An Anonymization Library for Rapid and Diverse Anonymization of Brazilian Personal Data

Stefano Luppi Sposito[1], **Raylan da Silva Sales**[1], **Edna Dias Canedo**[1], **Geovana Ramos Sousa Silva**[1]

[1]Departamento de Ciência da Computação
Universidade de Brasília – Brasília, DF – Brazil

{luppisposito, raylanwork, geovannna.1998}@gmail.com, ednacanedo@unb.br

Abstract. *The prevalence of personal data in the hands of large companies highlights the necessity for robust regulatory frameworks. The General Data Protection Law (LGPD) seeks to standardize data usage, emphasizing minimal ownership and, when needed, anonymization in line with regulations. The absence of a specific tool for anonymizing Brazilian personal data remains a significant challenge. The lack of a dedicated tool for anonymizing Brazilian personal data poses a hurdle in achieving LGPD compliance. This study proposes the development of a library tailored to anonymize personal data, considering the unique aspects of Brazilian regulations. The goal is to create an efficient and secure library for removing identifiable information from documents, aligning with the LGPD. Furthermore, the results obtained from the implementation and testing of the developed library provide significant contributions to the data privacy community. The successful integration of support for various document formats such as .PDF, .DOCX, and .XLSX, coupled with the ability to anonymize text strings, demonstrates the versatility and practicality of the library. Notably, the performance tests reveal promising outcomes, showcasing the effectiveness of each function and regular expression employed. These results not only validate the functionality of the library but also underscore its potential in aiding individuals and organizations in adhering to data protection regulations.*

1. Introduction

The General Data Protection Law (LGPD)[da República 2018], sanctioned in 2018 and in effect since September 2020, regulates how companies operating in Brazil should act regarding the collection, processing, and sharing of personal and sensitive data. In essence, the LGPD recommends that companies anonymize personal data concerning their customers. According to the law, personal data includes all information capable of identifying its owner, such as name, ID, phone number, date of birth, location, email, and data related to social media accounts, among other information that can lead to the identification of the data subject. It is important to note that, according to the LGPD, data is considered anonymized only when it has definitively lost the ability to identify a person.

Data anonymization can be applied in any context involving personal data of one or more individuals, being not only necessary but indispensable, especially in sectors such as health[Bild et al. 2020], social networks, and business. Data anonymization techniques for personal information, such as masking, generalization, and distortion of data, can be

employed to protect sensitive information like banking details and residential addresses [Murthy et al. 2019]. The specific techniques used may depend on the type of data being anonymized and the specific needs of the organization. Therefore, it is crucial to strike a balance between privacy protection and the usefulness of data for analysis and research.

According to the General Data Protection Law [da República 2018]:

> Article 46. Data processing agents must adopt security measures, both technical and administrative, capable of protecting personal data from unauthorized access and accidental or unlawful situations of destruction, loss, alteration, communication, or any form of inappropriate or illegal treatment.

These security measures, among many others, may consist of anonymization tools, whether individual software or existing programming language libraries. Currently, various data anonymization tools are available from various internet sources, each performing their respective operations using unique and effective techniques. However, the most efficient tools do not meet the needs of Brazilian cases, lacking support for both the Portuguese language and the format of Brazilian personal data, such as CPF and RG. Therefore, there is a need for an open-source library that is secure and efficient, usable by any organization or individual, ensuring standardized and secure anonymization in accordance with LGPD.

The primary aim of this study is to develop a library tailored for anonymizing personal data in alignment with the requirements outlined in the Brazilian General Data Protection Law (LGPD). Through a comprehensive process, including an analysis of existing anonymization techniques, examination of the LGPD, identification of Brazilian data intricacies, and subsequent development of the library, the goal is to offer an efficient and secure solution for removing sensitive information from documents while upholding privacy and confidentiality standards mandated by regulations. The resultant technological solution not only facilitates compliance with data protection laws but also enhances information security and privacy, thereby contributing significantly to safeguarding personal data in the Brazilian context.

2. Background

2.1. General Data Protection Law

In the context of the digital society, the increasing flow of personal information and the need to protect people's privacy have become issues of extreme relevance [Tomás 2022]. In this regard, the General Data Protection Law (LGPD) emerges as important regulation in Brazil, establishing guidelines for the processing of personal data by public and private organizations [Teixeira 2020].

The LGPD was enacted in 2018 with the primary goal of ensuring individuals' control over their personal data, establishing clear rules on how this data should be collected, stored, processed, and shared [da República 2018]. The law aims to protect the privacy, freedom, and dignity of citizens while promoting transparency in data processing activities.

Anonymization, as emphasized by the LGPD [Kalam et al. 2005, Carvalho et al. 2020, da República 2018], is pivotal for safeguarding personal data

in Brazil. It enables organizations to utilize information for legitimate purposes like research and analysis while preserving individuals' privacy [Tomás 2022]. The LGPD stipulates specific criteria for data anonymization, requiring technical processes that render data unidentifiable, alongside security measures to safeguard data during processing. This legal framework underscores the importance of anonymization in upholding privacy rights [Teixeira 2020], facilitating data use in compliance with regulations while necessitating ongoing updates and security measures for effectiveness [Tomás 2022].

2.2. Anonymization

Anonymization is a process used in the protection of personal data, with the goal of rendering the anonymized information in such a way that it becomes impossible to identify the individual to whom the data refers [Tomás 2022]. This method seeks to eliminate or modify elements that could allow for the direct or indirect identification of data subjects, thereby ensuring the privacy and confidentiality of the information. Anonymization plays a fundamental role in compliance with data protection laws, such as the General Data Protection Regulation (GDPR) in the European Union [intersoft consulting 2018] and the General Data Protection Law (LGPD) in Brazil [da República 2018]. By applying appropriate anonymization techniques, organizations can use data for research, analysis, and product development purposes without compromising the privacy of the individuals involved.

The protection of privacy [Alves and Neves 2021] and personal data has become a paramount issue in the digital era. With the increasing volume of collected information, it is crucial to adopt appropriate measures to ensure the security and confidentiality of this data. In this context, anonymization emerges as a promising technique capable of preserving individuals' privacy while allowing the secure use of sensitive information. This chapter will address the benefits and challenges of anonymization, as well as its relationship with GDPR and LGPD [Tomás 2022].

2.3. Anonymization Techniques

There are various techniques available, each with its own advantages and disadvantages. For this work, some of the five techniques described in the paper by Murthy et al. [Murthy et al. 2019] were chosen: Suppression and Masking. However, it is essential to understand that numerous other techniques exist, which may be better applied in different situations but with the same ultimate purpose. Additionally, one must be aware of the limitations of anonymization, as in some cases, de-anonymization techniques can be employed to identify individuals through correlations and data cross-referencing [Murthy et al. 2019].

The data suppression technique involves the complete removal of a data point and its replacement with a value that lacks meaning, such as "XXXXX,". This technique is particularly useful when there is a need to conceal textual data, such as names and addresses. However, since this method completely eliminates the data, its application may interfere with the usability of the data, as they lose their meaning and value.Masking, on the other hand, entails obscuring data by replacing characters based on predefined rules, preserving certain characters like the first of a word while altering others like numbers and letters, as detailed in [Murthy et al. 2019].

2.4. Related Works

Several studies contribute to the field of data anonymization, providing insights into techniques, challenges, and tools.

The work by [Prasser et al. 2020] offers an overview of anonymization tools and algorithms, emphasizing the importance of balancing privacy and data utility.

Furthermore, [Jha et al. 2023] introduces z-anonymity and k-anonymity techniques, demonstrating their effectiveness in protecting user privacy, especially in continuous data streams.

[Murthy et al. 2019] presents five anonymization techniques and discusses their suitability for different types of data attributes, focusing on preserving data utility while mitigating privacy risks.

While existing techniques in programming languages like Python [Pelgrim 2023] provide solutions for data anonymization, there's a gap in tools specifically tailored for Brazilian personal data anonymization across various file types.

3. Library Proposal

The programming language chosen for this project is Python [FOUNDATION 2023], a high-level, interpreted, and general-purpose programming language. It is widely adopted in the data science community and has a wide range of libraries that facilitate data manipulation, processing, and analysis. The choice of Python is based on its ease of use, code readability, and the availability of a wide variety of libraries for data anonymization needs. For the effective functioning of the library, the Docx2txt, Docx, PDFMiner, and Pandas libraries were used.

3.1. Architecture

By adopting a modular approach, the architecture allows for easy extensibility and customization of the components according to the specific requirements of the anonymization context. The interconnected components work together to ensure an efficient flow of data throughout the anonymization process. Additionally, the modularity of the architecture facilitates the maintenance and continuous improvement of the solution, making it more adaptable to future changes and advancements in data protection.

The architecture can be divided into modules and components for better visual understanding, as follows:

- External Data Import Component: Responsible for importing data from external sources, such as CSV files, databases, etc.
- Personal Data Anonymization Module: Performs the general coordination of the anonymization process, calling specific modules for different types of data.
- Regex Component for Personal Data: Contains regular expressions to identify and anonymize different types of personal data (CPF, telephone, date, zip code, email).
- Anonymization Module for PDF, DOCX, XLSX and text files: Each specific module performs data anonymization for a specific file format or for text.
- Anonymization Module for Text Strings: Performs anonymization on text strings.
- Anonymized Data Output Module: Responsible for storing or sending anonymized data to the desired destination.

3.2. Anonymization Approches

To create the code, it was first necessary to decide which anonymization technique would be used in the library, based on the techniques proposed by Murthy *et al.* [Murthy et al. 2019]. For reasons of practicality and ease, the Suppression technique was chosen, also aiming at data security, since there is no way to recover anonymized data with this technique [Murthy et al. 2019]. In this case, the choosen aproach consists in completely erase the personal data contained in documents, once the objective consists in data anonymization in a way that the documents could be made public if needed, whithout the risk of having personal data leaks.

Considering that all quasi-identifying attributes to be anonymized had a pattern, xxx.xxx.xxx-xx for CPF for example, it was decided to create a class containing regular expressions that would serve to locate each data pattern throughout the text received. Each of the expressions searches the text for a specific format, regardless of the value that the attribute has. In the library code [Raylan Da Silva Sales 2023], the 'RegexDadosPessoais' class was created with the sole purpose of storing regular expressions and compiling them into variables through the 're' library [Foundation 2024]. Each variable: regexCPF, regexTelef, regexData, regexCEP and regexEmail, receives a regular expression responsible for finding the format of CPF, Telephone (in this case, cell phone number), Date, CEP and Email, respectively.

The 'pdfminer.six'[Valvekens 2024] library was used to process .PDF files, where the 'extract_text' function searches for the file through the provided path, extracts the text and stores it in string format in the text variable. After the pattern variable is created and receives the regular expression referring to the type of data to be anonymized, the library function 're': 'sub' is used, where it receives a regular expression, a replacement string and a string search. The function searches the search string for all occurrences recognized by the regular expression and replaces them with the replacement string. In this case, the replacement string used was #########. Finally, the new string with the anonymized data is returned to the main program.

The 'anonimiza_string' function was created so that the library could offer simple support for text strings that were already stored within variables in the program. The operation of this function is extremely similar to the 'anonimiza_pdf' function, since it receives a text string along with the type of data you wish to anonymize, and passes the arguments as parameters of the 'sub' function that changes and returns the text with the desired parts anonymized.

The 'anonimiza_docx2txt' function was created so that it is capable of reading a file in the .DOCX format and extracting its data. However, it is not capable of creating or overwriting a new file, since the docx library — used to create the docx file — is not compatible with PyPI. Therefore, the chosen approach was to return a text string, containing the contents of the anonymized file according to the user's choice.

Support for .XLSX files was added relying mainly on the 'pandas'[pandas via NumFOCUS 2024] library. The library retrieves data from the file using the 'read_excel' function and stores the dataset within the 'text' variable. To change the data within the dataset, it was necessary to use the 'apply' function, which goes through the entire dataset changing the values passed as an argument. In this case,

the values passed as arguments consist of: a regular expression, according to the user input and the anonymization string already used previously.

4. Validation

The test cases were made to ensure the integrity of the library and verify it works correctly, simulating a normal use of the application, where the previous presented anonymization functions were called, passing the archives paths, that would be stored in the variable ”arquivo” and the desired personal data to be anonymized in a string format: ”CPF”, ”Telefone”, ”Data”, ”CEP” and ”Email”. This string was then stored in the ”Flag” variable, which is used in the ”retorna_pattern” function, that serves the purpose of retrieving the regex function that contains the format of that specific data format. All the results are availabe in the GitHub page of the project [Raylan Da Silva Sales 2023], in the folder ”testResults” on the ”develop” branch.

It is worth noting that in all cases of examples presented in the images, all names or possible non-anonymized quasi-identifiers are censored with a black band, to preserve the identity and integrity of the individuals. All data anonymized by the library becomes the string #########, which does not need to be censored. The tests were carried out on machines with the Windows 10 operating system, Python versions 3.10 and 3.11 and each with 16GB of RAM.

4.1. PDF Anonymization

For test cases using files in the .PDF format, the strategy of recovering the anonymized string and storing it in a .TXT file was used, for easy viewing, since the .PDF anonymization function is not capable of generating a new file in the same format, maintaining the original text formatting. The tests carried out were carried out on documents found through internet searches, using the full name of individuals in single quotation marks as the search string, which only returns results that contain exactly the string between the quotation marks. Tests that had some type of personal data were used.

For this test, Covid-19 vaccination data from a municipality in the state of Rio De Janeiro [de São João da Barra 2024] were used. In this case, the data to be anonymized are the dates of birth and dates of vaccination, as shown in Figure 1.

This list was chosen due to the amount of data contained in the file, since it presents more than 6000 lines of information, which was considered to be a good amount of data to evaluate the capabilities of the library created, since The library’s ability to extract data from the file and process it, in addition to returning anonymized data, could be verified.

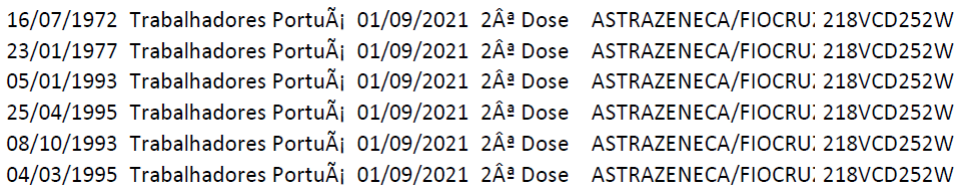
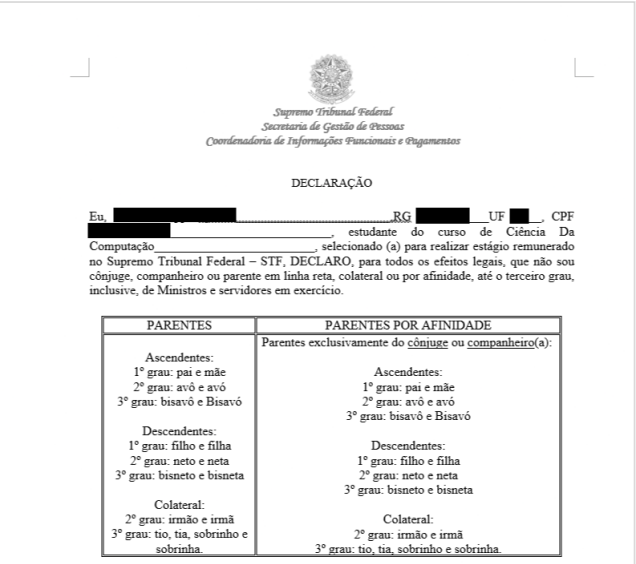


Figure 1. Non-Anonymized Vaccination Data

4.2. DOCX Anonymization

The first test consisted of applying the anonymization of CPF and RG, using the regular expression already contained in the library and another entered manually. The application was made in a .DOCX document that consisted of a declaration of nepotism that was used by one of the authors of this study to join the Federal Supreme Court as an intern. The file is not available to the public, but is shown in Figure 2.

Due to the fact that it is an official document, we wanted to evaluate whether the technologies used in the created library would be able to maintain the original formatting of the file, since if they were, the library could easily be used by government agencies to anonymize official documents without worry about changing the original formatting.



Supremo Tribunal Federal
Secretaria de Gestão de Pessoas
Coordenadoria de Informações Funcionais e Pagamentos

DECLARAÇÃO

Eu, [redacted] RG [redacted] UF [redacted], CPF [redacted], estudante do curso de Ciência Da Computação_____, selecionado (a) para realizar estágio remunerado no Supremo Tribunal Federal – STF, DECLARO, para todos os efeitos legais, que não sou cônjuge, companheiro ou parente em linha reta, colateral ou por afinidade, até o terceiro grau, inclusive, de Ministros e servidores em exercício.

PARENTES	PARENTES POR AFINIDADE
Ascendentes: 1° grau: pai e mãe 2° grau: avô e avó 3° grau: bisavô e Bisavó Descendentes: 1° grau: filho e filha 2° grau: neto e neta 3° grau: bisneto e bisneta Colateral: 2° grau: irmão e irmã 3° grau: tio, tia, sobrinho e sobrinha.	Parentes exclusivamente do cônjuge ou companheiro(a): Ascendentes: 1° grau: pai e mãe 2° grau: avô e avó 3° grau: bisavô e Bisavó Descendentes: 1° grau: filho e filha 2° grau: neto e neta 3° grau: bisneto e bisneta Colateral: 2° grau: irmão e irmã 3° grau: tio, tia, sobrinho e sobrinha.

Figure 2. Non-Anonymous Nepotism Statement

The code used for testing remained basically the same for the library version and the version contained in the 'develop' branch. Differing only in the fact that in the library version it was necessary to recover the anonymized string and, to facilitate visualization, a .txt file was created to store the string.

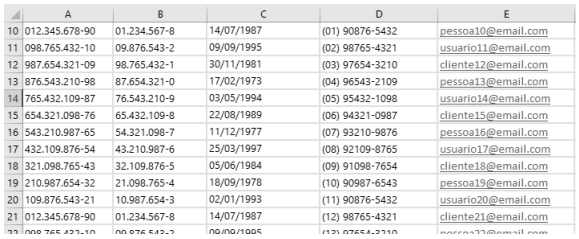
Using the library, instead of overwriting the file, it was decided to first use the .DOCX document anonymization function, to anonymize the CPF and, after that, the returned string was passed to the string anonymization function, since the .DOCX file was not recreated, so it would not be possible to extract the text again. Finally, the result was saved in a .txt file for easy viewing of the result.

4.3. XLSX Anonymization

To anonymize data in .XLSX format, a spreadsheet was used with fictitious data generated by Artificial Intelligence ChatGPT [OpenAI 2024], with 50 lines of data and 5 columns, where each column contains a quasi-identifier, namely: CPF, ID, Date of Birth, Cell Phone Number and Email.

This test was created with the aim of verifying the library's behavior when using anonymization functions in a chained manner, where the output file was changed in each

of them. The number of data lines, for this specific test, was not considered of paramount importance, since the objective was to verify only the anonymization functions and the ability of the pandas library to quickly overwrite data from a single file. The worksheet is shown in Figure 3



	A	B	C	D	E
10	012.345.678-90	01.234.567-8	14/07/1987	(01) 90876-5432	pessoa10@email.com
11	098.765.432-10	09.876.543-2	09/09/1995	(02) 98765-4321	usuario11@email.com
12	987.654.321-09	98.765.432-1	30/11/1981	(03) 97654-3210	cliente12@email.com
13	876.543.210-98	87.654.321-0	17/02/1973	(04) 96543-2109	pessoa13@email.com
14	765.432.109-87	76.543.210-9	03/05/1994	(05) 95432-1098	usuario14@email.com
15	654.321.098-76	65.432.109-8	22/08/1989	(06) 94321-0987	cliente15@email.com
16	543.210.987-65	54.321.098-7	11/12/1977	(07) 93210-9876	pessoa16@email.com
17	432.109.876-54	43.210.987-6	25/03/1997	(08) 92109-8765	usuario17@email.com
18	321.098.765-43	32.109.876-5	05/06/1984	(09) 91098-7654	cliente18@email.com
19	210.987.654-32	21.098.765-4	18/09/1978	(10) 90987-6543	pessoa19@email.com
20	109.876.543-21	10.987.654-3	02/01/1993	(11) 90876-5432	usuario20@email.com
21	012.345.678-90	01.234.567-8	14/07/1987	(12) 98765-4321	cliente21@email.com

Figure 3. Worksheet Containing Fictitious Quasi-Identifiers

4.4. Discussion of Results

The analysis of the results obtained during the test was extremely important, as its functionality and reliability could be verified. This section aims to interpret these results and provide a more in-depth look at what was observed.

The tests carried out demonstrated that the library has remarkable efficiency in anonymizing chained data, as was possible to notice mainly in the tests carried out with .XLSX documents. Tests on other types of files also proved satisfactory, as the library was able to anonymize and return text strings without complications. However, there was a delay in the anonymization of vaccination data, where the program execution took 48.78 seconds. Future optimizations will be carried out in an attempt to reduce execution time.

The library demonstrated robustness when dealing with a variety of data types and testing situations, having no problems with the option of manually inserting a regular expression to identify personal data not supported by default, further increasing the possibilities for anonymization in different contexts . The library, during the tests, anonymized all selected data in an impeccable way, so that they could not be recovered, guaranteeing the security and inviolability of the quasi-identifiers presented.

When comparing the library with other existing ones, it was possible to see comparable performance in several aspects, however, this performance is superior when it comes to the anonymization of Brazilian personal data. The library created proved to be superior in terms of data anonymization in relation to the Anonymizedf [Fridriksson 2020] library, since the library in this work has the ability to find data in specific types of documents and anonymize them, while the Anonymizedf library only has the ability to generate false data that can be used in manual text replacement. Concerning the Anonympy library [ArtLabs 2022], the library developed for this study has demonstrated superiority in terms of the range of data types that can be anonymized, coupled with comprehensive support for Brazilian datasets.

5. Threats to Validity

The library lacks the capability to recreate .PDF and .DOCX files with anonymized data, instead only providing the raw string via code. This limitation may lead to data structure and formatting loss when attempting to reconstruct documents from the returned string.

Efforts to find compatible libraries for .PDF and .DOCX files during the library's development proved unsuccessful, hindering automatic file recreation functionality.

6. Conclusion

This work aimed to create a library that serves as a personal data anonymizer for Brazilian data. The library was designed to receive data from files in different formats and apply the suppression technique to the desired data. It was decided to create regular expressions that would serve as validators to traverse the texts and find the data to be anonymized. Additionally, support for .PDF, .DOCX, and .XLSX documents was successfully added to further enhance the library's utility, as personal data is not always contained in a single type of document. In the case of documents and files not supported by the library, an option for anonymizing text strings was also included to provide even more support.

One of the main contributions of this work is the aforementioned library, as it will assist and simplify the anonymization of data for individuals and even companies. The library comes with documentation that facilitates its use. According to the tests, it was possible to verify the functioning of each function and each regular expression inserted into the library. As observed, all tests yielded expected and even surprising results in terms of execution time and the quantity of sequentially anonymized data. Additionally, the library demonstrated superiority in anonymizing Brazilian personal data when compared to other tools.

Future works primarily involves improvements to the created library, specifically adding support for the modification and creation of .DOCX and .PDF documents, similar to what is already being done for .XLSX documents.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Alves, C. and Neves, M. (2021). Especificação de requisitos de privacidade em conformidade com a lgpd: Resultados de um estudo de caso. In *WER*.

ArtLabs (2022). anonympy 0.3.7. `https://pypi.org/project/anonympy/`.

Bild, R., kuhn, K. A., and Prasser, F. (2020). Better safe than sorry – implementing reliable health data anonymization. *PDigital Personalized Health and Medicine*, 270:68–72.

Carvalho, A. P., Canedo, E. D., Carvalho, F. P., and Carvalho, P. H. P. (2020). Anonymisation and compliance to protection data: Impacts and challenges into big data. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 22nd International Conference on Enterprise Information Systems, ICEIS 2020, Prague, Czech Republic, May 5-7, 2020, Volume 1*, pages 31–41. SCITEPRESS.

da República, P. (2018). Lei geral de proteção de dados pessoais (lgpd). *Secretaria-Geral, accessed in November 19, 2019*. `https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf`.

de São João da Barra, P. (2024). Dados de vacinação da covid-19. https://sjb.rj.gov.br/uploads/16b8b4fc9fde4cb8ba34bd4119b041644120a5b3.pdf.

FOUNDATION, P. S. (2023). Python documentation. https://www.python.org/doc/.

Foundation, T. P. S. (2024). Regular expression operations. https://docs.python.org/3/library/re.html.

Fridriksson, A. (2020). anonymizedf 1.0.1. https://pypi.org/project/anonymizedf/.

intersoft consulting (2018). General data protection regulation. https://gdpr-info.eu.

Jha, N., Vassio, L., Trevisan, M., Leonardi, E., and Mellia, M. (2023). Practical anonymization for data streams: z-anonymity and relation with k-anonymity. *Perform. Evaluation*, 159:102329.

Kalam, A. A. E., Deswarte, Y., Trouessin, G., and Cordonnier, E. (2005). Personal data anonymization for security and privacy in collaborative environments. In McQuay, W. K. and Smari, W. W., editors, *Proceedings of the 2005 International Symposium on Collaborative Technologies and Systems, CTS 2005, Saint Louis, Missouri, USA, May 15-20, 2005*, pages 56–61. IEEE Computer Society.

Murthy, S., Abu Bakar, A., Abdul Rahim, F., and Ramli, R. (2019). A comparative study of data anonymization techniques. In *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, pages 306–309.

OpenAI (2024). https://chat.openai.com.

pandas via NumFOCUS, I. H. b. O. (2024). Python pandas documentation. https://pandas.pydata.org.

Pelgrim, R. (2023). Data anonymization in python. https://mostly.ai/blog/data-anonymization-in-python.

Prasser, F., Eicher, J., Spengler, H., Bild, R., and Kuhn, K. A. (2020). Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*, 50(7):1277–1304.

Raylan Da Silva Sales, S. L. S. (2023). anonymization-library. https://github.com/Rayxan/anonymization-library.

Teixeira, G. C. (2020). O papel social da lei geral de proteção de dados no brasil. *UNIVERSIDADE DO SUL DE SANTA CATARINA*, pages 1–59.

Tomás, J. C. P. (2022). *Data anonymization: algorithms, techniques and tools*. PhD thesis, Instituto Politecnico de Coimbra.

Valvekens, M. (2024). Pdfminer.six documentation. https://github.com/pdfminer/pdfminer.six.