pangeanic

Everything you need to know about data anonymization

www.pangeanic.com

1. Introduction to data privacy

In today's world, **data is collected and transferred** continuously, and it's its governance and control are a complex process that can pose challenges in many areas.

Malicious attacks jeopardizing the **cybersecurity** of data are increasing in number and danger. Meanwhile, **consumers are growing more concerned** about their data as they become increasingly aware that companies hold information about them.

These issues have led to more **comprehensive legislation.** New regulations carry hefty fines and seek to guarantee citizens' rights concerning the privacy of their data and the obligation of those who collect this information to protect its security.

Faced with a continually evolving landscape, companies search for the **best solutions** to use databases to improve their business processes while complying with the law and **citizens' expectations. Data anonymization** allows for just this. Put simply, the process of data anonymization consists of **altering personal information** in a database so that it cannot be identified.

This means data can retain its value and become **an asset** for a company seeking to better understand its customers and the processes it carries out. Security of this information and **privacy** at the individual level is also guaranteed: anonymization makes it impossible for sensitive information to be attributed to a specific person.

This process can be carried out aggressively or superficially – but ultimately, some of the altered data may include personal names, company names, addresses, telephone numbers, or family relationships.

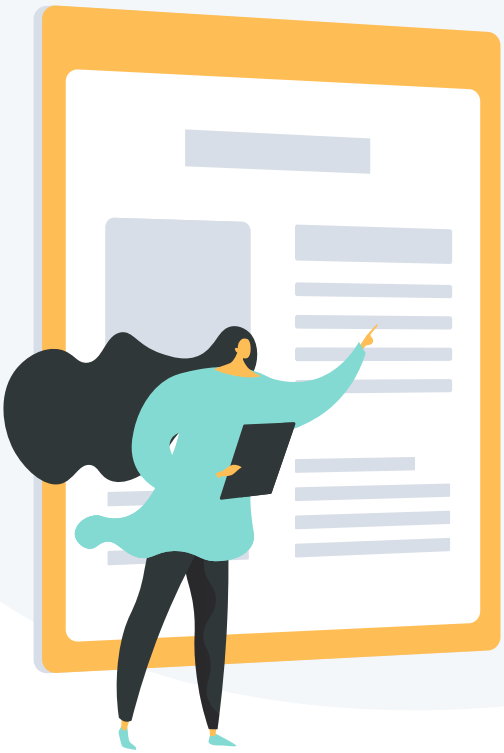
The aim is to minimize the identification of individuals within a database and the ability to access **sensitive information** that could be used for malicious ends.

Data anonymization allows companies to store and transfer this data more securely, satisfying at least three objectives:

1. Compliance with **privacy laws** (CCPA/CPRA, HIPAA, GDPR or APPI) and guaranteeing **fundamental rights.**

2. Achieving **secure data** storage so data maintains its analytical value while protecting its privacy and confidentiality.

3. Building consumer and employee trust in a company's ability to secure their personal information.



2. Importance of Data Anonymization

Anonymization is a necessity for any company with a data assets for the following reasons:

1. Companies can **store and transfer data** more **securely**. While most organizations have systems in place to protect the information they collect, it's impossible to eliminate every vulnerability or security breach from external attacks.

   Anonymization adds an extra layer of security and **minimizes the risk** of cyber-attacks, since anonymized data has no practical value and can't be used.

   Anonymization protects **data internally** too, increasing its confidentiality.

2. Increased **efficiency** and **simpler data exchange**, boosting productivity and synergies with third parties.

   The process is **automated** with the right anonymization software, meaning companies can share, use and even monetize this data.

   For example, cross-referencing information in cloud environments (in search of behavioral analysis, trends, correlations, or clustering), would only be possible in two ways: by obtaining user consent or by applying **anonymization techniques**. Any other scenario would be in breach of the law.

   Some programs possess multilingual capabilities, a useful feature when working with international institutions (e.g at the European level).

3. Complying with **legislation**, eliminating the possibility of any breaches.

4. Adhering to international **privacy laws**, including CCPA/CPRA, HIPAA, GDPR, or APPI.

5. Developing **consumer confidence** – especially relevant at a time of increased wariness when providing personal data.

6. Ability to **retain data**. Current legislation allows for data retention in the public interest, statistical purposes, or scientific or historical research. These actions can only be carried out if the data has undergone a process of anonymization (or pseudonymization).

Data Privacy Legislation

EUROPEAN LEVEL

GENERAL DATA PROTECTION REGULATION (GDPR)

The **General Data Protection Regulation (GDPR)** is mandatory in the European Union. It seeks to ensure personal data is kept private at all levels.

The regulation came into force in May 2018 and seeks to **protect** citizens' data as part of the **fundamental rights** of natural persons (right to honor and personal and family privacy).

This regulation defines personal data as "any information about an **identified or identifiable** natural person," including physical, physiological, and other types of data. This also includes **ARCO rights** (access, rectification, cancellation, and opposition), the right to data portability, and the **right to be forgotten.**

The latter is crucial for data anonymization. The law states citizens have the right to request the deletion of their data in three scenarios: when they are used beyond the purpose for which they were collected, when they have been collected unlawfully or when they did not give their consent for its use.

If a citizen requests for their data to be deleted, the organization is obliged to comply. But if the data has been anonymized, you won't run into this problem - as you'll be able to guarantee its confidentiality and therefore **safeguard its value and attributes.**

Anonymized data is no longer "information about an identified or identifiable natural person" and can therefore be kept for statistical, research, or analytical purposes.

THE WG 29 OPINION 05/2014

EU law **complements** the legislation set out in GDPR by referencing anonymization. It defines anonymization as "the result of a processing personal data to irreversibly prevent identification".

It also establishes three data privacy features vulnerable to **risk** (uniqueness, inference, and linkability) and highlights the need for organizations to generate processes to avoid data identification.

While the legislation doesn't specify how to carry out this process, it specifies anonymization and pseudonymization **best practices.**

NATIONAL LEVEL

NEW ORGANIC LAW ON DATA PROTECTION AND DIGITAL RIGHTS GUARANTEE (LOPDGDD)

This new regulation came into force in December 2018 and seeks to replace the previous regulation, Organic Law 15/1999 on Personal Data Protection.

Its objective is to **adjust Spanish legislation** to the EU's GDPR.

Spain's regulation, like GDPR, seeks to **protect the right to privacy,** in line with Article 18.4 of the Spanish Constitution.

Among other aspects, this law defines personal data as any information, in text, image or audio format, that allows a person's identification.

It then states some data is considered low risk (a person's name or e-mail address) and other information is of higher risk (health or religious affiliation data).

In line with EU regulation, data that does not identify a person is not considered personal data.

Anonymized data, therefore, are **not subject to the same legal obligations** as the rest, but are governed by The Regulation on the free flow of non-personal data.

Some of the most relevant aspects of this regulation include:

- Establishing the protection in data **transfer** processes, taking special care in the protection of personal data on the internet.
- Includes the **right to be forgotten** and the **right to portability,** as well as the European standard.
- Introduces **new obligations and requirements** for organizations collecting and using personal information, focusing on user consent. These requirements affect both the collection of data and its storage and transfer.
- Establishing an obligation to **notify** the Spanish Data Protection Agency of **security breaches** within a maximum of 72 hours.
- Establishing that organizations must understand their **obligations** regarding the treatment of user data and data citizen's rights.
- Establishing the concept of **proactive prevention or accountability:** companies must maintain the security of their databases proactively (before problems arise). This includes cybersecurity aspects and anonymization and pseudonymization processes.
- Specific rights surround the data of deceased persons. Thus, citizens will be able to **request access, rectification, and deletion** of this information.

LAW 14/2007, OF JULY 3, 2007, ON BIOMEDICAL RESEARCH

Given the growing use of databases in the biomedical research sector, Law 14/2007, of July 3, 2007, on Biomedical Research includes criteria to protect citizens' privacy in this context.

Obtaining and analyzing statistical data is essential for medical knowledge, as recognized by this regulation. But this law also focuses on the risks associated with collecting and storing medical data and seeks to **protect citizens** against possible breaches of their rights.

On the one hand, it guarantees freedom of research and scientific production under Article 20 of the Spanish Constitution. On the other hand, it protects a person's dignity and identity, for instance, by defining the use of genetic data or biological samples in research.

The law regulates biomedical research at all levels, from the use of cells and tissues of human embryonic origin to the right not to be discriminated against, along with regulating donation.

These are some aspects it includes on **data privacy:**

- An **individual's autonomy** to grant the consent of data used in the context of biomedical research, plus the right to obtain information before its publication.
- Duty of **confidentiality** for any person with access to personal information.
- It sets the **standards for traceability** of human cells and tissues.

3. Anonymization

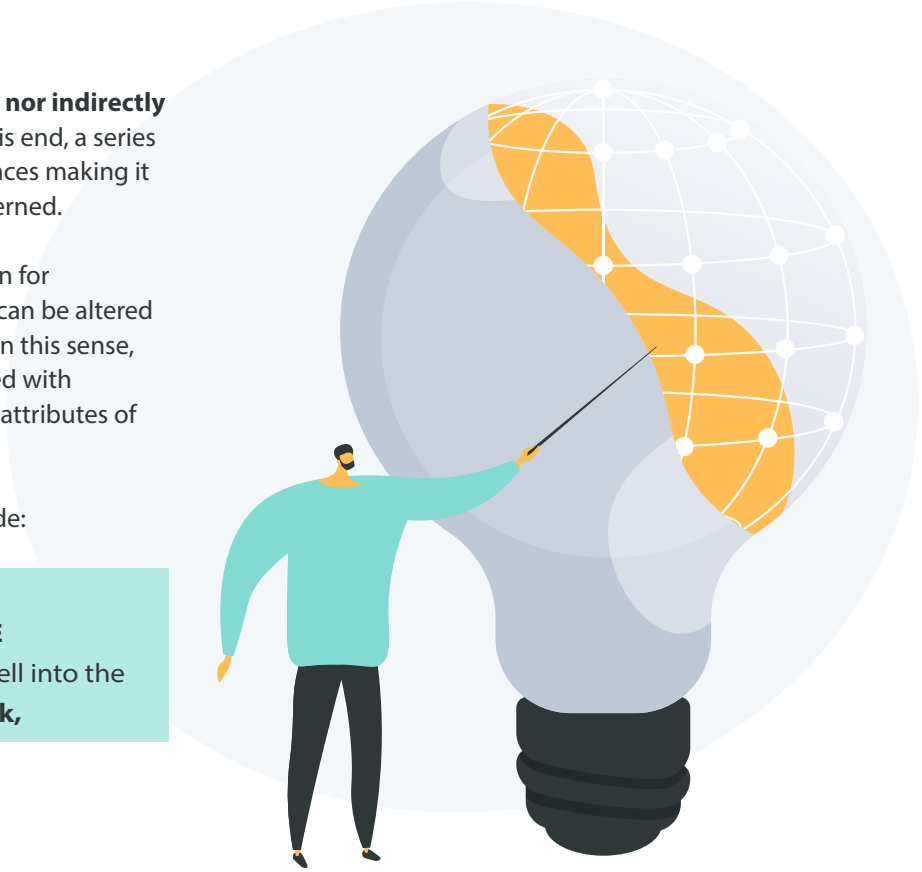
Types of anonymization methods

Anonymization consists of **altering** data in such a

way that the person is **neither directly nor indirectly and irreversibly identifiable.** To do this end, a series of data is exchanged for symbols or spaces making it impossible to identify the person concerned.

Depending on the configuration chosen for anonymization, different types of data can be altered (names, locations, professions, dates.). In this sense, the need for anonymization is combined with organizations' ability to retain valuable attributes of the data they store.

Some anonymization techniques include:



ORIGINAL SENTENCE

In **1988,** ownership of the club fell into the hands of **Standard Bank,** which repossessed the club from **Zola Mahobe.**

PSEUDONYMIZATION

In **1539,** ownership of the club fell to **British Petroleum,** which repossessed the club from **Richard Nixon.**

ID TAG

In **[[DATE]],** ownership of the club fell to **[[ORG]],** which repossessed the club from **[[PER]].**

REDACTION

In **XXXX,** ownership of the club fell into the hands of **XXXXXXXXXXXXXXXXXXXXXXX,** who foreclosed on the club to **XXXXXXXXXXXXXXXX.**

These techniques are based on the classification of named entities, and other techniques known as masking (e.g social security numbers, phones, emails, credit cards, etc.) Pangeanic has also developed a mixed technique based on neural models and anonymization profiles. Pangeanic's method is customizable to each company – customizing will always improve the model's results. This profiling technique incorporates regular expressions and

anonymization dictionaries.

Personal names, emails, and dates can be modified so that they cannot be accessed except with a password. When, say, a client needs to add a new word dictionary, we can **safely reverse** the identification process, i.e. allowing the original data to be recovered. This data only becomes identifiable when the data controller deems it necessary – the data remains inaccessible to anyone not authorized to access it. However, full re-identification can also be blocked with a pseudonymization process if necessary, or if the password has been destroyed.

In any case, it's important to know that the legislation linked to **GDPR** defines data that can be re-identified as personal data. This is because it considers that, using **additional information** (i.e. knowing the password) it is possible for this data to be attributed to a natural person.

Anonymization Tools

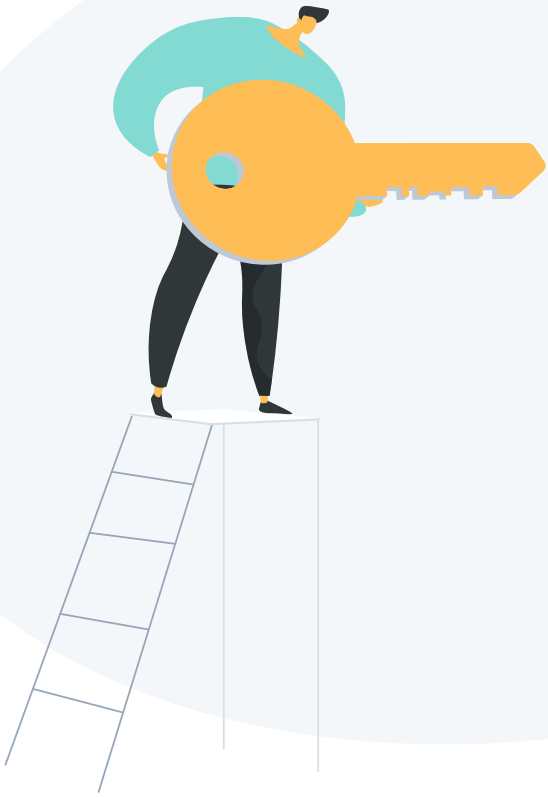
At least two requirements are needed to anonymize data sets:

1. **Professional teams**
   In the case of anonymization, usually different teams work independently from each other, so that the activities of one do not intersect with the other.

   In addition, to make the process as secure as possible, teams should follow specific steps: **A risk assessment, deciding which anonymization techniques are most appropriate** for a data set, and identifying **security measures** to maintain data confidentiality at all times.

   Each team should have a person in charge who in turn establishes who will be authorized to access the data.

2. **The right software**
   Nowadays, tools with artificial intelligence capabilities enable companies to carry out their own anonymization processes in a totally secure way, also incorporating capabilities for secure data storage and transfer.



Main industries affected by data anonymization

Many industries today depend on storing and transferring data to generate the most efficient processes possible.

For this reason, they must proactively implement security measures, including the ability to anonymize the information at their disposal. Only in this way can they ensure compliance with the law while using intelligence provided by data analytics.
Some of the main industries using anonymization include:

- **Banking:** Anonymization makes it possible to secure sensitive information such as customer's account numbers, their banking activity, customers' transactions, addresses, among many other data.

- **Healthcare:** Hospitals and research institutions hold sensitive data like medical records often transferred to third parties. This includes private and public companies and others related to the sector, such as insurance companies, psychology centers, physiotherapy centers, etc. For example, a patient's personal information or medical history may form part of a study aimed at developing a drug or curing a disease. With anonymization,

laboratories receiving information can't access sample donor's identities, making this process completely safe and legal.

- **Legal:** lawyers, lawsuits, and consultants work with legal documents that often contain personal data. This information must be anonymized from the source.
- **Public administrations:** In many cases, governments and public administrations are obliged to keep data open and transparent. To data confidentiality breaches, data must be anonymized. Thus, it is possible to make information available to citizens and enhance institutional transparency, without these data being identifiable.

4. Data security, a key factor

The lack of control and protection of personal data by organizations is a major risk, no matter whether they are small or large organizations.

The volume of data to be stored and managed is growing, yet companies and decision-makers are often **unaware** of the status of this information within their organizations, and its importance. This means they are not taking **preventative measures** and are increasing the risk of a cybersecurity breach.

This lack of diligence is compounded by increasingly **stringent data privacy legislation,** carrying severe penalties.

In this regard, avoiding problems inevitably requires **impeccable management of confidential information and data,** including the correct anonymization processes.

5. Anonymization of structured and un-structured data

Companies' data assets are usually stored in databases and data-lakes or repositories. The information stored in databases is often **structured**, as the storage includes lots of meta data and internal references usually in the Data Definition structure of

the Database.

Documents, image files or even plain text are typically stored and kept as individual items with simple organization and metadata.
Pangeanic provides anonymization solutions for both structured and unstructured assets.

For structured information the anonymization solution provides DB masking in two stages, allowing users:

1. To know how PII-sensible a Database is, where the private data is, which type of data. This is the process of **Discovery**.
2. To Generate an **anonymized copy** of the database where the identified private information has been masked using one of the available methods (redaction, indexing, gaps,…)

The two functionalities, Discovery and Anonymization, are offered eventually in two modes:

- **On-line.**
  The discovery and anonymization process is done accessing an on-line connected database. It can be done over the full database or over a partial list of tables and fields.

- **On-dump.**
  The process run on a dump file of the database, a plain text Json or SQL format. The anonymization output being a dump-file where the PII or sensible info has been anonymized.

6. Masker, a 360° solution for businesses

Masker is an artificial **intelligence-powered anonymization tool** that enables the transfer and storage of de-identified data securely. By de-identification we mean anonymization that hides sensitive or personal data. Masker includes human-monitoring techniques for model output to enhance

anonymization.

Through this system, companies can ensure **compliance with privacy regulations,** as well as **store data** that cannot be identified and share information that cannot be tracked, with the highest security standards always in mind.

With Artificial Intelligence, Masker can **identify personal data** (personally identifiable information or PII) and **replace the most sensitive information** using different anonymization techniques.

Companies can adjust this tool according to their needs. This includes the level of **sensitivity** of the personal data or the **specific techniques** to be used to anonymize it.

Among the benefits of Masker, the following stand out:

1. The possibility to choose between **permanent or reversible anonymization.** Companies can irreversibly delete personal data or decide to re-anonymize it. Although anonymized content often cannot be recovered; Masker can be configured to generate an index document containing the necessary information for the software to reconstruct the text.
2. Choose the **level of anonymity.** The user can choose how aggressive the process needs to be. This includes a spectrum of possibilities ranging from removing direct identifiers (e.g. names) to all secondary information (e.g. family relationships or job description).
3. It integrates as an **API** with back-ends such as MySQL, SQL Server, Oracle, MongoDB and most database formats. In addition, custom systems can be integrated on-premises or through private SaaS deployments.
4. Available for most **data formats,** including structured and unstructured. This ranges from text documents, to emails, social networks or business applications, such as MS Office documents...).
5. It is a **multilingual platform** that uses language-specific or multilingual anonymization models and is integrated with the machine translation tool, PangeaMT.
6. Ability to **train, configure and clone the artificial intelligence** engine around specific languages, vocabulary or terms.
7. **Flexible deployment,** using a secure cloud-based service, a service-agnostic API and the ability to deploy on-premises.

In short, Pangeanic allows you to **improve the efficiency** of communication processes between organizations, protect customer data, automate processes and manage database risks.

Our **decades of experience** applying **artificial intelligence to language** has allowed us to create a secure and efficient system already trusted by companies in the legal, financial and banking, government, pharmaceutical and medical sectors. We are also part of the European Commission's MAPA project, which aims to create a **multilingual anonymization platform** for the benefit of financial, healthcare and public institutions.

7. Success story: MAPA Project of the European Commission

MAPA is an integration project aimed at introducing natural language processing (NLP) tools and **developing a set of tools** for efficient and reliable anonymization of texts in the medical and legal fields. The proposal addresses all official EU languages,

including those with scarce resources (such as Latvian, Lithuanian, Estonian, Slovenian and Croatian).

The project aims to **build a complete, open source, multilingual anonymization toolkit** capable of detecting personal data (name, addresses, emails, credit cards, bank accounts, etc.) and anonymizing them, which will help public administrations to comply with the GDPR.

What experience does Pangeanic's have in anonymization?

Pangeanic is the **designer and manager of the MAPA** project, as well as the consortium composed of companies and public entities from Spain, France, Malta and Latvia, with Polish public administrations as observers.

Pangeanic's role is that of **data collector and annotator of entities** (persons, places, dates, streets...) in **9 European languages** that form the basis of the recognition by the neural network of the information to be anonymized in a text.

In addition, **Pangeanic has developed its own anonymization models** to include features beyond what is of interest to public administrations and to include languages such as Japanese.

What results have been achieved?

Currently, anonymization is being **used by European Complaints Watch offices** in each country in each language, anonymizing thousands of European user and consumer information documents, as well as legal firms.

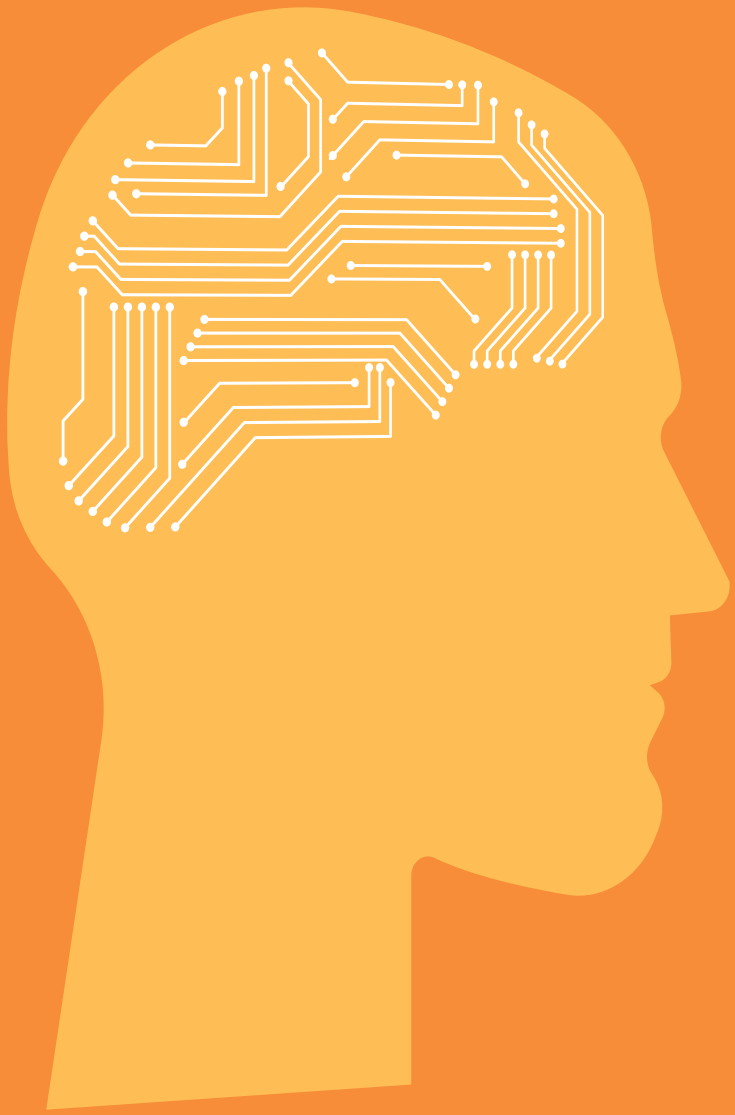
Also, major companies in the **financial sector, public finances, hospitals and health services** are considering its use.

Do you need more information about our products?

Contact our team and receive personalized advice.

CONTACT

pangeanic



Follow us:

www.pangeanic.com

Illustration resources from freepik.com