, more safeguards are expected to be put into place, such as PbD, which will be later discussed.

Some of the epidemiological data transparency challenges are being overcome with a more robust transparent legislative framework, which includes Brazilian accountability public authorities like the Office of the Controller General (in Portuguese, CGU), the Federal General Accounting Office (TCU) and their subnational representatives, alongside OGP and NGOs transparency initiatives, such as the COVID-19 Transparency Index. However, the development of a data protection framework tends to be more fragile since Brazil still does not have an operative DPA. Hence, establishing data protection benchmarks relies mostly on these same stakeholders’ efforts to communicate with other privacy advocates in order to merge interests and achieve the adequate balance between transparency and data protection to COVID-19 data.

Therefore, in the next section, we present some key concepts of PbD that can be implemented to improve data management on pandemic databases, ensuring transparency while respecting privacy and data protection principles. PbD should be seen as an important element of the concept of transparency by design and the idea that both are key to democracy within the concept of open government.58

One approach to ensure that individual's privacy and governments' transparency comes together: privacy by design

PbD is a framework that prescribes that privacy should be built directly into the design and operation of

information technologies, business practices and network infrastructures.59 Thus, it relates to the idea that data controllers must be proactive in embedding privacy requirements throughout the entire lifecycle of any personal information processing, from data collection until its erasure,60 with special vigour to sensitive data.61 PbD also goes side by side with the notion of Privacy by Default, whereby privacy protections built into the system are activated automatically, and that no action is required from the individual for enhancing her privacy.62

PbD methodologies encompass the embedment of an ethical dimension towards the development of products and services, and relate to the creation of technological measures for ensuring privacy. They are a constant feature of data protection regulations worldwide. In the European General Data Protection Regulation—GDPR, it was established implicitly under the concept of ‘Data Protection by Design’ (DPbD), in its Article 25, whereby the controller shall implement appropriate technical and organizational measures taking into account the state of the art, the cost of implementation, nature, scope, context and purposes, as well as its risks when processing personal data.63 When applied in the fight against COVID-19, the combination of both the ethical approach in PbD and the rule-based one in DPbD are of great importance for ensuring that data made public for transparency and research purposes do not allow for individual identification and discrimination among one’s peers.

Among the safeguards to be adopted, the GDPR mentions pseudonymization and data minimization. Under Article 83, the data protection by design obligation is one of the general conditions for supervisory authorities to assess the controllers or processors’ degree of responsibility when imposing administrative fines.

The LGPD takes a similar approach by stating a PbD-like provision under Article 46. Paragraph 2 specifies that
such safeguards should be adopted throughout the whole lifecycle of the data processing. By its turn Article 52, §1, VIII establishes PbD as one of the criteria that the national DPA will consider when assessing the severity of the sanctions.

Considering that COVID-19 data include sensitive information, it is of utmost importance that the data categories being made public are indeed essential for the purposes of combatting the pandemic. Hence, the idea of protecting privacy throughout the whole life cycle of the data processing should apply from the moment the data is collected and recorded by the health professional until it is made public, for instance, on the national health authority website. In this sense, professionals who are in direct contact with ill individuals should be instructed properly to register only essential data, applying data minimization from the outset of the data processing.

Furthermore, at the individual level, only pseudonymized information should be made available, avoiding references to particular traits of the individual. Health professionals should also promote some sort of anonymization from the start of the processing by applying techniques such as k-anonymity (further described below), to decrease chances of re-identification. Moreover, data should be retained only for the period necessary for the combat of the COVID-19 pandemic.

Ann Cavoukian, who first described the concept of PbD, developed seven principles to guide its implementation. The first establishes that (i) controllers should be 'proactive, not reactive', in ensuring privacy and data protection. The second, named 'Privacy by Default', posits that (ii) personal data should be automatically protected in any given IT system or business practice, without the need of any from the individual to protect her privacy.

Privacy should also be (iii) embedded into a system’s design, as one of its essential components. In addition, (iv) false dichotomies, such as the one that considers privacy and security as irreconcilable, should be avoided, to ensure both privacy and security. Protecting privacy throughout the full lifecycle of data processing by ensuring (v) end-to-end security, (vi) transparency, and by (vii) keeping user-centric solutions are also pivotal requisites for ensuring PbD.

Therefore, following these seven principles of PbD is an important step towards ensuring data protection. Special consideration should be given to end-to-end security, whereby security measures are applied in data processing from start to finish, and that transparency and control are ensured for the data subject, in order to understand clearly how data is being treated.

The ‘six protection goals for privacy engineering’ described by Hansen et al. are effective means to implement PbD appropriately. The first three, ‘confidentiality’, ‘integrity’ and ‘availability’, also known as the ‘CIA triad’ and first developed in the 1980s, are of utmost importance for ensuring information systems’ security. ‘Confidentiality’ relates to secrecy, to the non-disclosure of certain information to certain entities. ‘Integrity’ regards the protection from unauthorised modification of the information processed. ‘Availability’, in its turn, ‘represents the need of data to be accessible, comprehensible, and processable in a timely fashion’. Besides the CIA triad, in order to better ensure privacy and data protection, the authors describe three additional principles: ‘unlinkability’, ‘transparency’, and ‘intervenability’.

64 Law no 13,709/2018, Art 46. Data processing agents must adopt security, technical and administrative measures capable of protecting personal data from unauthorized access and accidental or unlawful situations of destruction, loss, alteration, communication or any form of improper or illicit treatment.

65 Law no 13,709/2018, Art 52. The data processing agents, due to the infractions committed to the norms foreseen in this Law, are subject to the following administrative sanctions applicable by the national authority:

66 EDPS (n 63).

67 It is also important to mention that security is also an essential principle of data protection under the LGPD, as stated in its art 6, VII.

68 EDPS (n 63).

69 Cavoukian (n 61) 2.

70 M Hansen et al., Protection Goals for Privacy Engineering (2015). IEEE CS Security and Privacy Workshops.

71 Ibid 160.
When approaching mechanisms for ensuring privacy-oriented open-data resources, ‘unlinkability’ and ‘transparency’ are key principles to orient systems’ design. The first relates to the system’s ability to prevent that pieces of information be connected to each other in different domains and, lately, to an individual. ‘Unlinkability’ is key for this paper’s purposes, since avoiding that individuals are identified from their health information is an important measure to protect individuals’ privacy when rendering personal data available. In that regard, avoiding that certain data be linked to its subject is essential. Since anonymization consists of processing data so its subject is no longer identifiable, it is a useful means to ensure unlinkability for infected individuals’ data.

On its turn, ‘transparency’, which is also described by both Article 5 of the GDPR and Article 6 of the LGPD as a key data protection principle, prescribes that all personal data processing must be understood and reconstructed at any time by its subject, but also how this treatment may affect one’s individuality. Since their data are being publicized, it is pivotal for individuals to be informed as to how their data are being processed in COVID-19 open databases. Finally, the main objective of ‘intervenability’ is to effectively allow changes and corrective measures to the data being processed, and ‘reflects the individuals’ rights to rectification and erasure of data, the right to withdraw consent, and the right to lodge a claim or to raise a dispute to achieve remedy’.  

Results
Ensuring unlinkability on COVID-19 statistical data through anonymization
As mentioned above, anonymization is one set of measures which allows for unlinkability. Recital 26 of the GDPR defines anonymous information as ‘information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable’. Regarding identifiability, the same recital states that ‘account should be taken of all the means reasonably likely to be used (…) such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments’.

Hence, the GDPR ‘embodies a test based on the respective risk of identification’, an approach which is also taken by the LGPD under its Article 5, III, which defines anonymized data as ‘data related to a data subject who cannot be identified, considering the use of reasonable and available technical means at the time of the processing’.

Anonymization is different from another method mentioned by both LGPD and GDPR, pseudonymization. Article 13, §4, LGPD defines pseudonymization as the ‘processing by means of which data can no longer be directly or indirectly associated with an individual, except by using additional information kept separately by the controller in a controlled and secure environment’.

On its turn, Article 4(5) of the GDPR, defines pseudonymization as the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

Unlike anonymized data, which are considered by both LGPD and the GDPR as non-personal information, pseudonymized data should still be regarded as personal data and thus subject to data protection regulations. However, such data protection legislation treat pseudonymization not as a form of anonymization, but as a means for guaranteeing a greater level of security of processing and PbD (Articles 25(1) and 32(1)(a) of the
GDPR), and thus a form of ensuring compliance which is taken into consideration when applying sanctions by supervisory authorities (Article 83 of the GDPR). Even though the LGPD is not as straightforward as the European framework in this aspect, Article 13 allows for an interpretation which also considers pseudonymization as an essential step for compliance.

Both the European Union and the Brazilian data protection frameworks assume a risk-based approach towards anonymization. This means that anonymized data are the ones that may not lead to the re-identification of the data subject through the use of ‘reasonable and available’ technical measures at the time of the processing.

However, ‘re-identification of individuals is an increasingly common and present threat’, especially when a third party has access to further background information which, through a combination of different datasets, become sufficient to de-anonymize the dataset. On the same direction, Finck and Pallas argue that ‘perfect anonymisation is impossible and that the legal definition thereof needs to embrace the remaining risk’.

Borgesius, Gray, and Echoud also agree with this view, thus explaining that anonymization by itself is not enough to guarantee the balance between open data transparency and privacy. Thus, the scholars have listed four categories of data: (i) raw personal data, (ii) pseudonymised data, (iii) anonymous data, and (iv) non-personal data. Furthermore, some rules of thumb were provided. First, non-personal data can generally be released without restrictions as fully open data while raw personal data should not be released as fully open data. As for pseudonymised and anonymized datasets, they could be disclosed as open data, with some access and reuse restrictions (the former having stricter rules than the latter). No matter how many anonymization techniques have been applied, there is always some risk of re-identification.

In that sense, the UK Information Commissioner’s Office (ICO) argues that there are two main ways for re-identification to come about: (i) ‘if an intruder takes personal data it already has and searches an anonymised dataset for a match’, or (ii) ‘if an intruder takes a record from an anonymised dataset and seeks a match in publicly available information’. The ICO goes on saying that ‘[i]n either case though it can be difficult, even impossible, to assess [identification] risk with certainty’.

Taking into consideration that there are no complete technical solutions to the de-anonymization problem, anonymization should thus be understood as a means of reducing the risks of re-identification of data subjects. With that in mind, the role of the controller becomes to make sure that appropriate measures are adopted in order to render the re-identification process impracticable or extremely expensive. To mitigate risks, it is essential to make sure that only anonymized data which is strictly necessary for a particular purpose is released.

Governments and organizations which aim to establish open-data resources for fighting COVID-19 should develop systems which have in their design mechanisms for lowering as much as possible risks of re-identification. The deployment of the techniques mentioned above will be useful depending on the context: publicizing data from one small municipality, with less than 100,000 inhabitants will require different measures than the ones applied in a city like São Paulo, with over 12 million inhabitants, since the smaller population may provide further challenges to anonymize a given individual among her peers. Therefore, regulators should take into consideration such specificities when creating open-data resources in order to protect individuals’ privacy, as will be further described in this article.

**The Toolbox: a sample of data anonymization techniques for a COVID-19 transparency database**

As explained, many techniques should be implemented together to guarantee anonymization. There is no one-size-fits-all solution, and context is fundamental to achieve the best result in a given case. The purpose of the processing plays an important role since it will delimit the legal basis and which data may be collected, as well as how they should be processed. In this article, we assume that COVID-19 transparency databases, differences apart, have a common goal of ‘processing health data to inform society on the history and current status of the pandemics in a given region’.

---

80 Article 29 Working Party, ‘Opinion 03/2013 On Purpose Limitation’ (2013) 31.
81 UK Cabinet Office, ‘Transparent Government, Not Transparent Citizens - A Report On Privacy And Transparency For The Cabinet Office’ (2011) <http://www.cabinetoffice.gov.uk/sites/default/files/resources/transparency-and-privacy-review-annex-b.pdf> accessed 12 July 2020, 42.
82 Finck and Pallas (n 13) 2.
83 F Zuiderveen Borgesius and others, ‘Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework’ (2015)
In order to assist data controllers in developing PbD solutions, experts and privacy-oriented institutions have created lists of ‘Privacy Patterns’. Some examples are the catalogue of the EU-funded project PRIPARE, a repository created by Vienna’s University of Business and Economics, and an academics collaborative project led by UC Berkeley. These lists present measures which may provide for any of the three goals of privacy protection—unlinkability, transparency, and interventional. As already mentioned, we focus here on anonymization, which aims to achieve the first goal.

Furthermore, DPAs have also put effort into creating guidelines to orient data controllers and other stakeholders on how to implement PbD and anonymization techniques. In this article, we highlight the approaches offered by three institutions: the European Data Protection Board (EDPB), the Spanish DPA—Agencia Española de Protección de Datos (AEPD), and the Singaporean DPA—Personal Data Protection Commission (PDPC). Therefore, the insights here presented are inspired by the guidelines provided by these three organizations.

To start with, we call attention to the structure proposed by AEPD in its ‘Guía de Privacidad desde el Diseño’, which presents eight strategies to achieve privacy protection. From these, the first four aim to guarantee unlinkability: (i) to minimize, (ii) to separate, (iii) to abstract, and (iv) to obfuscate. In the following paragraphs, techniques will be discussed accordingly to how they implement these four strategies.

These techniques can be implemented either in identifier attributes (ie fields that uniquely identify a data subject, such as name, passport number, ID) or in quasi-identifiers (ie fields that, by themselves, do not identify a data subject, but when put together may reveal one’s identity, such as date of birth and zip code).

To minimize
This strategy is tightly connected with the data minimization principle, which is core to many data protection frameworks. It means that any data that is not necessary for the purpose of processing should not be collected, or if wrongly collected, be eliminated from the database, during the data cleaning phase. In the LGPD, this principle can be interpreted from two other principles—adequacy and necessity. When further processing, the controller should periodically consider whether processed personal data still is adequate, relevant and necessary, or if the data shall be deleted or any other anonymization technique should be implemented.

Two common approaches for this strategy are attribute and record suppression. Assuming that, for any reason, the collected data has not passed the necessity test (ie the particular data is not necessary to achieve the purpose of processing), it should be suppressed, as long as this will not affect the accuracy of the results of the database.

While attribute suppression removes an entire category of data from a database (ie an attribute, such as ‘date of birth’ or ‘zip code’, usually a ‘column’ in a spreadsheet), record suppression removes a set of attributes collected from one or more particular subjects/objects (ie the ‘row’ of the spreadsheet).

In the context of a COVID-19 transparency database, while attribute suppression might be an option, record suppression seldom is. In the former, it is clear that several attributes should be suppressed, such as identifying numbers (ie social security, passport, ID) and names.

---

88 ‘Privacypatterns.eu - Collecting Patterns for Better Privacy’ (Privacypatterns.eu, 2020) <https://privacypatterns.eu/?flimit=6&offset=0> accessed 12 July 2020

89 ‘Privacy Patterns Catalog’ (Privacypatterns.wu.ac.at, 2020) <http://privacypatterns.wu.ac.at:8080/catalog/> accessed 12 July 2020

90 ‘Privacy Patterns’ (Privacypatterns.org, 2020) <https://privacypatterns.org/> accessed 29 July 2020.

91 European Data Protection Supervisor (EDPS), ‘Opinion 5/2018: Preliminary Opinion on Privacy by Design’ (2018) <https://edps.europa.eu/sites/edp/files/publication/18-05-31_preliminary_opinion_on_privacy_by_design_en_0.pdf> accessed 12 July 2020

92 Some of the guides mentioned were drafted by the former Article 29 Working Party, and later adopted by the EDPB.

93 AEPD (n60).

94 Agencia Española de Protección de Datos (AEPD), ‘La □ Anonimidad como medida de la privacidad’ (2019) <https://www.aepd.es/sites/default/files/2019-09/nota-technica-anonimidad.pdf> accessed 12 July 2020 3.

95 For example, art 6(II) in the Brazilian LGPD, and art 5(1)(c) in the EU’s GDPR.

96 Law no 13,079/2018, Art 6. Personal data processing activities must observe good faith and the following principles:

---
Also, dates of birth may not be necessary, as long as the age was also obtained (and, as we will discuss later on, the age range).

To separate

This strategy aims to avoid or minimize the profiling of an individual due to the crossing of personal data between two different processing operations within the same organization or between two different ones. The focus here is to create independent processing contexts that hamper the correlation of different data points of the same individual.

Although different approaches may exist to implement this strategy, we highlight pseudonymization techniques, since they are mentioned in some data protection frameworks, such as the EU’s GDPR (article 4(5)) and the Brazilian’s LGPD (article 13, § 4°), and are another way of applying the data minimization principle.

Pseudonymization, which may also be referred to by technicians as coding, is the replacement of identifying data with made-up values. Some pseudonymization techniques may be irreversible, that is, once implemented, the original value cannot be obtained back, or reversible, where the original values are securely kept but can be retrieved and linked back to the pseudonym. For an example of irreversible pseudonymization we cite hash functions, and for the latter, encryption with secret keys (preferably asymmetric).

In the context of COVID-19 databases, pseudonymization may prove to be useful when sharing non-aggregate data between two different institutions (eg a municipality and a federal state). By keeping identifiers within a given organization and sharing only pseudo-identifiers, the risk of identification during data transfers is diminished. Furthermore, even within the same governmental level, data could be pseudonymized, based on the roles of different institutions. For example, while health institutions may need to keep the complete profile of individuals for diagnosis and proper treatment, they may keep their main identifiers (such as name and social security numbers) and only share with the administration pseudo-identifiers. Therefore, several layers of pseudonymization may be applied, one above the other, guaranteeing further protection.

To abstract

Another set of approaches to avoid the easy identification of individuals from data that could not be eliminated is abstraction, which is limiting the level of detail of processed data. This can be achieved by implementing techniques such as aggregation or adding noise.

Data aggregation can be achieved by deliberately reducing the precision of data by using ranges or using conceptual hierarchies. For example, a COVID-19 database could present cases only under 10-years age ranges. Such measure could keep the usefulness of the data allowing to visualise what is the proportionality of elder people being infected (who are supposedly more prone to getting severely ill or dying from the virus) while helping to avoid infected patients from being identified. Another interesting approach would be using location hierarchies: for example, instead of registering datasets with the ZIP code, aggregate addresses to only reveal larger neighbourhoods or municipalities. The aggregation level will depend both on the population’s density in a given region and the number of potential cases detected. According to the EDPB, data aggregation is one way of applying the data minimization principle.

Another abstraction approach is the use of noise, such as adding fictional synthetic data to the database, data perturbation (eg rounding up/down values) and/or swapping some attributes between different records, so as to scramble the rows. Unfortunately, in the context of COVID-19 databases, there is a need for high accuracy on the real number of infected patients and illness, and the use of these techniques may end up spreading false information. Therefore, they should be avoided.

To obfuscate

The fourth strategy to provide unlinkability is obfuscation, that is, limiting the amount of information of a given data, increasing its confidentiality. The aggregation techniques above mentioned also help to obfuscate data. Another technique not yet mentioned is character
masking, which is the change of the characters of a data value (e.g., by using a constant symbol, such as ‘*’ or ‘x’). Masking is typically partial, being applied only to some characters in the attribute.\textsuperscript{109} It is usually applied in non-aggregated data, such as ZIP codes or identification numbers (e.g., Social Security and ID).

In a COVID-19 database, masking may be applied to hide location data such as ZIP code. However, as already mentioned, it should be considered if there is not a better technique that could be applied, such as data aggregation, since solely masking a ZIP code may not be enough for passing in a k-anonymity test.

**Verifying anonymization robustness: K-anonymity and similar methods**

In order to verify whether anonymization is strong enough, it is suitable to apply a field of study known as Statistical Disclosure Control (SDC).\textsuperscript{110} Among the SDC methods, the most common is probably k-anonymity, which is a property of anonymized data that allows quantifying to what extent the anonymity of subjects present is preserved in a dataset where identifiers have been removed.\textsuperscript{111}

Therefore, assuming that a table has been stripped off all its identifiers (either by eliminating, masking them, or applying any other suitable technique as mentioned before), the probability of identifying a specific individual based on a given set of quasi-identifiers (i.e., all the non-uniquely identifiable attributes of a record) is a maximum of 1/k.\textsuperscript{112}

For example, in a COVID-19 database divided by region where on a given day there are 55 infected individuals in Region A, 27 in Region B, and 3 in Region C, and assuming this is the minimum granularity of the chart, we can say the table is 3-anonymized (k = 3). If the same chart now reveals the number of deaths on a given day, and we have 5 deaths in Region A, 3 in Region B, and 1 in Region C, the chart now is 1-anonymized (k = 1).

The issue of a low number of k is that, even if the anonymity techniques implemented have been good enough to avoid singling out (i.e., there are no identifiers in the chart), inferences may still occur: in the example above, we can infer that 1 out of 3 infected individuals of Region C has deceased on that given day.

Usually, in COVID-19 databases, where numbers are quite large, K will probably be high enough to keep inferencing as low as possible, to avoid identification. In order to ensure that, it is fundamental to do a double check on the more granular data revealed in the charts and take into consideration less dense and populous regions. If K is low, some of the above-mentioned techniques not yet implemented should be applied, to raise its value.

As a general rule, it should be decided a minimum size of a given region, where authorities should not share non-aggregated data below that threshold. Furthermore, age ranges and sensitive health data (such as chronic diseases which makes the individual more vulnerable to the coronavirus) should be clustered as much as possible, balancing between the usefulness of the data and the anonymity of individuals.

Two other methods to further guarantee anonymization, and mitigate inference attacks, are l-diversity and t-closeness. L-diversity extends k-anonymity by making sure that in each ‘column’ every attribute has at least ‘l’ different values, while t-closeness goes even beyond, aiming to create ‘columns’ that resemble the initial distribution of attributes in the table.\textsuperscript{113}

In the context of COVID-19 databases of very populous regions, implementing these supplementary methods may not be necessary, since due to the high number of individuals, k may be big enough to guarantee l-diversity automatically. However, as mentioned before, due to the heterogeneity of population within subregions, it may be necessary to apply a limitation on the disclosure of non-aggregated data in less crowded areas.

**Brazil’s COVID-19 Open databases: struggling between transparency and privacy**

Based on the suggestions mentioned above, we finish this study by briefly analysing two COVID-19 Transparency Portals and their respective databases. Due to this article’s size limitations, it will not be possible to go in-depth since this would probably require a specific analysis to identify all the issues within each database.

The first one is the Federal Government online portal ‘COVID-19 Panel’,\textsuperscript{114} which has been updated daily since June 2020. The portal provides data on the total number of cases and deaths and periodicity (weekly and daily). Furthermore, it can categorize the number of cases by Federal States. When downloading the available

---

\textsuperscript{109} PDPC (n 98) 13.

\textsuperscript{110} International Household Survey Network, 'Introduction to Statistical Disclosure Control (SDC)' (2014) <https://www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf> accessed 12 July 2020.

\textsuperscript{111} Accordingly to AEPD, an individual is said to be k-anonymous within the dataset in which it is included if, and only if, for any combination of the associated quasi-identifier attributes, there are at least others k - 1 individuals who share with him the same values for those same attributes. See more on AEPD (n 94) 2–3.

\textsuperscript{112} WP29 (n 102) 17.

\textsuperscript{113} Ibid.

\textsuperscript{114} BRAZIL. COVID-19 Painel Coronavirus <https://covid.saude.gov.br/> accessed 12 October 2020.
CSV spreadsheet, more data can be found, such as the municipality and total population in a given State. However, this spreadsheet provides a mere overview with aggregated data of COVID-19 infections and does not contain microdata. In other words, the level of transparency of the Federal Government database is below what should be expected for open data standards. Microdata can only be found on e-SUS VE and SIVEP Gripe databases at the Open DataSUS portal. These databases, by their turns, provide too many attributes which may easily allow for the identification of individuals, by cross-referencing these data with information available in other public databases.

The second database we have selected for this brief analysis is the one from the State of Espirito Santo, which occupies the first position at Open Knowledge Brazil transparency ranking. At first glance, the data available on the web portal achieves transparency while respecting privacy: the graphs presented cannot be cross-related to further segment a given case, and the division by age is presented in ranges.

However, the available microdata, while not using any direct identifier, may raise some privacy concerns. First of all, the regional granularity is at district level and Espirito Santo is a relatively small state, where some districts have less than 500 individuals. Furthermore, the microdata provides specific age, gender, racial identity, education level and type of clinical assessment for each given case. When cross-relating these data with other governmental databases it is highly probable that some individuals may be identified and can be tracked by health insurance providers, for example.

By briefly analysing the selected microdata structure, the relevance of a Data Protection Impact Assessment (DPIA) becomes evident. According to the LGPD, the ANPD may request a data controller to implement a DPIA, especially when processing sensitive data. Even though the Brazilian DPA is still in its initial steps, it seems to be good practice for a data controller who processes sensitive data to implement a DPIA, regardless of a request. In this sense, it is preoccupying that DATASUS still fails at providing a reliable information management infrastructure and guidance to public health agencies on transparency and data protection.

Conclusions

In this article, we chose Brazil as a case study to propose a model on how to implement transparent COVID-19 government databases while respecting the privacy of individuals.

During the pandemic, multiple legislations were enacted in Brazil, including rules regarding personal health data sharing to control the spread of the virus. Although these provisions aimed to provide more transparency to the pandemic, privacy issues have risen, since some of these legislative measures curtailed the rights to access public information, have not considered data protection safeguards and postponed the applicability of the Brazilian data protection law, the LGPD. Therefore, this article aimed to propose an approach on how to share transparent data on the COVID-19 while ensuring the protection of individual’s privacy. The measures suggested herein referred to a set of anonymization techniques to be implemented on any personal data processing activity related to the pandemic.

This would be the first step to implement PbD, which is much broader than this sole approach. Fully implementing PbD would guarantee respect to the principles of unlinkability, transparency, and intervenability. That would mean preventing individuals’ re-identification, providing transparency regarding personal data processing and ensuring that effective remedies exist to protect their rights.

The PbD measures suggested in this article consist mainly of data anonymization techniques. However, they should not be considered a complete solution to the issue, rather a means of risk reduction, as there is always a risk that anonymized data may still be reidentified. Furthermore, organizational measures, which go beyond the analysis of this article, should also be put in place, such as access control, privacy policies, and data protection impact assessments.

Regarding anonymization, this study highlighted different techniques proposed by some DPAs to implement the unlinkability principle. These techniques may be clustered into four groups:

(i) to minimise data processing of any data that is not strictly necessary for the purpose, including the attribute of suppression for data collected excessively;

115 See footnotes 46 and 47.
116 GOVERNO DO ESTADO DO ESPÍRITO SANTO. Painel COVID-19, Estado do Espirito Santo. <https://coronavirus.es.gov.br/painel-covid-19-es> accessed 12 October 2020.
117 OKBR (n 18).
118 According to IBGE census in 2014, the State of Espirito Santo has 3,88 million residents, which accounts for less than 2% of the country’s population.
119 GOVERNO DO ESTADO DO ESPÍRITO SANTO, Indicadores Socioeconômicos dos Bairros dos Municípios do Estado do Espírito Santo – Censo Demográfico 2010. (2012) <http://www.ijsn.es.gov.br/component/attachments/download/1621> accessed 2 October 2020.
(ii) to separate through the processing of data collected from different contexts in order to reduce the correlation possibilities;

(iii) to abstract by reducing data details (e.g., data aggregation);

(iv) to obfuscate by limiting data processed with aggregation techniques or masking characters of data attributes.

In other words, it is possible to provide public information on health data, such as the ones related to the COVID-19 pandemic, while respecting individual’s privacy, by implementing PbD measures, such as anonymization techniques. However, we state once again, that anonymization methods are not one solution by themselves. Furthermore, finding the right balance between open data transparency and privacy may not prove easy, as it can be seen in the examples of the federal government and the State of Espirito Santo COVID-19 databases.

In the case of Brazil, another relevant issue is the lack of an operative DPA. The existence of a body able to audit and control data processing, especially during the pandemic, would guarantee the enforcement of anonymization techniques and safeguard transparency of public information without violating privacy. Furthermore, financial, technical, and administrative independence of the DPA is equally relevant since it refers to the capability of the audit body to supervise data processing activities made by or shared with public bodies.

The COVID-19 crisis has posed many challenges, but the reinforcement and implementation of data protection frameworks should be listed as a high priority by every country. Treating data anonymization as a tool that allows more transparency is not only a technical issue, but also a political one.

*doi:10.1093/idpl/ipaa025*

*Advance Access Publication 22 April 2021*