# Development of an application to anonymize data to be shared in the cloud

DOI: 10.46932/sfjdv3n4-097

Received in: April 14th, 2022
Accepted in: June 30th, 2022

**David González Marrón**
Ph.D. in Computer Sciences
Institution: Tecnológico Nacional de México - Instituto Tecnológico de Pachuca (ITP)
Address: Carretera México-Pachuca, Km 87.5, CP: 42080
E-mail: david.gm@pachuca.tecnm.mx

**Verónica Paola Corona Ramírez**
Master in Information Technologies
Institution: Tecnológico Nacional de México - Instituto Tecnológico de Pachuca (ITP)
Address: Carretera México-Pachuca, Km 87.5, CP: 42080
E-mail: veronica.cr@pachuca.tecnm.mx

**Angélica Enciso González**
Master in Software Engineering Management
Institution: Tecnológico Nacional de México - Instituto Tecnológico de Pachuca (ITP)
Address: Carretera México-Pachuca, Km 87.5, CP: 42080
E-mail: angelica.eg@pachuca.tecnm.mx

**Alejandro Márquez Callejas**
Engineering in Information and Communication Technologies
Institution: Tecnológico Nacional de México - Instituto Tecnológico de Pachuca (ITP)
Address: Carretera México-Pachuca, Km 87.5, CP: 42080
E-mail: alejandromarqueztec@gmail.com

**Iridian Sandivel Pérez Hernández**
Engineering in Information and Communication Technologies
Institution: Tecnológico Nacional de México - Instituto Tecnológico de Pachuca (ITP)
Address: Carretera México-Pachuca, Km 87.5, CP: 42080
E-mail: iridianperez012@gmail.com

**ABSTRACT**
A design of an interface is proposed to be developed as a lightweight application that allows the use of different algorithms to anonymize proposed by various authors, the application focuses on adapting data from distributed applications that must interact using the JSON lightweight Data Interchange Format. The user interaction is minimized for the anonymization process, an interface is provided to facilitate the selection process of the anonymization algorithms that users choose.

**Keywords:** anonymization, data security, data privacity.

## 1 INTRODUCTION

With the boom that the Internet and the Internet of things have had, thousands or even millions of data provided daily by ourselves are created, this massive data is called Big Data, which according to the Bioethics and Law Observatory of the University of Barcelona (2015), is the processing of large volumes of information in order to establish correlations between them, predict trends and make decisions. If used responsibly, this data can be a powerful asset. It is clear that smart applications in general are a great boost to innovation in areas such as health, inclusion, environment and business (Kroes, N., 2010), all this thanks to Internet-enabled services, which contribute to the development However, the central point of this research is to protect the sensitive information contained within large amounts of information, since these have generated a security alert, implying a challenge to secure and protect them.

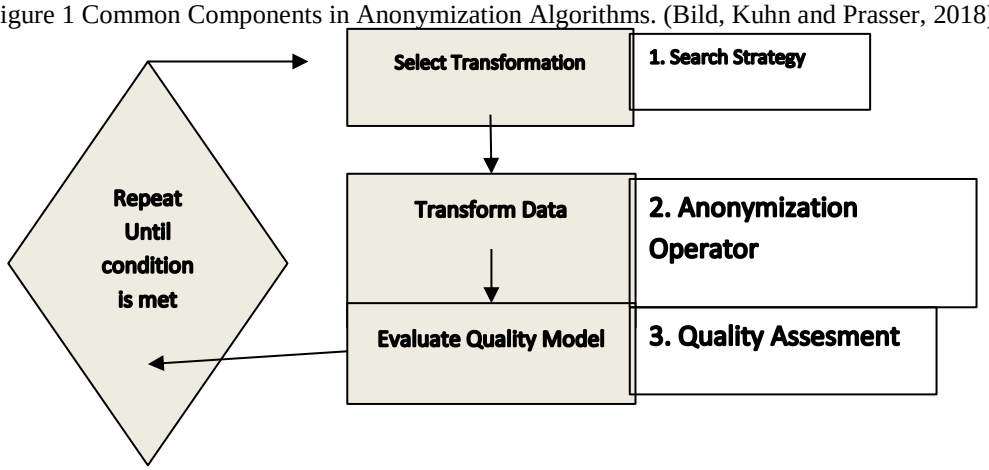
In this work, some strategies are proposed to respect privacy options through an architecture that meets current protection expectations. Unfortunately, as explained in the article by Ramonet (2016), large corporations such as Google and Facebook obtain information from all over the world on a daily basis, the question is how aware are we of what we share. The security of personal data must be guaranteed, protecting it against attacks, treatment or unauthorized use, loss and destruction of data, however, the term security applied in real life is a complex concept.

## 2 METHOD DESCRIPTION

According to the Spanish Data Agency (2016), anonymization is a technique that is responsible for minimizing the risks of re-identification of anonymized data, that is, that the origin of that information cannot be obtained, allowing publication without violating the protection rights of individuals and organizations. These anonymization techniques must be approached from the concept of data protection and various principles have been designed for this purpose that must be applied in the anonymization processes.

Figure 1 shows the common components that are implemented in anonymization algorithms regardless of their nature.

Figure 1 Common Components in Anonymization Algorithms. (Bild, Kuhn and Prasser, 2018).




## 3 PROBLEMATIC

Assuming having sensitive and personally identifiable information from different users, as shown in Table 1, and that it is going to published for survey analysis, it is necessary consider that, as Greenleaf (2017) explains, there are global laws on data and personal privacy, so the identification of who are the owners of that information must be avoided, for which, what is required is to eliminate the information that allows personal identification as shown in Table 2.

Table 1 Example of a Table with sensitive information

| Id | Name | Age | Postal Code | Sex | Location | Activity | Search |
|---|---|---|---|---|---|---|---|
| 47677 | Lynton Tadeo | 27 | 15153 | Male | -24.52202, -56.71208 | Activity1 | Search 1 |
| 47906 | Camilla Van | 52 | 41398 | Female | -13.21192, 133.57312 | Activity 2 | Search 2 |
| 47677 | Alexander Hayes | 24 | 47905 | Male | 35.89210, -95.58568 | Activity 3 | Search 3 |

Table 2 Example of a table with removal of sensitive information

| Id | Name | Age | Postal Code | Sex | Location | Activity | Search |
|---|---|---|---|---|---|---|---|
| 47677 | Lynton Tadeo | 27 | 15153 | Male | -24.52202, -56.71208 | Activity1 | Search 1 |
| 47906 | Camilla Van | 52 | 41398 | Female | -13.21192, 133.57312 | Activity 2 | Search 2 |
| 47677 | Alexander Hayes | 24 | 47905 | Male | 35.89210, -95.58568 | Activity 3 | Search 3 |

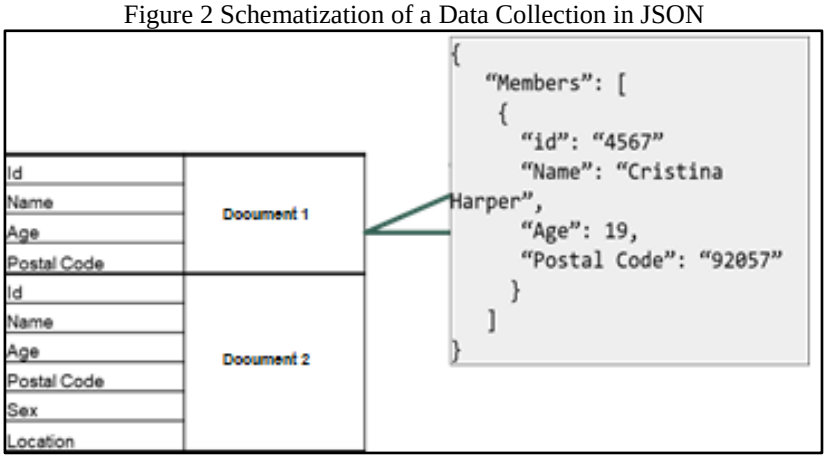
This means that if these data are subjected to an analysis, they should not differ from the information that could be obtained. In the anonymization process, the direct or indirect identification chain must be broken.

However, as Anderson (2009) explains, with only three information parameters such as age, sex and postal code, a person can be identified, resulting in 87% of the population being identified with these parameters, which means that doing only the process shown in Table 2 may be insufficient to achieve information privacy. For this reason, it is essential to apply this type of techniques in areas where the flow of information is provided by a very large group of users who continuously enter sensitive data.

## 4 METHODOLOGY

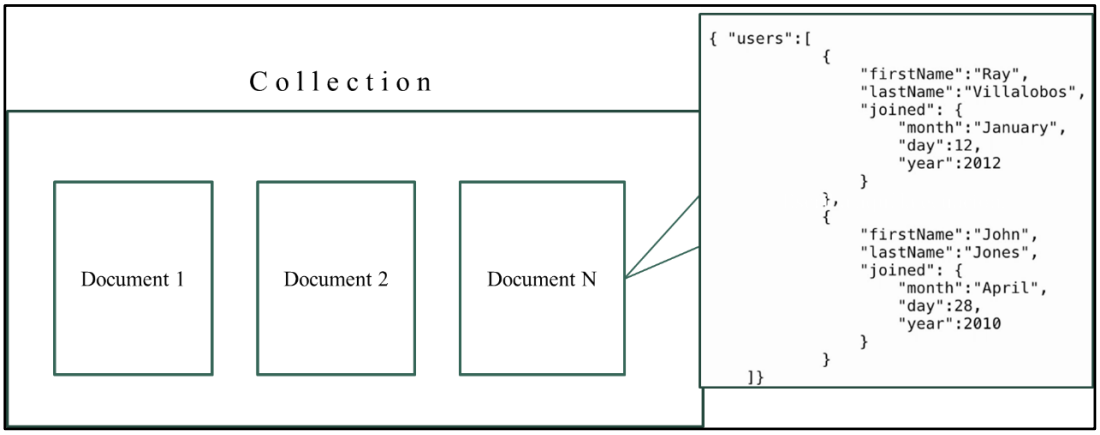
The proposal regarding this process is that the anonymization be carried out on JSON-type data, a format that is widely used for data exchange on the Internet. A collection of JSON files, as shown in Figure 2, is a set of documents, in which each document has different fields or attributes that can be in a different order.

Figure 2 Schematization of a Data Collection in JSON




As can be seen in Table 3, a collection in JSON files can contain different documents, these in turn can have a non-standardized number of attributes, for example, in the case of Document 1 shown in Figure 2, it can be seen which is made up of 4 attributes and Document 2 by 6 attributes.

Table 3 Attributes on a JSON Document




As can be seen in Figure 3, in the first column (Element 1) is the data source section, in this module the elements that provide information from sensors, networks and users are involved. After that, the second column is made up of a collection of files with JSON type format (Element 2), which have inside data composed of the owner of the information, which, as mentioned at the beginning of this article, contains fields with sensitive information.

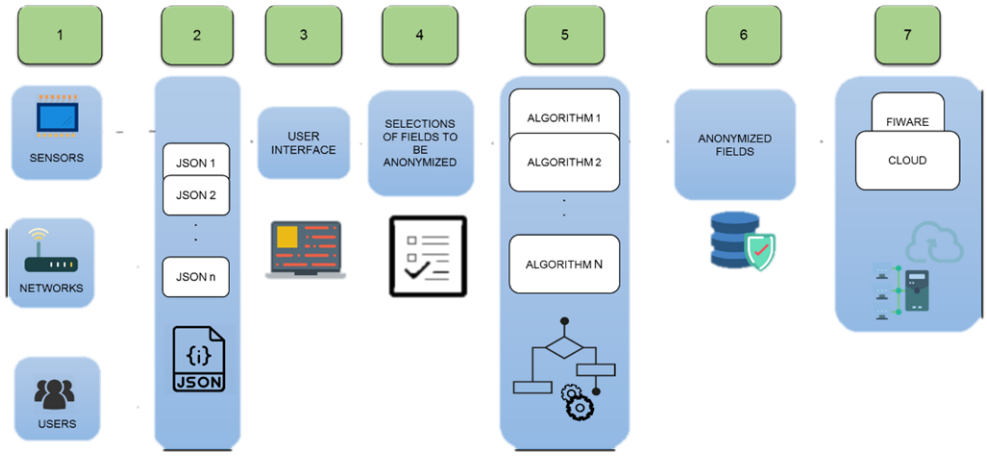
In order for the information to be processed, this proposal proposes using a platform with a graphical interface (Element 3) where users can know all the attributes that make up the JSON file to be anonymized. The fourth column (Element 4) allows you to select the attributes to anonymize and the desired anonymization algorithms. Due to the complexity of selecting the appropriate data and anonymization algorithms, this work has been left to be carried out by human experts, since it is currently very complex to be able to automatically anonymize the information, since there may be omission or loss of relevant information.

Once the fields to be anonymized have been selected, the respective algorithms (Element 5) of the fifth column must be executed to transform the sensitive data into anonymized information, then the modified fields will be displayed avoiding the identification of the owner of the information (Element 6) to be uploaded to the cloud (Element 7).

Elements Involved

1. Origin of data
2. Information stored in JSON format.
3. User Interaction.
4. Use of the proposed tool.
5. Transformation Process.
6. Result Obtained.
7. Transfer to Cloud.

Figure 3 Anonymization Proposal




## 5 FUNCTIONAL DESCRIPTION OF ALGORITHMS

The proposal focuses on the development being able to distinguish the attributes that are not repeated, and display them to the user so that he/she can decide which procedure to execute in each case. To exemplify the above, Table 4 shows how the unique attributes are selected.

Table 4 Unique or non-repeating attributes within the file to be anonymized

| Attributes |
|---|
| Id |
| Name |
| Age |
| Postal Code |
| Sex |
| Location |

Next, the type of treatment to be given to the data must be specified and the anonymization algorithms that the user considers relevant must be selected based on the type of data that each attribute has, these may be Numerical or Nominal, in case of choosing to anonymize those data will select which algorithm should be executed depending on the type of attribute of the data. Table 5 shows the treatment that can be given to numerical attributes and Table 6 shows the treatment of nominal values. The anonymization algorithms respond to the needs described in the Guidelines for the Anonymization of Microdata (2014).

Table 5 Anonymization Algorithms for Numerical Values

| Algorithm | Description | Example | Explication |
|---|---|---|---|
| Numeric_1 | Put zero in a fraction of the number | 1075 → **1000** | The last two numbers to zeros |
| Numeric_2 | Generalize the number using a range | 1075 → **[999-1100]** | The value is understood to lie between the two given numbers |
| Numeric_3 | Approximate the number using comparison symbols (>=<) | 1075 → **>850 <1150** | It is understood that the number is greater than 850 and less than 1150 |
| Numeric_4 | Replace a fraction of the number with asterisks (*) | 1075 → **10**** | In this case the last two numbers are replaced with asterisks |

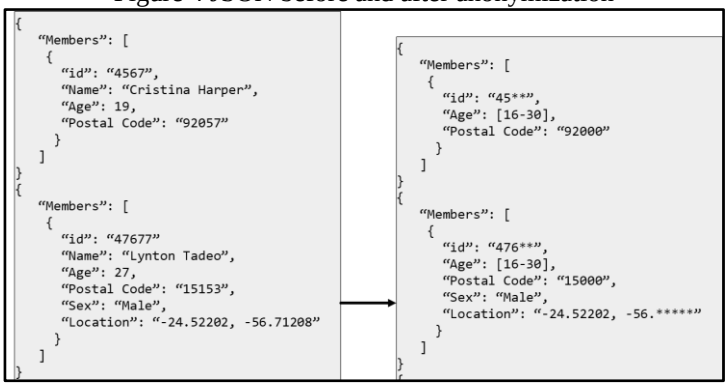
Table 6 Anonymization Algorithms for Nominal Values

| Algorithm | Description | Example | Explication |
|---|---|---|---|
| Nominal_1 | Truncate text by defining the number of characters | Juan Pérez ↓ **Juan P** | The last four letters of the text were removed |
| Nominal_2 | Concatenate the uppercase | Juan Luis Hernández López ↓ **JLHL** | Only the initials remain. |
| Nominal_3 | Put the first letter of each word | Calle Antonio de la Cruz ↓ **CAdlC** | Only the first letters of each word are considered |
| Nominal_4 | Replace the text with another defined | Calle Porfirio Diaz ↓ **Restringido** | The *Restringido* word is selected to be the word to be replaced in the text |
| Nominal_5 | Extract vowels | Calle Porfirio Diaz ↓ **Ae oiio ia** | Only vowels are preserved in the text |

## 6 FINAL COMMENTS

An example of the result obtained during the expected anonymization process is shown in Figure 4.

Figure 4 JSON before and after anonymization



## 7 CONCLUSIONS

Through this work, we have sought to introduce some anonymization algorithms that have been developed in the area, which allow the production of anonymous data that are resilient against re-identification and with privacy guarantees, and that also retains its usefulness for use in different contexts.

A seven-step approach has been presented to acquire data, anonymize it and publish it in the cloud with the guarantee of respecting the privacy of the data source.

Currently, the automation of the process using the Python language is being completed, the anonymization algorithms have been developed, and only one of the algorithms mentioned in the numerical and nominal data types remains to be implemented. The times obtained for anonymization by selecting various algorithms have been satisfactory since for small files of 500 records, 0.078 sec. has been required, for medium files of 10,000 records 1.8 sec. and for large 100,000 records 14.48 sec.

## REFERENCES

Agencia española de protección de datos. (2016). Orientaciones y garantías en los procedimientos de anonimización de datos personales. Recuperado de https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2016/Orientaciones_y_garantias_Anonimizacion.pdf

Anderson, N. (2009, agosto). Anonymized" data really isn't—and here's why not. arstechnica. Recuperado de https://arstechnica.com/tech-policy/2009/09/your-secrets-live-online-in-databases-of-ruin/

Bild, R., Kuhn, K. A., & Prasser, F. (2018). SafePub: A Truthful Data Anonymization Algorithm With Strong Privacy Guarantees. Proceedings on Privacy Enhancing Technologies, 2018(1), 67–87. https://doi.org/10.1515/popets-2018-0004

Dirección de Regulación, Planeación, Estandarización y Normalización. (2014, 29 agosto). Lineamientos para la Anonimización de microdatos. Recuperado 22 agosto, 2018, de https://www.dane.gov.co/files/sen/lineamientos/DSO_020_LIN_08.pdf

Greenleaf, G. Global Data Privacy Laws 2017: 120 National Data Privacy Laws, Including Indonesia and Turkey (January 30, 2017). (2017) 145 Privacy Laws & Business International Report, 10-13; UNSW Law Research Paper No. 17-45. Recuperado de: https://ssrn.com/abstract=2993035

Kroes, N.: European Commissioner for Digital agenda, The critical role of cities in making the Digital Agenda a reality. Closing speech to Global Cities Dialogue Spring Summit of Mayors Brussels, 28 May (2010)

Observatorio de Bioética y Derecho, Documento sobre bioética y Big Data de salud: explotación y comercialización de los datos de los usuarios de la sanidad pública, Universidad de Barcelona, 2015. Recuperado de: http://www.publicacions.ub.edu/refs/observatoriBioEticaDret/documents/08209.pdf

Ramonet, I. (2016, 6 febrero). Google sabe todo de ti. La Jornada. Recuperado de http://www.jornada.com.mx/2016/02/06/mundo/018a1mun