







Big Data, Anonymisation and Governance to Personal Data Protection

Artur Potiguara Carvalho
Electrical Engineering Department (ENE), Technology
College, University of Brasília (UnB)
Brasília, DF, Brazil ^a
artur.carvalho@redes.unb.br

Edna Dias Canedo
Department of Computer Science, University of Brasília
(UnB), P.O. Box 4466, CEP 70910-900, Brazil
Brasília-DF, Brazil ^c
ednacanedo@unb.br

Fernanda Potiguara Carvalho
Law School (FD), University of Brasília (UnB)
Brasília-DF, Brazil ^b
fernanda.carvalho@unb.br

Pedro Henrique Potiguara Carvalho
University of Brasília (UnB)
Brasília-DF, Brazil ^d
pedrohpcarvalho@gmail.com

ABSTRACT


In a massive processing data era, an emerging impasse has taking scenario: privacy. In this context, personal data receive particular attention, with its laws and guidelines that ensure better and legal use of data. The General Data Protection Regulation (GDPR) - in the European Union - and the Brazilian General Data Protection Law (LGPD) - in Brazil - lead to anonymisation (and its processes and techniques) as a way to reach secure use of personal data. However, expectations placed on this tool must be reconsidered according to risks and limits of its use, mainly when this technique is applied to Big Data. We discussed whether anonymisation used in conjunction with good data governance practices could provide greater protection for privacy. We conclude that good governance practices can strengthen privacy in anonymous data belonging to a Big Data, and we present a suggestive governance framework aimed at privacy.


CCS CONCEPTS


• **Security and privacy** Domain-specific security and privacy architectures; • **Information systems** Data management systems.


KEYWORDS

Anonymisation; Big Data; Privacy; Governance; Personal Data Protection

^a: <https://orcid.org/0000-0001-7463-1487>.

^b: <https://orcid.org/0000-0003-4934-5176>.

^c: <https://orcid.org/0000-0002-2159-339X>.

^d: <https://orcid.org/0000-0002-0110-4069>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

dg.o '20, June 15–19, 2020, Seoul, Republic of Korea

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8791-0/20/06...\$15.00

<https://doi.org/10.1145/3396956.3398253>

ACM Reference Format:

Artur Potiguara Carvalho, Fernanda Potiguara Carvalho, Edna Dias Canedo, and Pedro Henrique Potiguara Carvalho. 2020. Big Data, Anonymisation and Governance to Personal Data Protection. In *The 21st Annual International Conference on Digital Government Research (dg.o '20)*, June 15–19, 2020, Seoul, Republic of Korea. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3396956.3398253>

1 INTRODUCTION

Data protection is a concern that has become popular [3, 4, 14], with the increasingly common news about data leaks [16, 22]. At the heart of these concerns are personal data, which threatens the privacy of countless people [4, 6, 9, 14, 18, 21, 27, 28, 31].

Thus, several countries have created specific data protection laws and rights guarantees to set limits on the use of personal data. As we focus on the Brazilian scenario, we elected two regulations to guide legal compliance in the use of personal data in this work: the Brazilian General Data Protection Law (LGPD) [5] and the European General Data Protection Regulation (GDPR) [30], as presented in the comparative Table 1. LGPD rose in August 2018, and until the date of publication of this work, it is still in legal vacancy. It was based on European legislation, which was a pioneer in the matter, and whose most recent law is the GDPR. It is possible to highlight that both regulations bring data anonymisation¹ as an effective privacy technique and a way to undo the personal character of the data.

Despite that, anonymisation is not a risk-free mechanism, as experts have increasingly warned it [2, 22, 26, 27]. We address these risks in more depth, both in the background section and in the hypothetical case study proposed by the research. These risks are even more evident in Big Data environments [4, 10, 19, 32].

Big Data is a massive data processing technology [7], while frequently not providing clear guidelines to store data. Often Big Data includes all kinds of personal data, that impact data protection directly [3, 11, 17, 25, 27]. Therefore, precisely because it is a context of a large volume of data, re-identification is facilitated, due to the crossing of information and the possibility of inference. So, in

¹The term is spelled with two variants: "anonymisation", used in the European context; or "anonymization" used in the United States context. We adopt in this article the European variant because the work uses the GDPR [30] as reference.

general, mechanisms such as Big Data need to undergo specific adaptations. That is why, especially in Big Data environments, anonymisation needs ally with other mechanisms for data protection.

Data Governance, in turn, presents itself as a possible ally in favor of privacy [11, 26, 34]. It aimed to promote standardization and quality control in internal data management, ensuring more considerable documentation, organization, and rationalizing costly and expansive data processing. The challenge is to promote greater data protection by mitigating the risks involved in anonymisation when processing data in the context of Big Data. For that, it is necessary identifying whether governance assists in better protection, mitigating the friction between the interests of companies and governments and the interests of privacy and protection of citizens.

To guide this work, we present the Background exploring the risks involved in anonymisation when this tool is used without the assistance of other privacy techniques, and in a Big Data Context. We will take good governance practices as a basis to assess whether their application, in Big Data environments with anonymised data, is consolidated as a factor for better information protection.

The justification for choosing the problem research point out the principles of data governance that can contribute to the reduction of risks that persist after anonymisation. We raised the hypothesis that the application of governance practices, when combined with anonymisation techniques, helps to reduce the risks of re-identification, providing greater protection to personal data. Therefore, governance is an essential ally in risk reduction. The main goal is to verify if the application of anonymisation tools in Big Data context in compliance with data governance can represent a lower risk to data protection.

In section Related Work, we raise the main bibliographical references for the subject to bring a brief discussion about the research's achievement on the subject. In section Results, we present the results obtained in the hypothetical case study, bringing discussions about the risks of anonymisation in Big Data environments. We also present a proposed framework oriented to privacy, which aims to reduce the risks presented. Finally, in section Threats and Validation, we present the limitations of this research and the aspects of validation. We point out as a research method, the literature review, and the study of a hypothetical case. We conclude that data governance can help to reduce the risks inherent to anonymisation techniques in Big Data context. Therefore, we suggest a framework for governance-oriented compliance, with practices geared to the specificities of anonymisation in Big Data. [2, 4, 10, 31, 34].

2 BACKGROUND

In the Brazilian scenario, many organizations have considered anonymisation to be the miraculous solution that will solve all data protection and privacy issues in Big Data [23]. This belief is due to absence of a massive culture of institutionalized data governance [2, 4, 6, 28, 31] (which points out other problems beyond the content of the data that make up the bases); and due to misinformation about the risks that persist after anonymisation.

A report by MicroStrategy in 2019, collecting data from Brazil, Germany, Japan, United Kingdom, and United States, found that only 38% of companies said they retained more than half of their

the controlled data. Only 16% of respondents say their “organization’s analytics technology deployment is at the maturity level to include a sophisticated architecture for self-service analytics with governance, security frameworks, access to Big Data, and mobile and predictive technologies supported by a center of excellence for training and support” [24]. Besides that, according to a Harvard Business Review, published in 2017, only 3% of Companies’ Data Meets Basic Quality Standards [33].

Despite the lack of governance, data processing in the Big Data environment has been growing. In Brazil, about 60% of Brazilian companies already use Data & Analytics to guide strategies and bring about necessary business changes. But, in this country, data governance is still an issue restricted to public entities, financial institutions, and large companies, in most cases. It undermines an efficient compliance of data protection processes and policies into organizations. Also, the Data Protection Regulation in Brazil (LGPD) [5], following the European Regulation (GDPR) [30], does not make clear the need to combine anonymity with governance and not even with other techniques to data protection. This lack does not occur, for example, when it comes to the collection, processing, and sharing of personal data. For these data, the regulations impose solid guidelines for data management.

Both Regulations states important principles of data protection should apply to any information concerning an identified or identifiable natural person (Article 6° [5]; Chapter II [30]). An example is the guiding principle of “data minimisation”, in the European context or “principle of necessity”, in Brazilian context (Article 5°(1)(c)[30]; Article 6° [5], respectively). The principle states that personal data need to be adequate, relevant, and limited to the purposes for which they are processed, that is, data processing should be limited to the minimum necessary to achieve its purposes.

Another principle that guides the processing of personal data is the principle of legitimate interests. It provides that data may be used, taking into consideration the reasonable expectations of data subjects based on their relationship with the controller. (Article 7°, IX, and 10° [5]; Article 6° (1)(f) [30]; Text Preceding GDPR, point 47 [30]). So, legitimate interest links the processing of personal data to the purposes for which it was collected. Therefore, these principles, among others, establish guidelines for the use and impose some limits on the processing of personal data. Its limits both the over-capture of data without a defined purpose, and its processing without legally established guidelines. For this reason, it is practically impossible to think of Big Data involving personal data without a series of management, control, and protection tools. The same does not happen when we talk about anonymous data.

Indeed, these principles are not applied to anonymous data, namely, to personal data rendered anonymous, once that the data subject is not or no longer identifiable. (Article 12° [5]; Text Preceding GDPR, Point 26 [30]). Once anonymised data can be collected, with no worries about maintaining a minimal database, and can be used even for purposes other than the original. So, regulations exploit the concept of anonymous data, as data that can no longer be linked to an identified or identifiable person, to assume that it cannot breach privacy. However, this premise poses some challenges.

We present four characteristics related to anonymisation that pose challenges to the use of this mechanism. The regulations assimilate the first three; however, the fourth challenge is the object of analysis of this research, on which we will focus. First, to be considered anonymous, it should not be possible to identify a person, even with a nameless profile. Regulations capture this situation by adopting the broad concept of personal data, as it identifies or allows one to identify a person. (Article 5°, I, [5]; Article 4°, I, [30]).

The second point of concern is the assumptions about what is considered an identified or identifiable to define anonymous data. Both regulations highlight that all reasonable means must be taken into account, namely, “all objective factors such as costs and the amount of time required must be considered for identification, taking into account the technology available at the time of processing and technological developments” (Article 5°, III, Article 12°, § 1°, [5]; Point 26 [30]). Therefore, even the legal structure, assumes that there is a reasonable margin to consider the data as anonymous, and that, by applying a higher amount of resources, the anonymised data can be re-identified. The third point would be the difficulties of determining the anonymity of a particular piece of overtime. This identification depends on criteria that changes according to technical advances or even by the specific analysis conditions. As mentioned in the previous paragraph, the laws take into account technical developments to determine reasonableness when defining anonymous data. It causes constant uncertainty regarding anonymisation.

Finally, as a fourth point of concern, anonymous data in Big Data Systems have a higher possibility of re-identification. Precisely because Big Data deals typically with massive data, the greater availability of data makes connecting information extremely easy, even when it comes to metadata or fragments of data. Thus, some known anonymisation techniques, such as masking, even when effective in smaller and closed databases, are hardly sufficient in a Big Data context [27]. These four characteristics, especially the last point, converge to the factor that it is not possible to sustain the unexamined belief in anonymisation as a surefire way of ensuring privacy in Big Data contexts, which leads us to the object of the present paper. Since, from the aforementioned regulations, it is inferred that anonymous data can no longer have its subject identified, the use and processing of anonymous data becomes most malleable. In this sense, it is not a legal requirement, the conscious management of these data. But the knowledge about risks involving anonymisation tools lead us to question the sufficiency of this technique for the protection of personal data, especially when dissociated from good management practices. Because it, we investigate the following research question:

RQ.1 Can data governance help to reduce the risks inherent to anonymisation techniques?

We take into account, for the resolution of this issue, the allocation of anonymous data in Big Data. As noted, anonymisation in Big Data involves risks, especially to user privacy. Therefore, we argue that anonymisation must be used with the assistance of other privacy mechanisms, especially with good governance practices aimed at privacy. Considering this, we can define the hypothesis of this research as follows:

HP. 1 The application of anonymisation tools coupled with data governance represents a lower risk to data protection.

Therefore, we list as Main Goal of this paper is to analyze the interaction between data governance and anonymisation and its effects on ensuring the privacy of personal data. We intend to expose privacy threats related to the use of anonymisation in Big Data context and to raise some of the governance characteristics that can assist in mitigating the risks related. We defend that governance is an excellent ally of anonymisation techniques to privacy when using Big Data platforms. The use of privacy-oriented governance can guide the adaptation in Big Data structure that is necessary to guarantee anonymisation as a robust tool for privacy on this bases. Thus, Big Data must also adapt to privacy through the two mechanisms: anonymisation and privacy oriented-governance.

As a research method, we use a literature review, exploring the evolution of the concept, classification, demands, improvements on anonymity. We also explored the weakness of anonymisation in Big Data with a hypothetical case study. Through it, we demonstrate that the main anonymisation techniques suffer an increased risk of re-identification in Big Data environments. Finally, we list some of the privacy-oriented governance practices that can contribute to reducing the weaknesses of anonymisation in massive data systems. We found that anonymisation becomes more fragile if used without the support of governance. We conclude that the concurrent use of these mechanisms strengthens data privacy, in addition to making Big Data Analytics systems more organized, auditable, and transparent, even favoring the business's organizational structure.

2.1 Related Work

In this section, we present a slice of some works in the area, which help us to understand the context involving anonymisation, Big Data and governance, and the relationship of these mechanisms to each other. By chronology, the Big Data Working Group, member of the Cloud Security Alliance [12], in 2013, described the Top Ten Big Data Security and Privacy Challenges, divided into those areas: 1) Secure Computations in Distributed Programming Frameworks; 2) Security Best Practices for Non-Relational Data Stores; 3) Secure Data Storage and Transactions Logs; 4) End-Point Input Validation/Filtering; 5) Real-Time Security Monitoring; 6) Scalable and Composable Privacy-Preserving Data Mining and Analytics; 7) Cryptographically Enforced Data-Centric Security; 8) Granular Access Control; 9) Granular Audits; 10) Data Provenance.

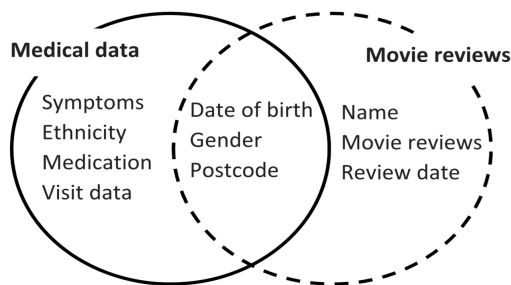
In 2015, H. Liu [21] announced the aspects involved in managing legal frameworks, privacy and security, subject enforcing, and data platforms. In 2016, Farvera and da Silva [6] discussed veiled threats to data privacy in the Big Data Era. In the same year, Mehmood et al. [22] conceptualized the methods and techniques to protection and encryption to data inside Big Data, as well they classified some ways to apply anonymisation. Therefore, we can observe that in this period, there were already studies, to encourage anonymisation in Big Data environments.

According to Mehmood et al. [22], it is possible to highlight two types of data: Personally Identifiable Information (“PII”) and Auxiliary Data (“AD”). The “PII” may include the quasi-identifiers, that is “the attributes that cannot uniquely identify a record by themselves, but if linked with some external dataset may be able

Table 1: Comparative table of regulations

Concepts	GDPR [30]	LGPD [5]
Personal data concept	Article 4(1)	Article 5(I)
Anonymisation concept	Text Preceding GDPR, Point 26	Article 5(III) (IX), Article 12
Exclusion of anonymous data from personal data classification	Text Preceding GDPR, Point 26	Article 12
Processing concept	Article 4(2)	Article 5(X)
Data minimization concept	Article 5(1)(c)	Article 6(III)
Legitimate interests concept	Article 6 (1)(f); Text Preceding GDPR, point 47	Article 7, IX, Article 10

to re-identify the records", therefore contains security liabilities concerning personal data. The Auxiliary Data ("AD") also can reveal the subjects referenced. These two types of data must be handled separately by anonymisation, according to the risks inherent to each. To exemplify that description, Mehmood et al. [22] showed an example (Figure 1) of link quasi-identifiers from records of medical application and movie reviews application.

**Figure 1: Quasi-identifiers and linking records [22]**

Still, in 2016, Lin et al. [20] presented a model considering differential privacy (another way to protect the data privacy). Quoting the weakness of the anonymisation methods, Lin et al. [20] applied the differential privacy to body sensor networks using sensitive Big Data. In their work, Lin et al. [20] combined strategies of anonymisation, aggregation, and Noise Addition strategies - numbers 3 and 4 (Figure 2) to hardening the privacy of a given dataset. But as shown, the scheme adopted by Lin et al. only considers the information given by the internal dataset, ignoring possible attacks using other auxiliaries data available on the Internet, for example. Lin et al. [20] also discussed the risk of data loss through the anonymisation process.

On 27 April 2016, in Brussels, the European Parliament legislated the GDPR. Since then, several studies have sought compliance mechanisms to the new legislation [4, 16, 28]. We highlight the work of Ryan and Brinkley [31], in 2017, that added the critical vision of the organization governance model to address the new protection data regulations issues. About 2017, Chandra and Goswami [32] listed the principal frameworks and algorithms to help to defend the Big Data security, even when the data has been stored outside world (for example, by storage on-demand, or by buying processing power) [32]. In this work, the authors exposed the importance of

data management to extract value from the Big Data environments [32].

Chandra and Goswami [32] also presented the data management as a challenge in Big Data context, because of data huge volume and variety and because its high-velocity intake makes it difficult to validate and process in real-time. In 2018, the legal aspects of personal data protection involving governance was discussed by Ventura and Coeli [34]. Jensen et al. [15] discussed how to get value from Big Data projects, reconciling processing with best data measurement and control practices. The study adds to the arguments of an increased need to manage data processed in Big Data. Still looking for ways to reconcile the extraction of value from Big Data platforms and complying with GDPR, Hintze and Eman [13], defended anonymisation as a possible solution. They advocate the generalized adoption of anonymisation after the processing of personal data, for the permanence of the data with the manager.

Regarding the use of anonymisation in Big Data, Brasher [2] brought some weaknesses of the current process of this tool in massive bases. Brasher's work [2] presents the five most common anonymisation techniques : (1) **Suppression**, (2) **Generalization**, (3) **Aggregation**, (4) **Noise Addition**, and (5) **Substitution**, as shown in Figure 2.

- 1) **Suppression** is the process that excludes any PII from the base.
- 2) **Generalization** shuffles PII identifiers, without excluding any information, reducing their link-ability.
- 3) In **Aggregation**, both data types (PII and AD) go through some reducing treatment that maintains some properties of data (average, statistical distribution, or others at choice) and also reduces their link-ability.
- 4) **Noise addition** adds some non-productive data to confuse the link between PII/AD and their subjects.
- 5) **Substitution**, that is similar to Generalization, while it differs in that: it shuffles not the identifier, but the value of the data itself, replacing the original dataset with other parameters. We can apply this process can to both Personally Identifiable Information and Auxiliary Data [2].

Finally, in 2019, Piras et al. [26], affirmed the importance of data management for extracting value in Big Data and presented the DEFEND, a tool to automate the processes related to compliance with GDPR. In the same year, the Brasher's review was resumed by Domingo-Ferrer [10], who presented the issues of anonymisation and its specificities in Big Data platforms.

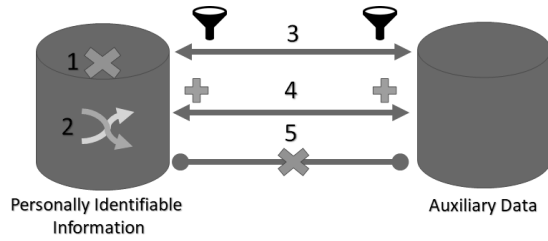


Figure 2: Anonymisation Techniques, adapted from [2, 22]

Domingo-Ferrer criticizes **Suppression** (strategy 1 in Figure 2). According to the author, anonymising data in Big Data is not enough because re-linking the deleted identifiers becomes trivial in this massive context, especially with the inclusion of external data in the analysis. The concerns about the social impact of this insufficient protection are as great as to have surfaced on mainstream media [10]. The author goes on to explain that the efficient privacy protection must consider balancing these two aspects: utility loss and privacy gain of Personally Identifiable Information-based data. Supposed privacy gains occur at the expense of utility loss. When a suppressed piece of data is discarded, less valuable information can be extracted [10]. So, Big Data anonymisation is still limited [10]. Domingo-Ferrer presents three main limitations to current Big Data anonymisation processes:

- 1) Trust in data controllers, granted by Regulations, is undermined by the lack of actionable management criteria for the treatment of confidentiality.
- 2) The utility cost of the process of data anonymisation, which may incur the difficulty of merging and exploring anonymised data.
- 3) The weakness of the anonymisation methods, which satisfy an insufficiently broad set of Statistic Disclosure Controls (SDC).

We will discuss some of these criticisms of Domingo-Ferrer [10] in more depth in the analysis of the hypothetical case proposed by this work. Important to mention that Mehmood et al. [22] and Domingo-Ferrer [10] agree about the trade-off between privacy by anonymisation and utility, and its negative relation mainly in the Big Data context.

Also, in 2019, Mustafa et al. [25] indicate a framework about privacy protection for application on Big Data in the health field. They present the threats of privacy involving medical data in the light of European Regulation.

From the studies presented, it is clear that the discussions involving anonymisation, Big Data, and governance are not recent. The papers point to the risks still present in the processing of anonymised data and alert to the specifics of the Big Data environment. Research also reveals a greater interest in organizing data in Big Data environments, through the construction of governance methods specific to these platforms. The need to better manage data in Big Data environments stems from experiences of less value extraction from disorganized data, the so-called "data swamp". Ventura and Claudia [34] and Jensen et al. [15] point out that, in addition to concerns about extracting value from a large

amount of data, legislative changes have promoted pre-existing discussions, highlighting concerns about the privacy of personal data. Thus, the demands for greater control over the capture and management of data have gained strength.

In this sense, the question raised in this paper is pertinent. Big Data environments have particularities that weaken anonymisation techniques; besides, data are not infallibly anonymised. This way, the use of anonymisation techniques that take into account data governance can hinder the violation of personal data protection?

Based on the concepts presented, and from those discussions, we present the reasons why we believe that governance can be a great ally in solving this problem.

3 RESULTS

To better illustrate the privacy risks that persist in the anonymisation data process, we propose a hypothetical case study, using the main anonymisation techniques presented. We will use a data repository proposal on a Big Data platform whose inserted data represents customers of a financial institution. The hypothetical example will use Big Data because, as already discussed, the large amount of data makes the re-identification of personal data more straightforward, since there is a higher possibility of inferring information and relating data.

In general, companies have customer registration databases that contain significant concentrations of personal data, sometimes even confidential. Besides, in the financial sector, it is possible to identify a customer using other unconventional data (considered quasi-identifiers), such as identity number, social registration, driver's license, bank account number, among others. As we mentioned earlier, both the nominal data and the quasi-identifiers are considered personal data by the reference legislation of this paper [5, 30].

Consider a certain data structure in a Big Data platform according to Figure 3:

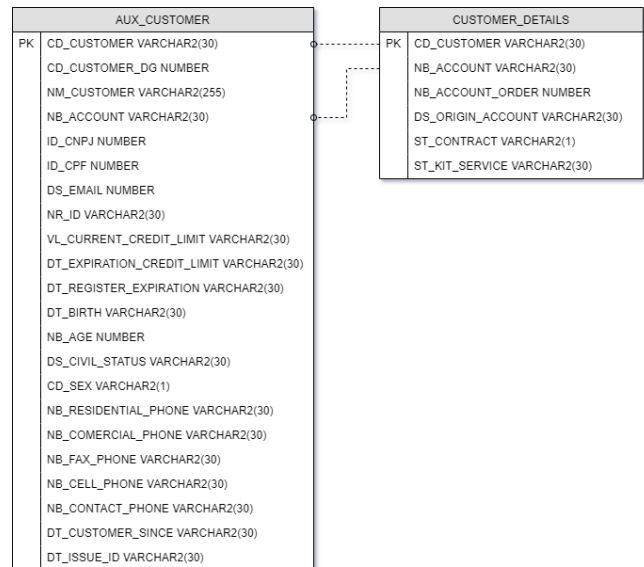


Figure 3: Hypothetical structure data model

This structure is implemented on a Data Base platform, to enable the analysis of the customer (current or potential) characteristics of a certain financial company. This analysis would contain personal data like filters by age, sex, or relationship time with the company and will support several departments in this organization. Also consider the dataset AUX_CUSTOMER and CUSTOMER_DETAIL, which were classified according to the Tables 2 and 3.

Table 2: Attributes/Classification of a example customer table

Personally Identifiable Information (PII) and Auxiliary Data (AD)	COLUMN NAME	DATA TYPE
PII	cd_customer	double
PII	cd_customer_dg	double
PII	nm_customer	string
PII	nb_account	double
PII	id_cnpj	string
PII	id_cpf	string
PII	ds_email	string
PII	nb_id	string
AD	vl_current_credit_limit	double
AD	dt_expiration_credit_limit	string
AD	dt_register_expiration	string
AD	dt_birth	string
AD	nb_age	double
AD	ds_civil_status	string
AD	cd_sex	string
AD	nb_residential_phone	string
AD	nb_comercial_phone	string
AD	nb_fax_phone	string
AD	nb_cell_phone	string
AD	nb_contact_phone	string
AD	dt_customer_since	string
AD	dt_issue_id	string

Table 3: Attributes/Classification of a example customer details Table

Personally Identifiable Information (PII) and Auxiliary Data (AD)	COLUMN NAME	DATA TYPE
PII	cd_customer	double
PII	nb_account	double
PII	nb_account_order	double
AD	ds_origin_account	string
AD	st_contract	string
AD	st_kit_service	string

Now, consider the anonymisation applied by combining the strategies 1-5 described before, according to the showing:

- 1) Using the strategy 1 (**Suppression**): Some registers were excluded from this table.
- 2) Using the strategy 2 (**Generalization**): In another register, the identification was weakened by shuffling the identifier.
- 3) Using the strategy 3 (**Aggregation**): The register with the same ID_CPF was converted to a unique register by the sum between attribute value VL_CURRENT_CREDIT_LIMIT and the max operation over attribute values DT_EXPIRATION_CREDIT_LIMIT, DT_REGISTER_EXPIRATION, NB_AGE and the min operation over attribute values DT_BIRTH, DT_CUSTOMER_SINCE and DT_ISSUE_ID.
- 4) Using strategy 4 (**Noise Addition**): It was included some register with random but full information.
- 5) Using strategy 5 (**Substitution**): It was divided into two registers groups (G1 and G2), and the AD attributes were shuffled between these two groups, preserving the original characteristics.

Based on the difficulty of transforming data privacy governance concepts into operational data protection actions (as described by Ventura and Coeli [34]), suppose that only part of the data in the structure shown by Figure 3 has been classified as identifiable of the respective subject. Only the data contained in the dataset AX_CUSTOMER will be anonymised, excluding the data present in the dataset CUSTOMER_DETAILS.

In the actual production environment, several reasons could lead to the Big Data information not being considered in providing anonymisation. As examples, data governance process failures, misinterpretation of regulation, the shadow in internal defining sensitive personal data, challenges to manage vast and several amounts of data, among others.

Using another dataset (about customer details) from the same schema which was extracted from the previous customer table, it is possible to undo or disturb the anonymisation (weakening the privacy protection) according to the shown:

- 1) Concerning strategy 1 (**Suppression**): The registers excluded were identified (being known as the application of the anonymisation method) by the referential integrity (not explicit) with the table CUSTOMER_DETAIL by the attribute CD_CUSTOMER. Besides, exclusion is the most aggressive strategy that produces the greatest loss of utility.
- 2) Concerning strategy 2 (**Generalization**): Using the attribute NB_ACCOUNT (not search index, but personal data), it was possible to identify the shuffling since this attribute can identify an individual.
- 3) Concerning strategy 3 (**Aggregation**): The presence of the old identifier register in the table CUSTOMER_DETAIL denounces that these registers were manipulated in the original table.
- 4) Concerning strategy 4 (**Noise Addition**): The absence of the register with the old identifier indicates that this register was added to the original table.
- 5) Concerning strategy 5 (**Substitution**): Combining the CD_CUSTOMER and the NB_ACCOUNT from these two tables, it is possible to identify the manipulation of these

data, even if it is hard to define with precision what was modified.

We clarified that all scripts used to create/populate the examples data structures are available below:

```
CREATE TABLE AUX_CUSTOMER(
    CD_CUSTOMER VARCHAR2(30) PRIMARY KEY,
    CD_CUSTOMER_DG NUMBER,
    NM_CUSTOMER VARCHAR2(255),
    NB_ACCOUNT VARCHAR2(30),
    ID_CNPJ NUMBER,
    ID_CPF NUMBER,
    DS_EMAIL NUMBER,
    NR_ID VARCHAR2(30),
    VL_CURRENT_CREDIT_LIMIT VARCHAR2(30),
    DT_EXPIRATION_CREDIT_LIMIT VARCHAR2(30),
    DT_REGISTER_EXPIRATION VARCHAR2(30),
    DT_BIRTH VARCHAR2(30),
    NB_AGE NUMBER,
    DS_CIVIL_STATUS VARCHAR2(30),
    CD_SEX VARCHAR2(1),
    NB_RESIDENTIAL_PHONE VARCHAR2(30),
    NB_COMERCIAL_PHONE VARCHAR2(30),
    NB_FAX_PHONE VARCHAR2(30),
    NB_CELL_PHONE VARCHAR2(30),
    NB_CONTACT_PHONE VARCHAR2(30),
    DT_CUSTOMER_SINCE VARCHAR2(30),
    DT_ISSUE_ID VARCHAR2(30));

CREATE TABLE CUSTOMER_DETAILS(
    CD_CUSTOMER VARCHAR2(30) PRIMARY KEY,
    NB_ACCOUNT VARCHAR2(30),
    NB_ACCOUNT_ORDER NUMBER,
    DS_ORIGIN_ACCOUNT VARCHAR2(30),
    ST_CONTRACT VARCHAR2(1),
    ST_KIT_SERVICE VARCHAR2(30)) ;
```

Note that the data used to detect the anonymisation process belonged to the same data schema as the original database. It is common for data generally to be considered anonymous to their self-platforms. Thus, it is possible that within the Big Data base, there are reliable data to guide the conclusions against anonymisation, as shown in the example. However, anonymity cannot neglect that, in the era of Big Data, a large amount of data is available from other sources. In this context, it would be easier to deduce information through quasi-identifiers, accessible on the internet, social network, another Big Data, or any other external data repository.

In addition to the results presented, we adhere to Domingo-Ferrer[10] criticism, which defines some limitations to current Big Data anonymisation processes, among which we highlight:

- 1) Confidence in data controllers: Legislation presupposes the reliability of controllers. However, in Big Data environments, especially those characterized as "data swamp", the reliability and audit of the data is impaired. In this sense, even if anonymised, the data would not be useful in guaranteeing privacy.

- 2) The cost of anonymising and maintaining anonymised data: In any of the techniques presented, anonymisation implies a reduction in the usefulness of the data. It is the well-known trade-off of loss of data utility in the anonymisation process. It makes anonymisation difficult when data needs to remain intact for business demands. For example, companies that provides personalized services needs to maintain the data users linkability, which ends up making anonymity a challenge.

Also, there is a cost in maintaining the anonymity of the data. As indicated, anonymisation requires continuous improvement of its processing, considering the evolution of the techniques. Therefore, in addition to the costs of storing the data, there are also the costs of remaining anonymous, requiring constant lack of data linkability verification. Its cost of the data anonymisation process can result in the difficulty of merging and exploiting anonymised data.

Having presented the considerations about anonymisation, let us analyze the question raised, that is, whether databases that take into account good governance practices are lower susceptible to the risks presented. The first analysis that we can do on this issue refers to the obstacles presented by Domingo-Ferrer[10]. Is it possible to identify impacts on the use of governance regarding the problems raised by the author? That is what we will check next.

- 1) Confidence in data controllers: As for the reliability of the controllers, governance is presented as a mechanism for streamlining processes, contributing to the transparency and data use audit. In this way, this is a tool to consider objective factors for measuring trust in data controllers. Governance comes to supply the lack of actionable management criteria for the treatment of confidentiality. It contributes to the correct processing of anonymised data, preventing faults from being detected only with data leakage. Also, governance defines the tasks of controllers, establishes accountability criteria for them, and provides for possible institutional sanctions resulting from poor management. On the other hand, the absence of governance contributes to an even greater distrust of the controllers. It is because, without defined guidelines for data management, there is a higher likelihood that massive data storage and processing will generate a real "data swamp", which we mentioned. Besides, the lack of a clear accountability definition for controllers resultant of the misgovernance contributes to an increase in lack of confidence.
- 2) The cost of anonymising and maintaining anonymised data: Governance, as a management tool, can assist controllers in the decision process on what data should be anonymised and what data must be maintained anonymously. Thus, the loss of data usefulness and the cost of maintenance are objective factors that must be taken into account in management. On the other hand, without efficient data management, the difficulties in choosing and measuring these costs increase considerably. So, knowing the data that make up the bases, therefore, becomes a differential factor.

Other factors of interaction between anonymity and governance can also be identified based on the expected characteristics of the

application of good practices. Greater knowledge about the data that make up Big Data can guide the establishment of criteria for data processing, since, due to the volume, sometimes the data cannot be processed in its entirety. Governance can also assist in the elimination of redundant data, reducing processing and storage costs, in addition to the risks of re-identification from remaining non-anonymised data. It can also guide the anonymisation process itself, taking into account other data platforms, such as public data, which can serve as a basis for re-identifying information.

Due to these characteristics, some authors such as Piras et. al. and Chandra and Goswami [26, 32], defends that the data privacy governance can be an ally to react to misuse of citizen data and lack of control over management and privacy issues of citizen data. Due to the risks presented, we understand that even anonymised data can be favored with the application of good practices. Also, data management appears on Chandra and Goswami [32] as an aspect relevant to the security of Big Data environments. In fact, data management is a powerful tool to hardening the data protection and data privacy on the Big Data environment. In this sense, we propose a possible framework to guide data governance to fulfill the issues of data privacy left of the applied anonymisation techniques, as shown in Figure 4.

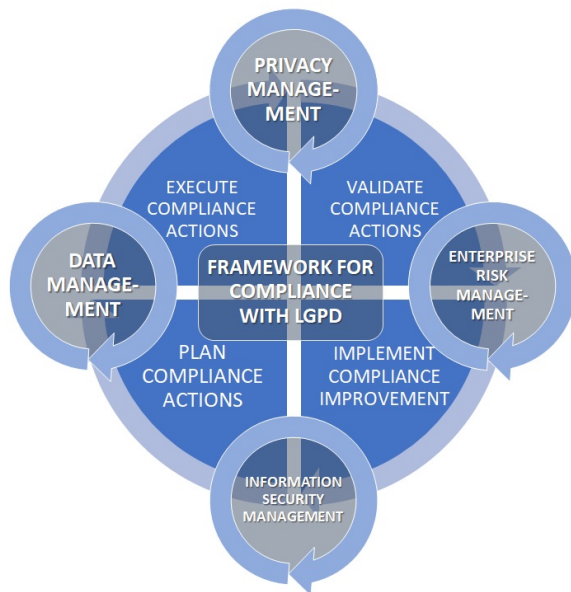


Figure 4: Hypothetical data management framework

The example of a framework (Figure 4) is composed of a Deming cycle (inner) [8] of actions for compliance (Plan, Execute, Validate, Improve compliance) rounded by the supports disciplines (Privacy Management, Data Management, Information Security Management, and Enterprise Risk Management). Each one of these components works together with the compliance to the Data Protection Law.

Expanding the discipline data management, we have the 10 functions proposed for Brackett and Earley [1], from the Data Management Body of Knowledge (DMBOK), according to the Figure 5.



Figure 5: Data management functions related to anonymisation [1]

Rego [29] describes that functions (Figure 5) according the following:

- **Data Management:** Represents the exercise of authority and control of strategy, policy, rules, procedures, roles, and tasks involved with data assets. This function is centered on this framework and influences all the others.
- **Data Architecture:** Defines the data needs (usually corporate) of the company, in addition to creating and maintaining the Corporate Data Architecture considering the company's strategic objectives.
- **Data Security:** Responsible for defining and maintaining security policies and procedures to provide adequate authentication, use, access, and audit.
- **Master Data Management:** Responsible for defining and controlling tasks to ensure the consistency and availability of unique views of the company's master and reference data.
- **Data Warehousing and BI:** Defines and controls processes to provide decision support data, generally available in analytical applications.
- **Documents and Content:** It is dedicated to planning, implementing, and controlling activities to store, protect, and access the company's unstructured data.
- **Metadata:** Responsible for manage and store the company's metadata, in addition to enabling forms of access.
- **Data Quality:** It is dedicated to managing tasks for technical application of data quality to measure, evaluate, improve, and ensure the company's data quality.
- **Data Modeling and Project:** Responsible for creating and maintain data models to avoid redundancy of information in the organization.

- **Data Storage:** It is dedicated to managing the infrastructure of data, valuing the adequate availability and performance of data queries.
- **Data Storage:** Managing and maintain the platform of data interoperability, controlling the flow of data between bases.

Of these functions, we separate six main that help the privacy protection of anonymised data from the hypothetical case. Are they: Metadata, Master Data Management, Data Integration, Data Security, Data Architecture and Data Modeling and Project. The functions Data Architecture and Data Security supports by default the process of anonymisation throughout the whole data life cycle (including, being able to comprise this tool explicitly in its processes). Another function will be detailed in the following explanation of the hypothetical example cited. To hardening the weaknesses found with the introduction of dataset customer details, consider the application of this theoretical framework. The possible benefits are showed forward:

- 1) Concerning the issue of strategy 1 (**Suppression**): The Metadata can provide some information about data lineage, the same as the Master Data Management. These two functions support the identification of the multiple data reference, even if it does not explicit on the base (like a missing reference constraint). Applying these functions of the data management, the exclusion of some registers in the main dataset with no corresponding exclusion in the details dataset would be detected and treated. Even so, with a reasonable dose of knowledge about the information present in the dataset customer (made possible by the Data Management, Data Modeling and Project and Metadata functions of Data Management), it would be possible to identify a data exclusion that would not affect the actual use of the information (if such exclusion existed).
- 2) Concerning the issue of strategy 2 (**Generalization**): The same that the item 1, the functions Metadata, Master Data Management, and Data Modeling and Project will provide inputs to identify all personal information (allied to the privacy management processes), and treat them. One of the artifacts generated by the Metadata function is the Data Catalog, where all these classifications (personal data, sensitive data, corporate data) are stored and continuously reviewed. It is a powerful method to control the constraints required for the LGPD [5] and GDPR [30] compliance.
- 3) Concerning the issue of strategy 3 (**Aggregation**): Mainly, the function Master Data Management reduces the redundancy of identifiers through the base. Perhaps, the application of this function minimizes the issue resultant of this technique.
- 4) Concerning the issue of strategy 4 (**Noise Addition**): If the application of noise addition respects the earlier step of data lineage, which will list all data related to this dataset (and its relations), the noise addition will be enforced and hard to identify.
- 5) Concerning the issue of strategy 5 (**Substitution**): The substitution which considers the Metadata and Master Data to update the data when necessary is the best form of anonymisation even in the Big Data context and considerably hard

to identify. The result of this anonymisation is quite similar to the productive data, strengthening protection.

4 THREATS AND VALIDATION

4.1 The Hypothetical case

We presented the risks involved anonymisation, in the Related Work section, and exemplified it in the Partial Results section. However, it is possible to identify threats in the hypothetical case study presented. The first threat reported is that the hypothetical case study did not set out to state that anonymisation failed in all its techniques, taking re-identification as always certain.

In some of the anonymisation techniques used, re-identification is a clear possibility in the comparative analysis with the table CUSTOMER_DETAILS, as occurred when attributes have been shuffled. But, in general, it was possible to conclude at least the existence of data processing, but not the complete re-identification, at less in this hypothetical case study. For example, when deleting data, comparison with the CUSTOMER_DETAILS table reveals that information has been suppressed. It means that the use of anonymisation is clear from a simple comparison with a table within the same database. It is true even with suppression, which is the most aggressive anonymisation technique.

So, it is possible to reveal which data has been modified, deleted, or shuffled, and provides a remnant base that maintains its integrity and can be used. Also, it provides information to complete or organize all bases through external reinforcement, as with a public base, as mentioned. Despite the threat presented, we emphasize that the objective of the hypothetical case study was not to present the complete fallibility of anonymisation techniques. The aim is to show the risks that Big Data environments provide to these techniques, mainly if we have a lack of data governance.

Knowing which data has been anonymised significantly weakens database protection. It is because data that has not undergone the anonymisation process, for example, or data that is reorganized within the platform, will constitute a remnant base that maintains its integrity. Thus, unchanged data is known to be intact and can be used to obtain relevant information. Obtaining such secure information besides being possible, it is likely, mainly in the context of Big Data, considering large databases that are stored without effective governance. The lack of management makes the leakage of this data risky, which serves as a subsidy for obtaining information.

The second threat is the fact that we do not consider the use of more than one anonymisation technique on the same data to assess the possible risks. We emphasize, again, that the objective of this work is not to disqualify anonymisation techniques, but pointing out their risks. In this sense, we believe that the combination of techniques, while reducing the risks of re-identification, does not eliminate them. Furthermore, we understand that the combination of techniques is normally used by databases concerned with governance, to identify which anonymisation techniques used together would be able to promote privacy while preserving the value of the data. It is because, usually, the association of anonymisation techniques implies a more significant loss of utility of the data than the exclusive use of one of the techniques.

4.2 Framework

Regarding the threats involved in the proposed structure, we highlight that this governance model has not been tested, and its validation is hypothetical. We seek to present a governance structure concerned with the privacy of anonymised data, based on the remaining risks to the application of anonymity presented in the related work section and the hypothetical case study. Therefore, it is possible that, in the application of the framework in a practical case, specificities not foreseen in this work arise. The proposal was only to introduce the need for governance and present a guiding model for these structures.

5 CONCLUSIONS

In the context of Big Data, even anonymous data does not ensure privacy without the support of other techniques. As highlighted, the tool has internal limits when exposed to a massive data stored. The expectations placed on this tool should be reconsidered according to the risks and limits of its use. In this sense, we seek to demonstrate why we believe that data governance has an essential role in measuring these risks and in managing the weaknesses of this tool.

Meantime, both GDPR [30] and LGPD [5] describe anonymisation as a useful technique for data protection, without, however, guiding that the use of these techniques must accompany good institutional governance of the data as a whole. These legislations were concerned with aspects of governance only for personal data. It excluded anonymised data, considered non-personal, from these concerns. The situation worsens by the absence of a widespread data governance culture.

Therefore, it is clear that anonymisation is not sufficient to reconcile Big Data compliance with Personal Data Regulations and data privacy if applied isolatedly. It does not mean that anonymisation is a useless tool, but it needs to be applied with the assistance of mechanisms developed by compliance-oriented governance. Besides, a Big Data-Driven framework is required to recommend best practices that, coupled with anonymisation tools, ensure data protection in Big Data environments and address this compliance issue.

To address these gaps, we propose a framework that can drive the best governance of anonymised data in a Big Data environment, focused on privacy. The model aims, through the compatibility of the two tools, providing greater security for anonymisation, and bringing greater organization and transparency to data management.

REFERENCES

- [1] Michael Brackett and Susan Earley. 2009. *The DAMA Guide to The Data Management Body of Knowledge (DAMA-DMBOK Guide)*. Vol. 2. DAMA, The Sun Building, New York, NY 10007, EUA. 406 pages.
- [2] Elizabeth A Brasher. 2018. Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation. *Colum. Bus. L. Rev.* 1 (2018), 209.
- [3] Maja Brkan. 2019. Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International journal of law and information technology* 27, 2 (2019), 91–121.
- [4] Pompeu Casanovas, Louis De Koker, Danuta Mendelson, and David Watts. 2017. Regulation of Big Data: Perspectives on strategy, policy, law and privacy. *Health and Technology* 7, 4 (2017), 335–349.
- [5] Presidência da República. 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). *Secretaria-Geral, Accessed in November 19, 2019* 1 (2018), 31. <https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf>.
- [6] Rafaela Bolson Dalla Favera and Rosane Leal da Silva. 2016. Cibersegurança na União Europeia e no Mercosul: Big Data e Surveillance Versus Privacidade e Proteção de Dados na Internet. *Revista de Direito, Governança e Novas Tecnologias* 2, 2 (2016), 112–134.
- [7] Andrea De Mauro, Marco Greco, and Michele Grimaldi. 2016. A formal definition of Big Data based on its essential features. *Library Review* 1 (2016), 122–135.
- [8] W Edwards Deming. 1993. The new economics for industry. *Government, Education, Massachusetts Institute of Technology, Cambridge, MA* 1 (1993), 235.
- [9] Edna Dias Canedo, Angelica Toffano Seidel Calazans, Eloisa Toffano Seidel Masson, Pedro Henrique Teixeira Costa, and Fernanda Lima. 2020. Perceptions of ICT Practitioners Regarding Software Privacy. *Entropy* 22, 4 (2020), 429.
- [10] Josep Domingo-Ferrer. 2019. Personal Big Data, GDPR and Anonymization. In *International Conference on Flexible Query Answering Systems*. Springer, Am Thalbach 22, 4600 - Thalheim bei Wels - AUSTRIA, 7–10.
- [11] Dr B Fothergill, William Knight, Bernd Carsten Stahl, and Inga Ulmican. 2019. Responsible Data Governance of Neuroscience Big Data. *Frontiers in neuroinformatics* 13 (2019), 28.
- [12] Cloud Security Alliance Big Data Working Group et al. 2013. Expanded top ten big data security and privacy challenges. *White Paper, Apr* 1 (2013), 39.
- [13] Mike Hintze and Khaled El Emam. 2018. Comparing the benefits of pseudonymisation and anonymisation under the GDPR. *Journal of Data Protection & Privacy* 2, 2 (2018), 145–158.
- [14] Dominik Huth, Laura Stojko, and Florian Matthes. 2019. A Service Definition for Data Portability. In *21st International Conference on Enterprise Information Systems*, Vol. 2. Springer, Avenida de S. Francisco Xavier, Lote 7 C.v. C, 2900-616 Setubal - Portugal, 169–176.
- [15] Maria Hoffmann Jensen, Peter Axel Nielsen, and John Stouby Persson. 2018. MANAGING BIG DATA ANALYTICS PROJECTS: THE CHALLENGES OF REALIZING VALUE. In 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden. *Managing Big Data Analytics Projects: the Challenges of Realizing Value* 1, 15.
- [16] Daniel Joyce. 2017. Data associations and the protection of reputation online in Australia. *Big Data & Society* 4, 1 (2017), 2053951717709829.
- [17] Maria Koutli, Natalia Theologou, Athanasios Tryferidis, Dimitrios Tzovaras, Aimilia Kagkini, Dimitrios Zandes, Konstantinos Karkaletsis, Konstantinos Kaggelides, Jorge Almela Miralles, Viktor Oravec, et al. 2019. Secure IoT e-Health Applications using VICINITY Framework and GDPR Guidelines. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. IEEE, Petros M. Nomikos A.E., 29 Vas. Sofias Av. 10674 Athens, Greece, 263–270.
- [18] Huang Lanying, Xiong Zenggang, Zhang Xuemin, Wang Guangwei, and Ye Conghuan. 2015. Research and Practice of DataRBAC-based Big Data Privacy Protection. *Open Cybernetics & Systemics Journal* 9 (2015), 669–673.
- [19] Lixiang Li, Kaoru Ota, Zonghua Zhang, and Yuhong Liu. 2018. Security and privacy protection of social networks in big data era. *Mathematical Problems in Engineering* 2018 (2018), 0–2.
- [20] Chi Lin, Pengyu Wang, Houbing Song, Yanhong Zhou, Qing Liu, and Guowei Wu. 2016. A differential privacy protection scheme for sensitive big data in body sensor networks. *Annals of Telecommunications* 71, 9–10 (2016), 465–475.
- [21] H Liu. 2015. Visions of Big Data and the risk of privacy protection: A case study from the Taiwan health databank project. *Annals of Global Health* 1, 81 (2015), 77–78.
- [22] Abid Mehmood, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, and Song Guo. 2016. Protection of big data privacy. *IEEE access* 4 (2016), 1821–1834.
- [23] Vinicius Mendes. 2019. Empresas afirmam que anonimização de dados pode ser solução mais rápida para adequação a LGPD. <https://www.olharconceito.com.br/noticias/exibir.asp?id=17994¬icia=empresas-afirmam-que-anonimizacao-de-dados-pode-ser-solucao-mais-rapida-para-adequacao-a-lgpd> (Date last accessed 20-April-2020).
- [24] Microstrategy. 2020. 2020 GLOBAL STATE OF ENTERPRISE ANALYTICS. <https://www.microstrategy.com/getmedia/db67a6c7-0bc5-41fa-82a9-bb14ec6868d6/2020-Global-State-of-Enterprise-Analytics.pdf> (Date last accessed 20-April-2020).
- [25] Uzma Mustafa, Eckhard Pflugel, and Nada Philip. 2019. A Novel Privacy Framework for Secure M-Health Applications: The Case of the GDPR. In *2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGS3)*. IEEE, Northumbria University London, 110 Middlesex Street, London, England, E1 7HT, 1–9.
- [26] Luca Piras, Mohammed Ghazi Al-Obeidallah, Andrea Praitano, Aggeliki Tsohou, Haralambos Mouratidis, Beatriz Gallego-Nicasio Crespo, Jean Baptiste Bernard, Marco Fiorani, Emmanouil Magkos, Andres Castillo Sanz, et al. 2019. DEFEND Architecture: A Privacy by Design Platform for GDPR Compliance. In *International Conference on Trust and Privacy in Digital Business*. Springer, Am Thalbach 22, 4600 - Thalheim bei Wels - AUSTRIA, 78–93.
- [27] Alexandra Pomares-Quimbaya, Alejandro Sierra-Múnera, Jaime Mendoza-Mendoza, Julián Malaver-Moreno, Hernán Carvajal, and Victor Moncayo. 2019. Anonymity: From a Small Data to a Big Data Anonymization System for Analytical Projects. In *21st International Conference on Enterprise Information Systems*. Springer, Avenida de S. Francisco Xavier, Lote 7 C.v. C 2900-616 Setubal - Portugal

- 38.524098, -8.905325, 61–71.
- [28] Christina Popovich, Francis Jeanson, Brendan Behan, Shannon Lefaiivre, and Aparna Shukla. 2017. Big Data, Big Responsibility! Building best-practice privacy strategies into a large-scale neuroinformatics platform.
 - [29] Bergson Lopes Rêgo. 2013. *Gestão e Governança de Dados: Promovendo dados como ativo de valor nas empresas*. Brasport, Rua Pardal Mallet, 23 - Tijuca, 20270-280 Rio de Janeiro-RJ.
 - [30] General Data Protection Regulation. 2018. EU data protection rules. , 1821–1834 pages. https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.
 - [31] Melissa Ryan and Mark Brinkley. 2017. Navigating privacy in a sea of change: new data protection regulations require thoughtful analysis and incorporation into the organization's governance model. *Internal Auditor* 74, 3 (2017), 61–63.
 - [32] K. Sushma, C. Naidu, Y. Sai, and K. Chandra. 2017. Design and Implementation of High Performance MIL-STD-1553B Bus Controller. In *2017 IEEE 7th International Advance Computing Conference (IACC)*. IEEE Computer Society, Los Alamitos, CA, USA, 266–269. <https://doi.org/10.1109/IACC.2017.0065>
 - [33] Thomas C. Redman Tadhg Nagle and David Sammon. 2020. Only 3% of Companies' Data Meets Basic Quality Standards. <https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards> (Date last accessed 20-April-2020).
 - [34] Miriam Ventura and Cláudia Medina Coeli. 2018. Para além da privacidade: direito à informação na saúde, proteção de dados pessoais e governança. *Cadernos de Saúde Pública* 34 (2018), e00106818.