

Integrating LGPD Requirements and Restrictions into Database Design

Patricia Vieira da Silva Barros¹, José Maria Monteiro¹, Javam Machado¹

¹LSBD – Computer Science Department – Federal University of Ceará

{patricia.barros, jose.monteiro, javam.machado}@lsbd.ufc.br

Nível: Doutorado

Ingresso: Março de 2020

Previsão de Término: Fevereiro de 2026

Etapas já concluídas: Revisão Bibliográfica Preliminar, Definição do Problema e Defesa de Qualificação

Defesa da Pré-Proposta: Dezembro de 2024

Defesa da Proposta: Fevereiro de 2025

Lista de Publicações:

- Estratégias para Modelagem e Avaliação da Conformidade entre Sistemas de Informação e a Lei Geral de Proteção de Dados Pessoais (LGPD) - SBSI 2023.
- Incorporando os Requisitos e as Restrições da LGPD ao Projeto de Banco de Dados - SBBDD 2024 (Full Paper).
- brModeloPD: Uma Extensão da Ferramenta BrModelo para Incorporar os Requisitos e as Restrições da LGPD ao Projeto de Banco de Dados - Demos 2024.

Abstract. *The Brazilian General Data Protection Law (LGPD) specifies how the processing, storage, and disposal of personal data should be conducted, conditioning it to the prior authorization of the data subject. On the other hand, current information systems are heavily reliant on the use of personal data and therefore need to comply with the LGPD. However, the methodologies and tools used for database design do not incorporate the requirements and constraints of the LGPD, making it difficult to ensure compliance between databases and current legislation. This article presents a methodology, called LGPDbyD, to incorporate the impositions and principles of the LGPD into the design of databases. To achieve this, we adapted the ER model, the Relational model, and the CREATE TABLE command.*

Resumo. *A Lei Geral de Proteção de Dados Pessoais (LGPD) tem por finalidade determinar como deve ser realizado o tratamento, o armazenamento e o descarte de dados pessoais, inclusive nos meios digitais, sempre com autorização prévia do concedente. Neste contexto, o sistema de bancos de dados passa a ser um componente ainda mais importante no desenvolvimento de software, uma vez que este é responsável pelo armazenamento, atualização e recuperação dos dados. Contudo, as metodologias e ferramentas utilizadas para o projeto de bancos de dados não incorporam os requisitos e as restrições da LGPD, dificultando assim a conformidade entre os bancos de dados e a legislação vigente. Este artigo apresenta uma estratégia, denominada LGPDbyD, para incorporar as imposições e preceitos da LGPD ao projeto de bancos de dados. Para isso, adaptamos o modelo ER, o modelo Relacional e o comando CREATE TABLE.*

1. Introduction

The General Data Protection Law (LGPD)¹ regulates how companies can use personal data as information related to an identified (or identifiable) natural person, in addition to determining how personal data should be processed, stored, and discarded, aiming to protect the fundamental rights of privacy and freedom. LGPD is a legislation that encompasses a change in processes, updating of documents and contracts in organizations, and mainly, a change in the culture of the day-to-day activities of companies, in the way they process personal data. The Brazilian law innovates by addressing aspects not mentioned by other data protection laws existing in Brazil, such as, for example, providing a more precise definition of the concept of personal data, an express provision of the legal bases that authorize the processing of such data, care in the processing of public data, the creation of the ANPD (National Data Protection Authority), the definition of sanctions, thus providing greater legal certainty to holders of personal data.

On the other hand, as current Information Systems (IS) are heavily based on the acquisition, storage, and processing of personal data, these systems must comply with the General Data Protection Law (LGPD). This law has a significant impact on the development of these systems, as they must now handle personal data in the manner stipulated by the legislation, in a more formal way, paying greater attention to the data life cycle. This cycle encompasses all operations performed on the information obtained by a company or institution, from its collection to its proper destruction.

In this context, the database system (DBS) becomes an even more important component in software development, as it is responsible for data storage, updating, and retrieval. More specifically, a database (DB), which represents a set of data and their inter-relationships, must comply with the General Data Protection Law (LGPD). For example, the result of an HIV test (which is used to diagnose an infection caused by the human immunodeficiency virus) should be stored in an encrypted form, making it inaccessible even to database professionals and system developers.

One of the alternatives to ensure database compliance with the LGPD is to incorporate the requirements and restrictions imposed by the legislation into the database design process. However, database design is a complex activity that involves four distinct phases: i) requirements gathering and analysis, ii) conceptual design, iii) logical design, and iv) physical design. Moreover, these phases produce various artifacts, use different notations, and are often supported by different software tools. Additionally, existing methodologies for database design do not incorporate the requirements and principles brought by the LGPD.

This article presents a strategy, called LGPDbyD, to incorporate the requirements and restrictions imposed by the LGPD into the database design process. To achieve this, we adapted the ER model, the Relational model, and the CREATE TABLE command. Additionally, we extended the brModelo tool to support the methodology proposed in this work. LGPDbyD aims to facilitate the processes of database design and auditing in compliance with the LGPD.

¹https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm

2. Related Work

[Khan et al. 2004] addresses the importance of integrating business requirements and constraints into conceptual database models, emphasizing the importance of effective communication between database developers and business stakeholders. This ensures that business requirements are understood and correctly translated into the database model. In this way, organizations can develop systems that are more aligned with their specific needs, resulting in greater efficiency, flexibility, and user satisfaction.

[Kamble 2008] proposes a conceptual model for dealing with data that have multiple dimensions, offering a structure that allows the representation, organization, and analysis of such data, in addition to understanding how different dimensions relate to each other and how to extract information from these relationships. The work proposed in [Dani and Getta 2005] presents a methodology and symbology for conceptual modeling focused on *Data Streams*, addressing the challenges of modeling in computing with continuous data flow, aiming to improve the efficiency and effectiveness of *Data Streams* processing in real-time.

In [de Abreu et al. 2021], the authors discuss how to ensure that query processing in databases respects user consents. The proposed solution is based on the development of SQL language extensions aimed at incorporating consent considerations during query processing. This allows developers to explicitly express in SQL commands the conditions under which data can be accessed and used.

In [Sarkar and Athanassoulis 2022], the authors present an extension for query languages that allows specifying data deletion policies. Thus, developers can define rules for the automatic deletion of data directly in an SQL command (generally in the INSERT clause) based on criteria such as the period of time the data remains stored. For example, when inserting a particular "tuple," it is possible to define that it should be automatically removed after five years. This strategy is of fundamental importance in contexts where privacy and compliance with different regulations are significant concerns.

In [Carvalho et al. 2023], an extension to the Entity Relationship (ER) model (ER+) is presented, which provides a more suitable framework for modeling distributed systems with multiple layers. The authors discuss how this new extension addresses scalability and performance issues in distributed systems, including techniques to distribute the workload among the different nodes of the system, optimize communication between layers, and ensure data consistency in a distributed environment.

In [Shastri et al. 2019], the authors seek to understand how SBDs may be affected by GDPR (General Data Protection Regulation), proposing a series of metrics to evaluate the performance of SBDs regarding GDPR compliance, such as query execution time, resource consumption, and the effectiveness of anonymization and pseudonymization techniques. Experiments were conducted using different types of databases and workloads to validate the proposed approach.

3. The Proposed Strategy: LGPDbyD

The proposed strategy, called LGPDbyD, aims to incorporate the requirements and restrictions of the LGPD into database design, adding small adaptations to the concepts and notations commonly used in conceptual, logical, and physical designs. The central idea is to change existing models as little as possible.

3.1. Conceptual Design

Initially, we propose an adaptation of the Entity-Relationship (ER) model, called ER-PD, with the aim of enabling the conceptual design of databases in compliance with the General Data Protection Law (LGPD). This adaptation allows the representation of key concepts present in the LGPD, such as personal data, consent, the type of processing to be performed on the data, and the data subject.

The Data Subject, as specified by the LGPD in its Article 5, refers to the individual whom the law intends to protect. Therefore, the Data Subject is a central concept in the ER-PD model, represented as a specific type of entity set that indicates the presence of personal attributes, which must be particularly protected. The notation used to represent this specific type of entity set, called “Subject,” is a rectangle with dashed lines (Figure 1).

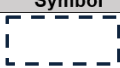











Symbol	Representation	Symbol	Representation
	Owner Entity		Consent Period Attribute
	Personal Attribute		Purpose Attribute
	Sensitive Attribute		Shared Attribute
	Anonymized Attribute		Child and Adolescent Attribute
	Encrypted Attribute		Identifier Attribute
	Consent Attribute		Semi-Identifier Attribute

Figure 1. ER-PD Model Notation.

According to the LGPD, some personal data are considered “sensitive” and must be treated with specific care, as highlighted in its Article 11. Additionally, one of the methods for handling sensitive data is anonymization. Although encryption is not explicitly mentioned in the LGPD, it is a commonly used alternative to ensure data anonymization. To represent the fact that an attribute stores personal data and the type of processing that must be performed on that data, the ER-PD model proposes the use of 11 new types of attributes (Figure 1), the main ones being: “personal attribute” (P), “sensitive attribute” (S), “anonymized attribute” (A), and “encrypted attribute” (C). Additionally, the ER-PD model also introduces two other new types of attributes: “identifier attribute” (I) and “semi-identifier attribute” (SI), to represent concepts commonly used in the field of data privacy. An Identifier attribute is one that can uniquely identify individuals. A Semi-Identifier attribute does not explicitly identify an individual but, when combined with other attributes, can enable the identification of an individual.

Besides, LGPD establishes specific rules for the processing of data related to Children and Adolescents, as outlined in Article 14 and its paragraphs. To represent this characteristic, the ER-PD model introduces a new type of attribute called “Child and Adolescent”. Additionally, the data subject can authorize their data, or part of it, to be shared with third parties. To represent this requirement, the ER-PD model proposes a new type of attribute called “Shared”. Figure 1 illustrates the notation added by the ER-PD model.

3.1.1. Running Example

Next, we will illustrate the use of the ER-PD model in the conceptual design of databases compliant with the LGPD. Initially, consider a medical clinic that wishes to design a database to store patient and examination information. Assume that a patient can undergo zero or more examinations, and an examination can be performed by zero or more patients.

For the patients, the following information is to be stored: patient ID (cod-paciente), CPF (Brazilian Individual Taxpayer Registry), name, date of birth, address, skin color, religion, gender, and sex. Note that all these attributes pertain to personal data. It is also considered that the Data Controller of the clinic has defined that the attributes patient ID and CPF should be encrypted. Additionally, it has been determined that the attributes name and date of birth should be anonymized. For the Data Controller, the attributes skin color, religion, gender, and sex are classified as sensitive personal attributes, meaning they require special care. However, no specific treatment has been defined for these attributes. The address attribute is considered personal data but is also a semi-identifier. Nonetheless, no special treatment has been defined for the address attribute. To model aspects related to consent, the Data Controller has requested the creation of the following attributes: consent-description, consent-start, and consent-end. To address the concept of purpose, the Data Controller has requested the creation of the attribute purpose-description. For the medical exams, the following information is to be stored: exam ID (cod-exam), description, and value. Note that none of these attributes pertain to personal data. Additionally, the relationship between patient and examination includes two attributes: exam-date and result. The Data Controller has defined that the exam result, which is personal data, should be encrypted, while the exam-date is a semi-identifier with unspecified treatment.

Figure 2 illustrates the conceptual schema, generated during the conceptual design phase, using the ER-PD model and modeled with the brModeloPD tool. Note that the entity set “Patient” is represented by a rectangle with a dashed line, indicating that it represents a “Data Subject”. Also, note that next to each attribute name, in square brackets, the model highlights the data type and the treatment that should be applied to it.

3.2. Logical Design

In this work, we propose an adaptation of the Relational model, called R-PD, with the aim of enabling the logical design of databases in compliance with the LGPD. The R-PD model allows for the representation of key concepts present in the LGPD, such as: personal data, the type of processing performed on the data, and the data subject. The R-PD model also facilitates the representation of attributes related to consent, purpose, as well as data concerning children and adolescents.

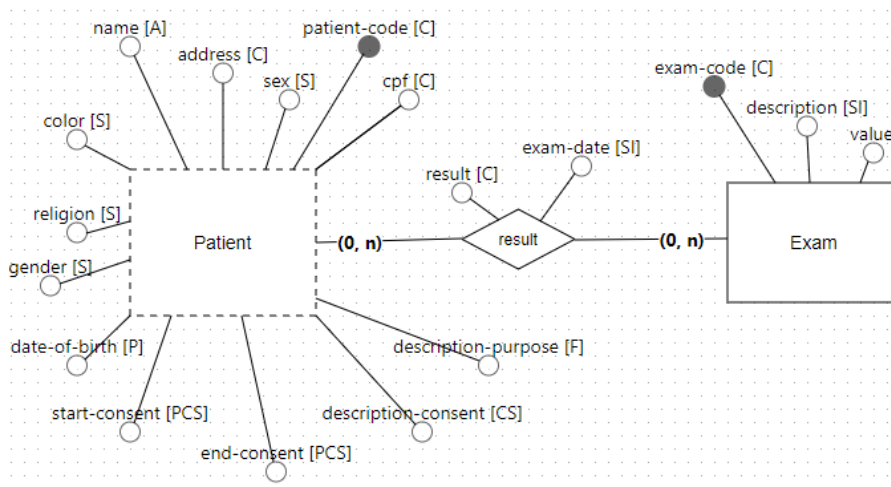


Figure 2. Conceptual Schema Using the brModeloPD Tool.

Figure 3 illustrates the logical schema for the previously described running example, using the brModeloPD tool. Note that the “Patient” entity is represented by a rectangle with dashed lines, indicating that it is a “Data Subject”.

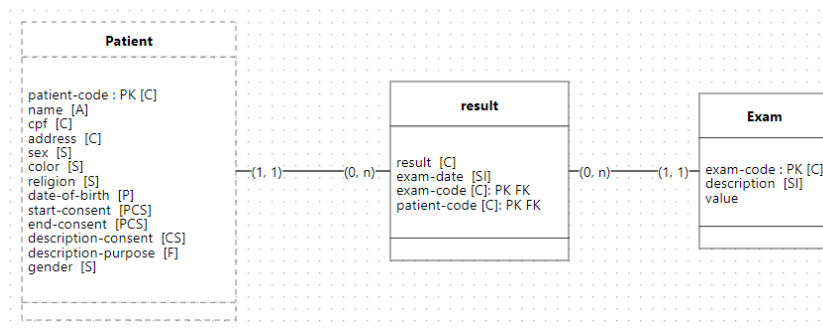


Figure 3. Logical Schema Using the brModeloPD Tool.

3.3. Physical Design

In this work, we propose an adaptation to the SQL CREATE TABLE command, called SQL-PD, with the aim of enabling the physical design of databases in compliance with the LGPD. This adaptation allows for the representation of key concepts present in the LGPD through metadata (comments in an SQL command). These metadata can be used for audits to ensure compliance with the LGPD. Listing 1 illustrates shows the CREATE TABLE command for the “Patient” table (which is part of the physical schema for the previously described running example). Note that the constraints arising from the LGPD are inserted through “comments” (in this example, following PostgreSQL syntax). Due to space constraints, we will not display the commands to create the “Exam” and “Result” tables. However, they can be easily conceived from Listing 1.

Listing 1. Comando CREATE TABLE Paciente (SQL-PD)

```

CREATE TABLE Paciente
  (patient-code integer NOT NULL,
   cpf char NOT NULL,
   name varchar NULL,
   address varchar NULL,
   date-of-birth date NULL,
   sex char NULL,
   color char NULL,
   religion char NULL,
   gender varchar NULL,
   start-consent date NULL,
   end-consent date NULL,
   description-consent varchar NULL,
   description-purpose varchar NULL,
   CONSTRAINT c1 PRIMARY KEY patient-code
   /* , */
   /* CONSTRAINT c2 Encrypted patient-code, */
   /* CONSTRAINT c3 Encrypted cpf, */
   /* CONSTRAINT c4 Sensitive sex, */
   /* CONSTRAINT c5 Sensitive color, */
   /* CONSTRAINT c6 Sensitive religion, */
   /* CONSTRAINT c7 Sensitive gender, */
   /* CONSTRAINT c8 Sensitive date-of-birth, */
   /* CONSTRAINT c9 Anonymized name, */
   /* CONSTRAINT c10 Anonymized address, */
   /* CONSTRAINT c11 Purpose description-purpose, */
   /* CONSTRAINT c12 Consent description-consent, */
   /* CONSTRAINT c13 Consent Period start-consent, */
   /* CONSTRAINT c14 Consent Period end-consent */
  )

```

4. Conclusion

In this work, we present a strategy called LGPDbyD for incorporating the requirements and restrictions of the LGPD into database design. To achieve this, we have made minor adaptations to the ER model, the Relational model, and the CREATE TABLE command. Additionally, we extended the brModelo tool to support the LGPDbyD methodology. The proposed methodology aims to facilitate the processes of database design and auditing in compliance with the LGPD.

References

- Carvalho, G., Bernardino, J., Pereira, V., and Cabral, B. (2023). Er+: A conceptual model for distributed multilayer systems. *IEEE Access*.
- Dani, A. and Getta, J. (2005). Conceptual modelling of computations on data streams.
- de Abreu, C., Praciano, F. D., Amora, P. R., and Machado, J. C. (2021). Consql: Consentimentos em sql para o processamento de consultas orientado a propósitos. In *Anais Estendidos do XXXVI Simpósio Brasileiro de Bancos de Dados*, pages 8–14. SBC.
- Kamble, A. S. (2008). A conceptual model for multidimensional data. In *APCCM*, volume 8, pages 29–38.
- Khan, K. M., Kapurubandara, M., and Chadha, U. (2004). Incorporating business requirements and constraints in database conceptual models. In *Proceedings of the first Asian-Pacific conference on Conceptual modelling-Volume 31*, pages 59–64.
- Sarkar, S. and Athanassoulis, M. (2022). Query language support for timely data deletion. In *Proceedings of the 25th International Conference on Extending Database Technology*, volume 2.
- Shastri, S., Banakar, V., Wasserman, M., Kumar, A., and Chidambaram, V. (2019). Understanding and benchmarking the impact of gdpr on database systems. *arXiv preprint arXiv:1910.00728*.