

AXA

How to apply anonymization in the context of GDPR?

© AXA

For Internal Use Only

- ***Purpose of the document***

This document provides guidelines in order to help AXA's entities to define their own approach to implement anonymization with regards to some GDPR requirements and available best practices. It will also define also what is pseudonymization compared to anonymization.

- ***Audience of the document***

This document can be used for awareness of executives and operationals on Anonymization. As the topic requires it, it provides some technical information which is rather intended towards Data Architects, Data Managers, Data Experts.

- ***Scope and limits of the document***

This document provides a consolidation of common knowledge and guidance required on anonymization across AXA entities. It should be used as a basis to build the capabilities adapted to specific entity's context. Each entity will have to customize and build its own step by step implementation roadmap.

The main environments in which anonymization is required in order to keep the maximum data range and value are BI environment whether it is a Data Warehouse or Data Lake. The main focus was indeed put on the "right to be forgotten" requirement for GDPR on those environments. However, this paper can be useful in other contexts like anonymizing/pseudonymizing ante-production environments or operational¹ environments. Or even, useful for other GDPR requirements such as secondary purpose limitation and data minimization.

¹ In operational environment, the only anonymization technique is deletion. See section 1

Revision History of working version

Version	Date	Author	Description
V0.1	26/03/2018	Fanny Vuillemin	Initialization of the document, with the contribution of Guillaume Hervouin, Céline Lescop, Fabrice Perrin, Fanny Vuillemin
V0.2	30/03/2018	Fanny Vuillemin	Taking into Account Celine Lescop revision comments
V0.3	16/04/2018	Fanny Vuillemin	Taking into account revision comments from Nicolas Shire, Fabrice Perrin, Celine Lescop and Gilles Coquet
V0.4	16/04/2018	Fanny Vuillemin	Current version
V0.5	07/06/2018	Fanny Vuillemin	Taking into account returns from Alain de Lamberterie, AXA France; Marie-Pierre VANDEN BERGHE, Benjamin DANIELS, Colette MAMMERICKX, Axa Belgium; Barbara Wolf, Axa Germany, Fuencisla GARZON-MERINO, Axa Spain

AXA Internal References

Title	Description
GDPR Technical Minimal Requirements for Anonymization, Pseudonymization and Deletion	Stream 4 - Minimal Requirements for Privacy techniques v1.0.docx
GDPR Architecture Patterns	GDPR Compliance project - ArchitecturePatterns.pptx
GDPR Definitions: encryption, anonymization, pseudonymizing	GDPR Compliance project - Definitions.pptx
Data anonymization Scoping Document	WP_0061 - Data Anonymization - Architecture description.pptx
Prepared data definition, categories and sub-categories to enable further data analytics	CrownJewelsRegister.pptx
Short introduction to Swiss GDPR project	SmartDataCircle-56-AXACH.pptx available on ONE
GDPR - AXA 25 policies - booklet - v1.0.xlsx	GDPR - AXA 25 policies - booklet - v1.0.xlsx

External References

Title	Description
Techniques d'anonymisation	Paper from Benjamin NGUYEN, Insa1 Centre Val de Loire et Inria2 Paris-Rocquencourt
G29 – Article 29 – DATA PROTECTION WORKING PARTY, 0829/14/EN WP216 – Opinion 05/2014 on Anonymization Techniques	G29 Opinion on Anonymization Techniques


Approval list

<i>Approved by</i>	<i>Entity</i>	<i>Approval Date</i>
Colette MAMMERICKX	<i>AXA Belgium</i>	13/06/2018
Marie-Pierre VANDEN BERGHE	<i>AXA Belgium</i>	13/06/2018
Benjamin DANIELS	<i>AXA Belgium</i>	13/06/2018
Alain DE LAMBERTERIE	<i>AXA France</i>	07/06/2018
Fuencisla GARZON MERINO	<i>AXA Spain</i>	15/06/2018
Barbara WOLF	<i>AXA Germany</i>	12/06/2018

Distribution list

Open to AXA internal

List of definitions

 Information	<p>About Definitions</p> <p>We have made explicit the following list of definitions that represent a certain level of consensus in the data world but that may vary depending on the context. One of the first steps you have to do when comparing your local situation with the vision proposed in this paper is to check that you share the same definition and if not to adjust.</p> <p>We have chosen to provide them along the document to help the reader and to provide the list below to find each of them.</p>
---	--

Anonymization.....	8
Confidential attributes.....	8
Data subject.....	8
DATASET / TABLE.....	8
Deletion	8
Encryption	8
Field	8
Identifiers	8
Inference.....	7
Linkability.....	7
Personal data / PII.....	8
Pseudonymization	8
Quasi-identifiers	8
Record	8
Singling out	7
Tokenization	9

Table of content

▪ PURPOSE OF THE DOCUMENT.....	1
▪ SCOPE AND LIMITS OF THE DOCUMENT	1
FOREWORD.....	6
WHAT ARE THE DIFFERENCES BETWEEN ANONYMIZATION, PSEUDONYMIZATION, ENCRYPTION AND DELETION? WHAT IS THE GLOBAL LIFE CYCLE OF A DATA? HOW TO APPLY “RIGHT TO BE FORGOTTEN” ON EACH ENVIRONMENT?	8
1. BASIC DEFINITIONS.....	8
2. WHAT IS THE DIFFERENCE BETWEEN ANONYMIZATION AND PSEUDONYMIZATION? WHY SHOULD I USE DELETION? IS ENCRYPTION A GOOD WAY OF ANONYMIZING?	9
3. HOW SHOULD I IMPLEMENT THE RIGHT TO BE FORGOTTEN ON MY ENVIRONMENTS AND WHAT KIND OF SOLUTION SHOULD I AIM AT IMPLEMENTING TO ANSWER THIS COMPLIANCE REQUIREMENT?	11
A. <i>Operational Environments</i>	11
B. <i>Analytical Environments</i>	12
GLOBAL PROCESS FOR UNDERTAKING ANONYMIZATION.....	14
1. STEP 1: STUDY THE BUSINESS USE CASES.....	14
2. STEP 2: IDENTIFY PERSONAL DATA AND BASELINE RISK.....	15
3. STEP 3: SELECT ANONYMIZATION TECHNIQUES WITH RELATED RISK REDUCTION.....	15
4. STEP 4: ASSESS THE ANONYMIZATION EFFECTIVENESS	16
5. STEP 5: ASSESS THE USE-CASE EFFECTIVE VALUE.....	17
6. STEP 6: APPLY ANONYMIZATION STRATEGY	17
WHAT ARE THE VARIOUS TECHNIQUES YOU CAN USE TO GET ANONYMIZED DATA?	19
1. ANONYMIZATION TECHNIQUES.....	20
A. <i>Data Masking</i>	20
B. <i>Basic Data Masking</i>	21
C. <i>Randomization</i>	21
D. <i>Generalization</i>	22
E. <i>Hashing</i>	23
2. ENSURING ANONYMITY: K-ANONYMITY, L-DIVERSITY	23
A. <i>K-anonymity</i>	23
B. <i>L-diversity</i>	24
LIST OF PERSONAL DATA AND SENSITIVE DATA AND PREFERRED ANONYMIZATION TECHNIQUES TO APPLY	26
VARIOUS WAYS TO CREATE AN ANONYMIZED DATASET	33
1. ANONYMIZING DATA WITHIN THE SAME DATASET: ANONYMIZATION AT A RECORD LEVEL	33
A. <i>Impacts</i>	34
B. <i>Pros and Cons of this technique</i>	35
2. ANONYMIZING DATA WITHIN A DEDICATED DATASET: ANONYMIZATION AT A DATASET LEVEL	35
A. <i>Impacts</i>	36
B. <i>Pros and Cons of this technique</i>	37
MAIN TAKEAWAYS.....	39

Foreword

With the GDPR came two major personal data management requirements:

- the right to be forgotten and the fact that personal data should be deleted from our systems after a defined amount of time;
- the necessity to limit personal data exposition to the processes (minimization)

However, personal data are now considered as the “new oil”, this deletion can be seen as a threat to unlock value in a digital world.

In the GDPR - AXA 25 policies booklet², it is said that:

“To ensure a basic maturity level, each entity should have a methodology to assess the level of appropriate safeguards with regard to a risk based approach of a project with scientific or historical research purposes or statistical purposes or for archiving purposes in the public interest.

This methodology should describe the various appropriate technical and organizational measures considering:

- the risks of re-identification such as Singling out, Linkability, Inference;
- the sensitivity of the data processed;
- the criticality of the project taking into account the scope (such as number of data subjects) and context (external providers, cloud solutions, etc).“

As an example, Privacy Impact Assessment and/or Security Risk assessment could be leveraged to address these measures.


In that context, anonymization – non-reversible masking of personal data – and pseudonymization – reversible masking of personal data - will allow to mitigate the risk of data loss. Indeed, they are solutions to investigate to balance value generation from data driven new businesses on the one side and compliance on the other side.

However, those technics are not widely used within AXA today. All AXA entities and most specifically CDO and IT people working with data need to better know and understand them, to find the right way to apply them in a real context.

Hence, one of the purpose of this document is to gather the best practices gathered in different experiences – AXA France who started GDPR program one year ago in 2016, DIL GDPR project on the Data Lake providing deletion services and actually working on anonymization services, Switzerland who will implement anonymization on their BI environments, Belgium who selected a Tool for anonymization and pseudonymization to cover BI environments, and market input to scale them across AXA. As a method of providing “appropriate safeguards”, the GDPR specifically mentions pseudonymization. “Each entity shall identify other equivalent technical measures such as noise addition, substitution, differential privacy, hashing, tokenization etc. Data subjects shall have a right to object to the data processing for specific reasons relating to their particular situations (article 19 of the GDPR). “

² See GDPR - AXA 25 policies - booklet - v1.0.xlsx – Rule 22

We will then focus on the simple technics that can be used on personal data to make a set of data “anonymized”, on the way to implement such technics by giving concrete examples on personal data types identified, the overall process to follow to undertake Anonymization of a dataset, and on the various tests to go through to ensure the minimum needed to consider the data are anonymized.


 Definitions	Singling out or Individualization: Possibility to identify the individual with one field (identifier)
	Linkability or Correlation: Possibility to identify the individual with the correlation of two or more field (quasi-identifier)
	Inference : Possibility to get closer to the individual with a high degree of probability with the correlation of fields (quasi-identifier)

Section

1

What are the differences between Anonymization, Pseudonymization, Encryption and Deletion? What is the global life Cycle of a data? how to apply “right to be forgotten” on each environment?

1. Basic definitions

 Definitions	Data subject: Any individual who has his or her data collected or processed
	Personal data / PII: Any information relating to an identifiable individual ('data subject') that can be used, either by itself or together with other personal data, to identify the natural person. Examples are name, an ID number, location data, birth data, IP address...
	DATASET / TABLE: A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.
	Record : a record (sometimes called a row) is a group of fields within a table that are relevant to a specific entry
	Field or Attribute: a field is a data structure for a single piece of data
	Identifiers (also known as direct personal data): These are attributes in the original data set that unambiguously identify the data subject to whom a record corresponds. Examples are passport number, social security number, full name, etc.
	Quasi-identifiers (also known as key attributes or indirect personal data) : These are attributes in the original data set that, in combination, can be linked with external information to re-identify (some of) the subjects to whom (some of) the records in the original data set refer. Examples are job, age, city of residence, etc.
	Confidential attributes : These are attributes that contain sensitive information on the data subject. Examples are salary, religion, health condition, etc.
	Anonymization : destruction of the identifiable data in a dataset. Anonymization irreversibly destroys any way of identifying the data subject
	Pseudonymization : a method to substitute identifiable data with a reversible, consistent value. Pseudonymization substitutes the identity of the data subject in such a way that additional information is required to re-identify the data subject
	Deletion : destruction of the data whether they are identifiable or not
	Encryption : conversion of data to totally unintelligible “text” to those who may try to access it, even in the case of data breaches.

Tokenization : process of substituting a sensitive data element with a non-sensitive equivalent, referred to as a token, that has no extrinsic or exploitable meaning or value. The token is a reference (i.e. identifier) that maps back to the sensitive data through a tokenization system. The mapping from original data to a token uses methods which render tokens infeasible to reverse in the absence of the tokenization system

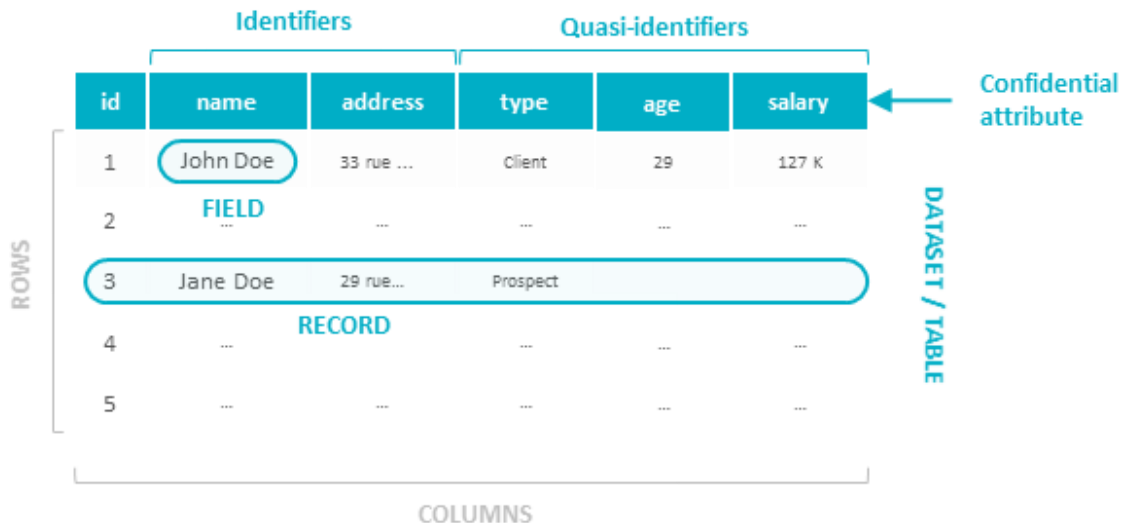


Figure 1: a dataset, its records and fields

These definitions are important because when you do anonymization, you have to start with a specific dataset. According to the type personal data (Identifier, Quasi-identifier), you will have to treat them in a different way, apply a different anonymization technique.

2. What is the difference between Anonymization and Pseudonymization? Why should I use deletion? Is encryption a good way of anonymizing?

If you look at simple definition, Anonymization is the destruction of the identifiable data in a dataset while Pseudonymization is a method to substitute identifiable data with a retrievable, consistent value.

By identifiable data, we mean personal data, data which defines a data subject either directly (name, surname...) or indirectly (color of hair, moustache, height, weight ...).

Another way of saying it, Anonymization irreversibly destroys any way of identifying the data subject while Pseudonymization substitutes the identity of the data subject in such a way that additional information is required to re-identify the data subject.

If you look at the last definition, you could compare pseudonymization and anonymization to a witness protection program. The objective of the “witsec³” is to hide a witness and make sure nobody will find him. Anonymization is the perfect way to hide a person, while Pseudonymization is hiding a person but somebody (including you) have ways to reach and find him. Your aim with anonymizing the dataset is to make sure that nobody will be able to find your witness, even you.

³ Witsec : United States Federal Witness Protection Program

So why do we talk about deletion and encryption. Well, you can have a look at the raw definitions:


- Deletion is the destruction of the data whether they are identifiable or not.
- Encryption is the conversion of data to totally unintelligible “text” to those who may try to access it, even in the case of data breaches.

If you look back at the metaphor, then deletion is getting rid of any traces of your witness. One of the solution would be to fake the death of your witness and recreate a whole new persona. By simulating the death, you get rid of all the data concerning the old persona and have to create all new data for the new persona. If your witness is supposedly not alive anymore why bother looking for him!

Encryption is like getting all your data concerning your witness inside a secret file for which you need a password to access. For pseudonymization, usually the witsec officer will have kept a phone to reach the witness or keep in mind the password (private key). So, if anyone can access the phone number or the password, no more witness protection.

Deletion and encryption are just techniques to undertake respectively anonymization and pseudonymization.

Imagine now: 20 years after the witness entered the protection program, someone recognize the face of the witness on a local newspaper then retrieves his new name. Your encryption has no more effect! Your anonymization became pseudonymization!

 Important	<p>There is a great difference between applying Anonymization/Pseudonymization to a dataset and Anonymizing/Pseudonymizing a set of fields.</p> <p>Indeed, Anonymization/Pseudonymization applies to all identifiable data in order to make the data subject unidentifiable whilst anonymizing/pseudonymizing a set of fields is only applying the techniques to some of the identifiable data.</p> <p>In other words, applying techniques at field level is a mandatory step, however you have to consider the full dataset to guaranty that identification of a subject is no more possible.</p>
---	--

3. *How should I implement the right to be forgotten on my environments and what kind of solution should I aim at implementing to answer this compliance requirement?*

The right to be forgotten should spread along all your data lifecycle and in each environment where the personal data are copied. This means either operational environments or BI/Analytics environments, but also ante-production environments.

Let's imagine we have the following data lineage: a customer is created in the CRM database. There is an output flow which sends the data to an Analytics environment. Of course, the operational database has ante-production databases in order to validate any new changes to the CRM software. There is the same need for the Analytics environment. Moreover, the Analytics environment has a Sandbox environment on which data scientist can develop their Analytics models.

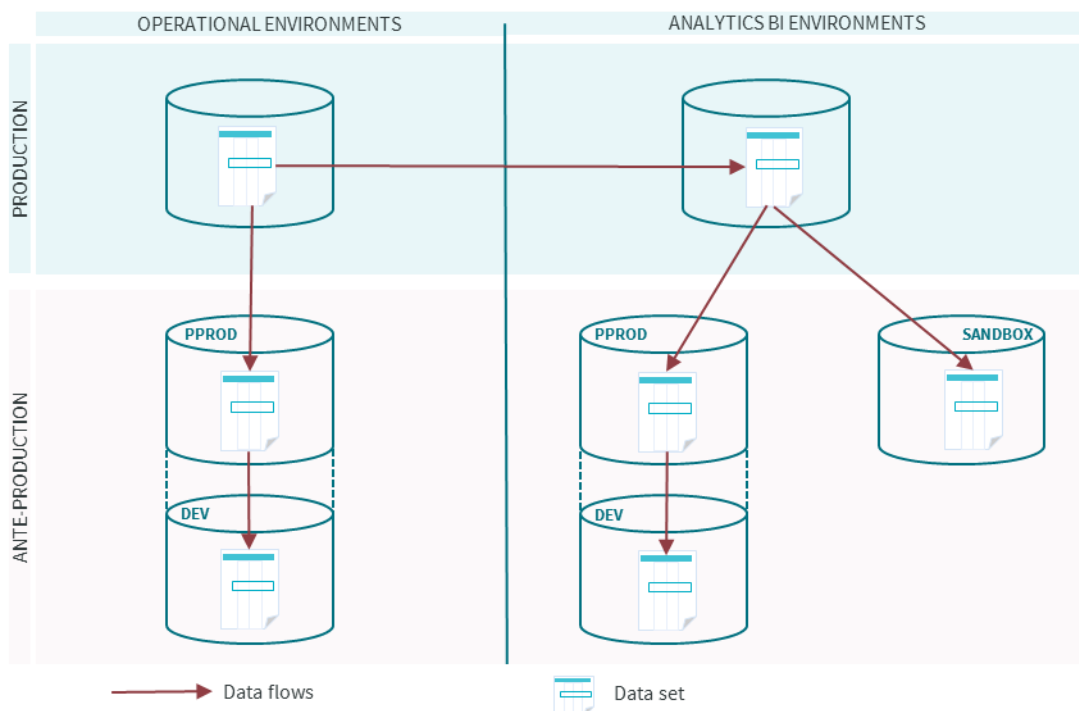


Figure 2 - sample data lineage

As you can see, for all the environments the data is copied in order to have real data to test upon in the dev env, the same amount of data to do load testing...⁴

Once the basics are settled, how will “the right to be forgotten” be distributed on the other environments.

A. Operational Environments

Operational environments are made for dealing with up to date data. You sometimes have archived data but as it has been seen in the GDPR, archiving is not a legal solution to apply the right to be forgotten, or at least, even if it increases the retention period, it has to have an end date. Is it worth anonymizing or pseudonymizing? The answer is usually no. According to the cost of

⁴ In real life, there are usually no automatic flows between production and ante-production environments. It is often a mix of copies, generated data, pseudonymized data. For the purpose of the exercise, the schema has been simplified with regular flows, automated or not.

anonymization/pseudonymization, the usage you will have of the data in these environments, the performance needs and sometimes storage capabilities, the only viable solution is DELETION.

So, you delete on the production environment, how about the ante-production environments. Well, you have also to delete the information on these environments.

Usually, in a normal lifecycle, the ante-production environments contain a certain amount of “real data”. For “risk reasons”, you can have pseudonymized/anonymized some of the personal data but not all the identifiable data (especially if you work with partners for maintenance). It is better to make sure the data in your ante-production reflects the production. So once the data has been deleted in your operational environment, you have to spread it to your ante-production environments (whatever the technic: replace data, destroy existing data and recreate data out of real ones or out of test data....).

B. Analytical Environments

Some use cases need personal data to be relevant in your analytical journey but some of them can work very well on anonymized data sets. For BI/Analytics environments, Anonymization/Pseudonymization is worth studying.

For non-anonymized dataset, the right to be forgotten can be spread by anonymizing the record (see section on Anonymizing data within the same dataset) or deleting the record. On anonymized dataset you do not have to consider the right to be forgotten, it is already applied. Anonymization is one way of applying Privacy By Design.

However, you also have ante production environments in BI. In the sandbox environment, you have sometimes many copies of data, for research needs, on which the right to be forgotten also needs to be applied. If Data Scientists only deal with anonymized data, there is no problem. In order to propagate on ante-production deletion (phase 4), one of the many solution could be simply to erase the old dataset on the ante-production (and sandbox) and copy it again from the production environment. Do not forget that ante-production environments should always be considered as temporary storage of data and cleaning them regularly should be in your operational process.

The following figure will show you the overall data lifecycle.

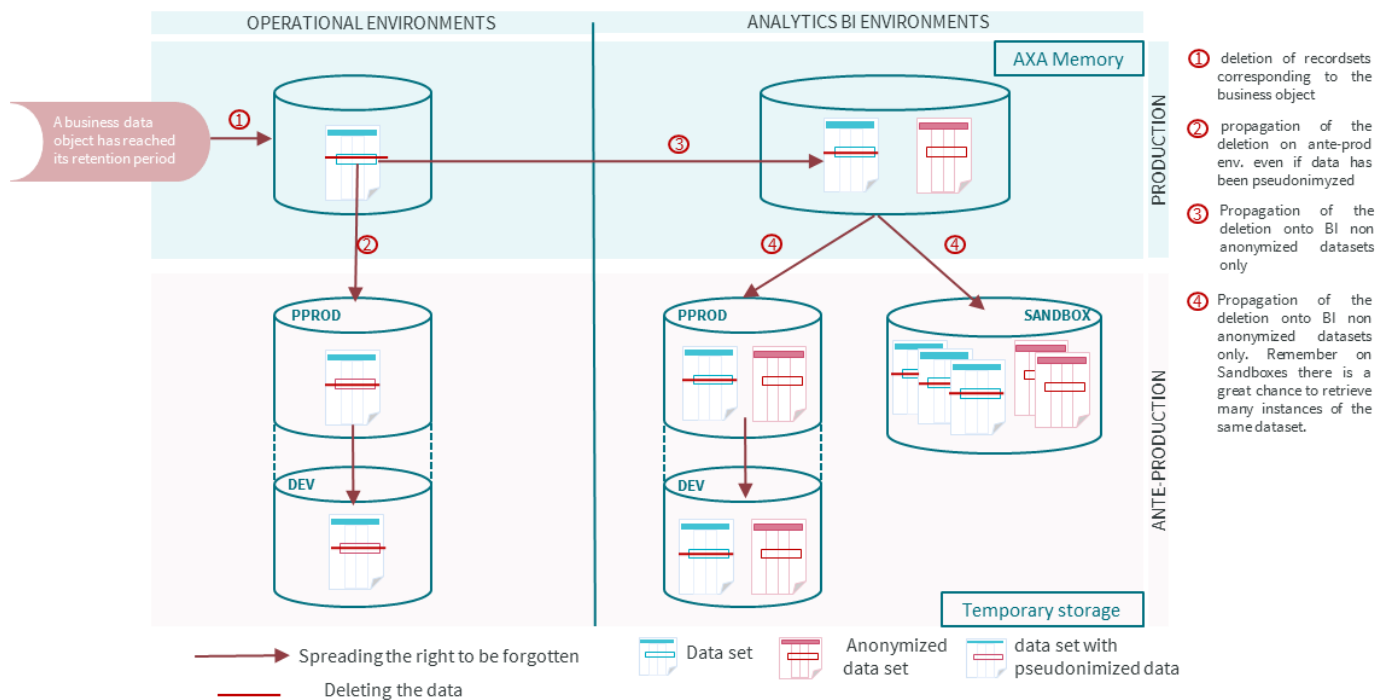


Figure 3 - Overall data lifecycle for GDPR

The BI/Analytics environments can be considered as the Memory of AXA whereas the ante-production should only contain temporary storage of data, time depending on the need to create and train your model. Usually you need a certain range of data to train your model and not all the data.

In the next section, we will see the overall process that has to be applied in order to assess the need for anonymization and then undertake anonymization.

<p>Important</p>	<p>To legally have an AXA Memory, then you should envisage Anonymization on your sources AND on the results of your Use Cases (usually a Business View). In order to enforce Privacy By Design, one method would be to get the results of your Use Cases in anonymized Business View datasets.</p> <p>If the Business Case allows it. For instance, for the next best offer Use Case, usually you need to have identifiable data to market your offer. In case of global reporting, you should not need to store direct personal data to calculate aggregates.</p>
-------------------------	--

<p>Information</p>	<p>This previous section was only focusing on the “right to be forgotten” and how anonymization can be a way to answer this need.</p> <p>However, Anonymization, and for this need, Pseudonymization can also be tools to answer the "Purpose limitation & Data Minimization" GDPR rule.</p> <p>Depending of the purpose and the associated Legal justification, the Individual has the Right to object or the consent should be taken into account. In the BI/analytical environments, we sometimes want to process personal data for a second purpose that could require a consent. Anonymization / Pseudonymization techniques are required in those cases.</p>
---------------------------	--

Section 2

Global Process for undertaking Anonymization

As a reminder, “Anonymization makes it **definitively impossible** to identify a particular individual”. It works by applying **irreversible** “techniques” on columns’ field identified as “quasi-identifier” or “unique identifiers”

We will see the various techniques in the next section. Now, we aim at understanding how we can apply anonymization by following specific steps in order to obtain an anonymized recordset/dataset. As you can see on Figure 4, it should be an iterative approach.

All along this process, you must keep in mind that there is a balance to find between the benefits in terms of risk, the cost of anonymization and the potential loss of value for the Business Case.

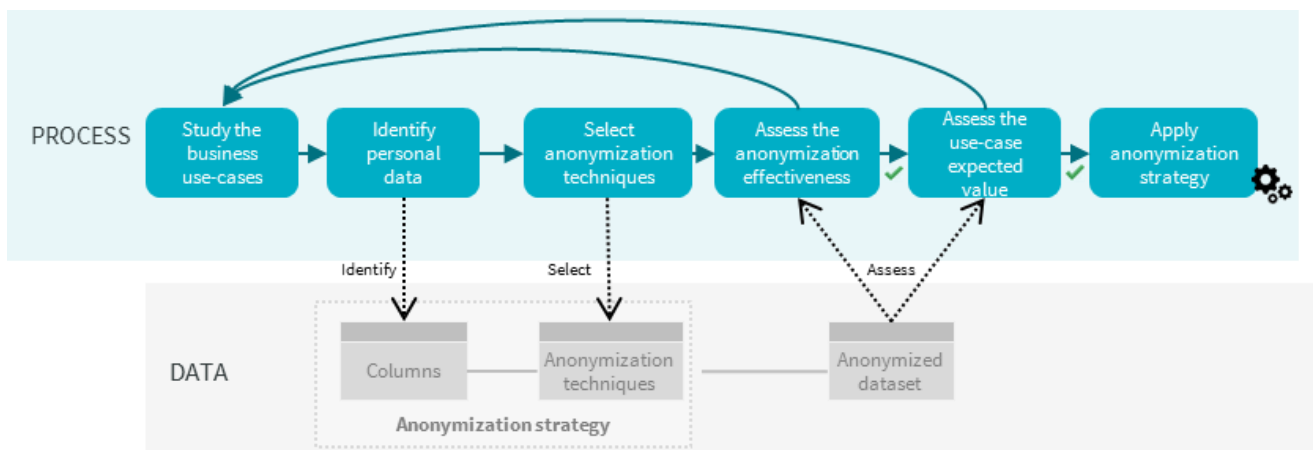


Figure 4 : Global process

1. Step 1: Study the business use cases

One of the major issue facing deletion of data is the fear of losing information. However, when you ask the simple question: “Why do you want to keep that information?”, the most common answer is: “In case” or “You never know”. In the case of GDPR, this answer cannot be and the regulator **will not** accept it.

Therefore, it is critical to clearly identify the business use case together with the related data to justify the need to keep personal data.

One of the common business use case encountered concerns operational reporting. The situation is this one:

an operational reporting on whatever subject is done on BI with a data-wizard (e.g. Business Object, Spotfire). Let us imagine we have a reporting done on a business object. This business object

has a retention period of 12 months. For the reporting, a sliding window of 13 months is necessary to follow the evolution of your indicators/KPI. The most common need will be: “I need at least 13 months of data”. The real need is: “I need to keep the results of my calculus for at least 13 months”. Whereas the first expression of need will indeed make you change the retention date, the second need only stores anonymized data since not pointing out a particular person. Of course, that will have an effect on the information system, another dataset to store, on the reporting itself, so that means an additional cost. However, if the reporting is of business key importance, it might be worth the cost of anonymization.

So, understanding the real need and therefore the potential data to be anonymized data is the entry point into the whole process⁵.

2. Step 2: Identify personal data and baseline risk

Once the use case(s) is(are) identified you can then identify the dataset(s) you need to anonymize or the anonymized dataset that you need to construct in order to run your use case(s).

In this(ese) dataset, you will have to identify the personal data, either identifiers (direct personal data) or quasi-identifiers (undirect personal data).

Be careful some data may not appear personal data at first but may contain some. For instance, comments field usually contains more personal data than expected. It is even considered as a sensitive data. GPS tracking coordinates are definitely personal data.

If you have datasets which should keep a link between them even after anonymizing the data, identify the primary key. If this primary key contains personal data, you will have to anonymize it. Our recommendation would be to use the Hash technique applied on both datasets to link together. Start your risk estimation based on the type of personal data you need to use and the sensitivity of the data, in the same kind of way you did it for the PIA (Privacy Impact Assessment) or the Information Security Classification Process.

3. Step 3: Select anonymization techniques with related risk reduction

For each selected personal data, then you will have to select the anonymization technique (cf Anonymization techniques section) you want to apply according to several factors:

- **The needs of the use case:** according to the data the need for the use case may be different. For example: in order to evaluate some samples in the test data, it is easier to talk about a person by saying his name and first name, even if they are bogus names rather than referring to that same person with his customer number. In that case randomization⁶ on the name and the first name is the right anonymization technique. Moreover, for some use cases, the salary may be an important feature to take into account, in that case you could use discretization⁷ (generalization technique applied to numbers by transforming defined numbers into ranges of numbers). In that case, you should see which ranges are needed for the use case to be efficient.
- **The tools you have chosen to perform your anonymization:** some techniques may not be available in all the anonymization tools. For some personal data, only techniques of adding noise to the data is really effective, but this technique is not spread in usual tools. Moreover, you usually have to deal

⁵ See Axa Belgium Approach

⁶ cf Randomization

⁷ cf Generalization

with both new data and legacy data. In that context, different tools might be available/used to achieve anonymization. Different tools mean different techniques and sometimes different results.

- **The type of personal data you are dealing with.** Indeed, identifiers should be removed or at least masked definitely. This means using either basic data masking or randomization. At one exception, the primary key of the dataset which should have been hashed⁸. For the quasi-identifiers, you have much maneuvering space because the technique should be applied according to the use case. Except in defined use cases, avoid keeping sensitive data directly readable or non-anonymized in your dataset. One of the method that could be applied would be to create a different dataset containing all the sensitive data and restrict access to this dataset.

In the section “List of personal data and sensitive data and preferred anonymization techniques to apply”, we will give you a list of personal data and sensitive data and will give you advices on which anonymization technique can be applied. However, this section gives only advices and cannot be a general rule. The business case must lead the anonymization technique to apply taking into account the limitations due to the used tool and the regulation.

4. Step 4: Assess the anonymization effectiveness

In order to assess the anonymization effectiveness, here are the points you need to focus on.

1. You have to keep in mind that the anonymization has to be effective within your environment and the data sources you deal with in your BI considering also the security measures that have been taken to reduce data leakage risks (what about if you are working with external providers, on cloud solutions....).
2. Your anonymization should reduce of the risks of re-identification such as Singling out, Linkability, Inference.
3. The data processed should lose sensitivity in case of data leakage

Let us assume that you have taken some safeguards against data leakage (risk 0 does not exists, anonymization is one way of reducing the risk).

You then have to assess points 2 and 3.

That is where you should ask you DPO to validate your anonymization strategy or see with a security expert whether they think your anonymization can guaranty a certain level of anonymity. He might also advise you to launch certain tests. You can not be the prosecutor and the judge, you have to get a second opinion.

However, there are certain tests based on the k-anonymity⁹ (and its sisters) that can be performed quite easily with basic sql techniques (group by, order by).

As seen before, only definite masking technique should be applied to the identifiers, however the question of singling out still stands on quasi-identifier.

In the Ensuring anonymity: K-anonymity, L-diversity paragraph, we will go through an example of de-anonymization which shows that which quasi-identifiers a researcher, Sweeney, have re-identified the governor of the state by using his birthdate, zip code and sex. That is why one of the

⁸ cf Hashing

⁹ Cf K-anonymity

main basic assessment would be to take all the quasi-identifiers within the list and see whether you can obtain a group of one person.

Take a sample of your dataset and try to apply the anonymization techniques you have defined. With simple sql, you can test the level of your k-anonymity with the count and group by function.

The minimum group count gives you level of anonymity you can hope to reach. If the minimum group count is 3, you are 3-anonymous...

The more quasi-identifiers you take to count your group, the less anonymous the groups are. If your level for anonymity is below a certain threshold (depending on the size of your dataset), then you should review your anonymization techniques. e.g. if your level is 3, while your dataset contains only 10 records, you have reached a very good level. However, for 1 000 000 records 3 is very low.

This level should be validated with your DPO in link with Risk Management.

5. Step 5: Assess the use-case effective value

This step is useful in order to assess two things:


First, is the cost of anonymization worthwhile taking into account the original use case. This is only a validation step because the ROI of the use case should have been calculated in step 1.


But most important matter, is you anonymized data worthwhile concerning your use case. Indeed, the data can be too anonymized and therefore the anonymous data is useless to get value out of your use case. So, check it before implementing the defined strategy.

Do not forget to confront that to the risk assessment, you should have started on step 2.

6. Step 6: Apply anonymization strategy

That's it! You have validated your approach! You can now implement it in your system. You will see in section Ways of Anonymizing data within the same dataset, different ways of implementing your anonymization strategy. It is really a matter of implementing anonymization at a resordset level or at a dataset level.

 Axa France Best Practice Ideas	<p>Every anonymization depends on the context (internal, open data, shared data, BI, data lake). In some contexts, you could apply a "light" anonymization (level of risk high) but this should be linked to a strong code of conduct regarding data for the professionals using these data (French actuaries).</p> <p>Moreover, your process for anonymization can be iterative. The aim is to avoid having such a strong anonymization that the data is value-less for the market (e.g. reinsurance). The idea behind it would also to be able to justify the approach towards regulators.</p>
---	--

 Axa Belgium Approach	<p>In BI and DL, usage of the data is generic. The approach towards anonymization of data should be very cautious and strongly linked to governance. Usually, Use Cases are not known so applying this process can be risky and lose the value of data if too much anonymization is applied.</p> <p>In order to answer the purpose limitation and data minimization need, Axa Belgium plans to use a tokenization tool. Through this tool, profiles can be set in advance depending on the use case purpose. To manage accesses through these profiles, a strong governance process has to be set-up.</p> <p>As for anonymizing the data, Belgium has chosen an iterative approach in order to grow in maturity along with the various use cases coming.</p>
---	--

Section

3

What are the various techniques you can use to get anonymized data?

Extract from [Wikipedia](#) :

"Data anonymization has been defined as "technology that converts clear text data into a nonhuman readable and irreversible form, including preimage resistant hashes (e.g., one-way hashes) and encryption techniques in which the decryption key has been discarded. ... Anonymized data refers to data from which the person cannot be identified by the recipient of the information. The name, address, and full post code must be removed, together with any other information which, in conjunction with other data held by or disclosed to the recipient, could identify the patient.^[2]

De-anonymization is the reverse process in which anonymous data is cross-referenced with other data sources to re-identify the anonymous data source."

If you consider this definition, there is a paradox. Indeed, if anonymization is successful, then de-anonymization cannot be foretaken since anonymization is irreversible.

However, it has become a challenge for some hacker or mathematician wiz kid to take a set of anonymized data and try to de-anonymize it. There are a lot of samples of this kind of de-anonymization:

- ✓ It takes now 3 points in time and space to identify a person
- ✓ By only taking the search's subjects on an engine (google, yahoo...), you have 80% chance of being identified.
- ✓ The MD5 algorithm has been cracked and is now considered as not secure enough.

If you want to anonymize a set of data, there is a great chance that if you want to make it irreversible forever you will end up with no data at all. Depending on your context, the anonymization you will apply will be more or less string. If a data has a low sensibility, you could stick to a low-level anonymization (e.g name, surname, address). Your business context could lead you to propose only low-level anonymization (e.g. anti-money laundering, fraud). The anonymization level should be validated through a risk assessment -at least oral- with your DPO.

We will then start from the postulate, that the anonymization done on data is irreversible within AXA's entities data sources.

Moreover, even if anonymization is not perfect, it is important to first focus on PII (Personally Identifiable Information)¹⁰, then, according to the context or the risk, focus on indirect data. The following paragraph will give you a definition of all the anonymization/pseudonymization techniques that we will advise you to use during the whole process.

Indeed, some techniques although very efficient may have some drawbacks (no tools will provide such a technique, difficulty to implement it...).


1. Anonymization techniques

The aim of this paragraph is to describe various types of models of anonymization, seeking to hide or break the connection between a person in the real world, and its personal or sensitive data. We will not address all the different models that can exist.

Two main families stand out from anonymization algorithms depending on their capacity to refer to unique and single records or not:

Anonymization deterministic algorithms are those which for a given data/data set to anonymize provide an identifiable/unique anonymized data/data set. As a consequence, data related to an individual are no longer readable but it is still possible to associate it to an anonymized data subject. NB: the algorithm is not persistent, otherwise it would be pseudonymization

Anonymization undeterministic algorithms are those which for a given data/data set to anonymize provide an aggregated/generalized anonymized data/data set. As a consequence, it isn't possible to be able to isolate a given data subject within a data set.

 Information	<p>Samples of Anonymization deterministic algorithms: K-anonymity and its extensions L-diversity, T-closeness, Perturbative masking such as noise addition, microaggregation, data swapping or post randomization</p> <p>Samples of Anonymization undeterministic algorithms: Differential privacy, Non-perturbative masking like Sampling, Generalization, Top/bottom coding, Local suppression</p>
--	--

A. Data Masking

Data masking means replacing a data by another. There are 3 types of data masking :

- **Definitive masking**: the data in a field is definitely changed inside the record by changing the value or masking part of the data (anonymization technique)
- **Reversible masking**: the data in a field is changed inside the record by changing the value or masking part of the data but you have a way of retrieving the original value (pseudonymization technique)
- **Dynamic masking**: the dynamic masking is a way of restricting access to certain fields in a table by adding a table of rights. This masking technique is relative to the user (UI level) and does not imply any change

¹⁰ Can be referred as desensitization

to the original data (storage level). It is not an anonymization technique, but rather a way of restricting accesses to sensitive data.

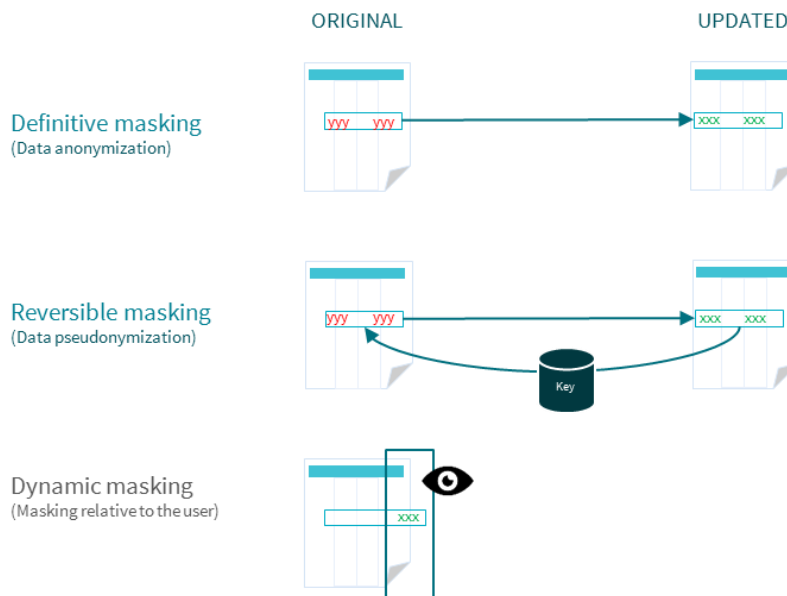


Figure 5 : Typology of data masking

B. Basic Data Masking

Among the deterministic algorithm you can find the basic data masking technique, which is simply to avoid seeing the data value by masking its value by either nothing or always the same value for each record inside the table.

ORIGINAL						ANONYMIZED					
id	name	address	type	age	salary	id	name	address	type	age	salary
1	John Doe	33 rue ...	Client	29	127 K	1	Unknown		Client	29	12 K
2	2	Unknown		12 K
3	Jane Doe	29 rue...	Prospect	...	138 K	3	Unknown		Prospect	...	12 K
4	4	Unknown		12 K
5	5	Unknown		12 K

Figure 6 : Basic data masking

In Figure 3, the name has been masked by the value “Unknown”, the salary by the value “12 K” for each record of the dataset, while the address has been erased/replaced by an empty string.

C. Randomization

Randomization is the process of exchanging the values of attributes by other values contained inside another list of records. You can also talk about data swapping if your list of definite records is based on your original table.

ORIGINAL								ANONYMIZED							
id	First name	Last name	address	zipcode	type	age	salary	id	First name	Last name	address	zipcode	type	age	salary
1	John	Doe	33 rue ...	92400	Client	29	127 K	1	John	Bell	33 rue ...	92400	Client	29	127 K
2	2
3	Jane	Martin	29 rue...	75015	Prospect	46	138 K	3	Jane	Donny	29 rue...	75015	Prospect	46	138 K
4	4
5	Roger	Dumont	89 street...	75020	Client	22	89 K	5	Roger	Sector	89 street...	75020	Client	22	89 K

Figure 7 : Sample of Randomization

In figure 6, all the last names have been changed at random by another list of names. This replacing list of names can be created out of any source of names, as long as the picking algorithm (algorithm which chooses the replacing name) is not reproducible and does not depend on the source data. Randomization can be applied at record or at table level.

ORIGINAL								ANONYMIZED							
id	First name	Last name	address	zipcode	type	age	salary	id	First name	Last name	address	zipcode	type	age	salary
1	John	Doe	33 rue ...	92400	Client	29	127 K	1	John	Dumont	33 rue ...	92400	Client	29	127 K
2	2
3	Jane	Martin	29 rue...	75015	Prospect	46	138 K	3	Jane	Doe	29 rue...	75015	Prospect	46	138 K
4	4
5	Roger	Dumont	89 street...	75020	Client	22	89 K	5	Roger	Martin	89 street...	75020	Client	22	89 K

Figure 8 : Sample of Data-swapping

The data-swapping is rather applied at table level.

D. Generalization

Generalizing means removing a “degree of precision” to certain fields. This technique has the effect of grouping the data according to a more general value than the original value.

ORIGINAL								ANONYMIZED							
id	First name	Last name	address	zipcode	type	age	salary	id	First name	Last name	address	zipcode	type	age	salary
1	John	Doe	33 rue ...	92400	Client	29	127 K	1	John	Doe	33 rue ...	92XXX	Client	[0-30]	[100-150 K]
2	2
3	Jane	Martin	29 rue...	75015	Prospect	46	138 K	3	Jane	Martin	29 rue...	75XXX	Prospect	[31-50]	[100-150 K]
4	4
5	Roger	Dumont	89 street...	75020	Client	22	89 K	5	Roger	Dumont	89 street...	75XXX	Client	[0-30]	<100K

Part of the data has been masked in order to change the zipcode into the department code

Data has been changed for groups of data rather than specific data

Figure 9 : Samples of generalization

In this sample, the age field is replaced by an interval of age (discretization).

With this technique, some data types may not be the same. Usually, you would create another data field with the right type and apply your generalization to this new column.
For instance, a department column could be added and take as a value only the first two digits of the zip code.

Generalization is one of the key technique of the **K-anonymity** and its sister **L-diversity**.

E. Hashing

The hashing technique is based on specific deterministic algorithms, it is a pseudonymization technique.

A hash function is a function that allows to transform a data value into a specific character set that keeps the relationships between the source objects (equality, order...) but without accessing the value directly.

One of the most common hash algorithm is called **MD5**. This algorithm is no longer considered as safe, so we would advise you to use at a minimum **SHA2**.

2. Ensuring anonymity: K-anonymity, L-diversity¹¹

A. K-anonymity

In 2002, an American researcher, Sweeney, could prove that the combination of other fields than direct identifiers can help retrieve the individual concerned.

The strategy was simple. Sweeney took 2 different sets of data (one of which was an anonymized health data, the other one came from the list of voters in the state) and linked them together by a triplet of value (zip code, the date of birth and sex) which is unique for almost 80% of the population in the US. She could then link health data to individuals (one of them being the governor of the state)

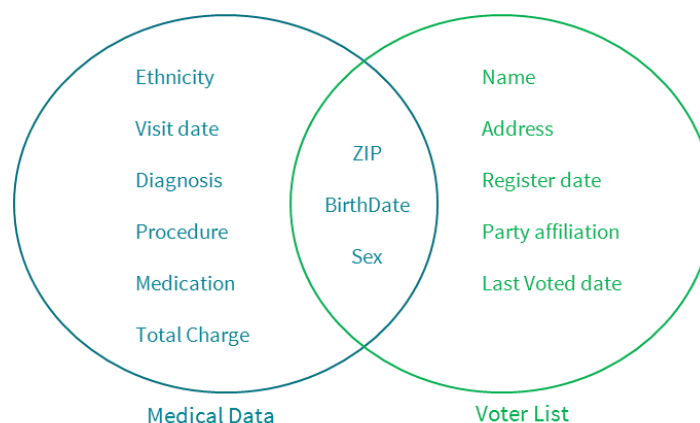


Figure 10 Sample of singling out in an anonymous dataset (source Sweeney 2002)

To protect against this type of attack, called record linkage, Sweeney has proposed the technique of k-anonymity. This will blur the ability to link a n-anonymous tuple to a n-tuple not anonymous in the following way:

- 1) determine the sets of attributes (referred to as quasi-identifiers) that can be used to cross the anonymous data with identifying data; then

¹¹ Almost all of this section is based on the paper “Techniques d’anonymisation, by Benjamin NGUYEN” available on the web, examples included

- 2) reduce the level of detail of the data so that there are at least k n-different tuples that have the same value of quasi-identifier (apply generalization techniques on the quasi-identifier)

The advantage of the k -anonymity is that analysis of the data will continue to provide accurate results, this close that one cannot separate individuals in a group.

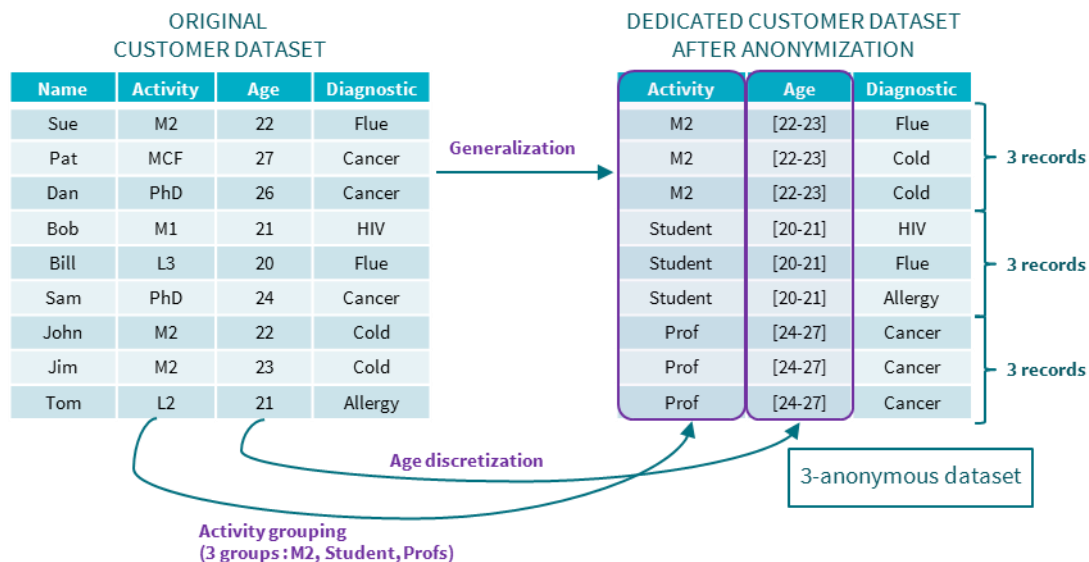


Figure 11 Anonymization on data from a university

An important technical problem remains achieving k -anonymity: be able to determine generalizations to apply on quasi-identifiers, what can be done either by a human expert who knows the area, or by a computing, often very costly in a real database.

B. L-diversity

In example showed in figure 8, you can easily deduce that Sam who is a PhD aged 24 has a Cancer because all the professorial team has a cancer. It is thus possible to infer information in some cases, without making the slightest crossing, for example if all the individuals in a class have the same value. The L-diversity model addresses this problem, by adding an additional constraint on the equivalence classes: not only at least k n-tuples must appear in an equivalence class, but also the sensitive field associated with the equivalence class must take at least L distinct values.

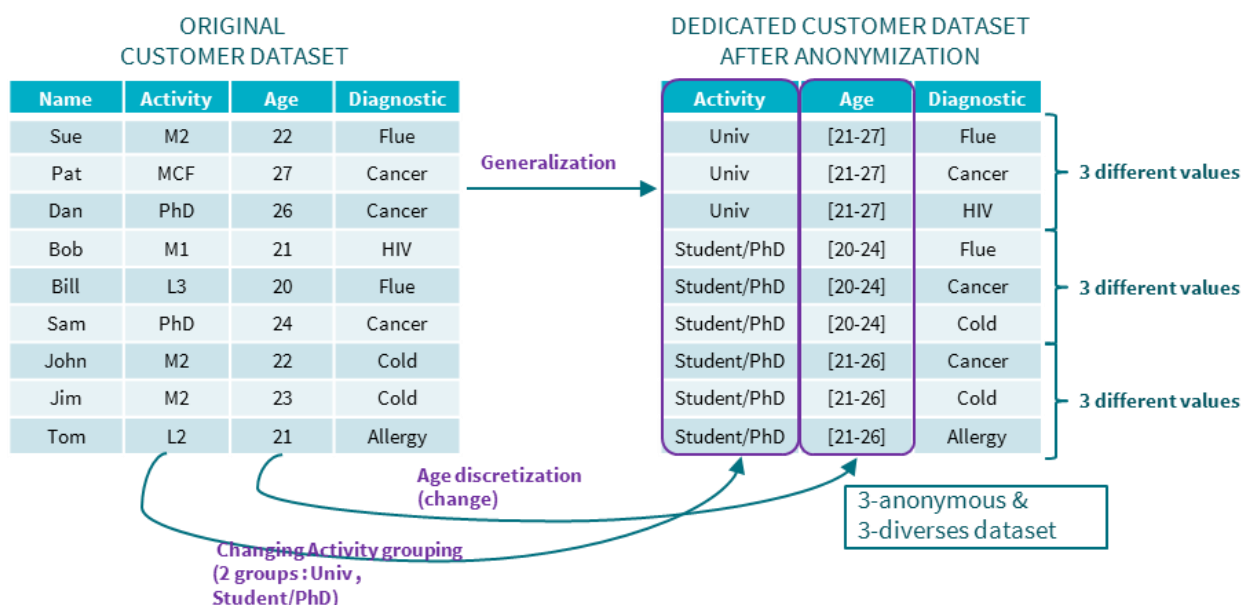



Figure 12 L-diversity

However, by leading an attack by crossing of the same type as that of Sweeney, there is still a possibility to deduce information. In the L-diversity sample, we can then deduce that a 20-year-old student will have a 33% chance to have the flu, 33% chance to have the cancer and 33% chance to have a cold, and especially no chance of having another pathology. If we know that Bill is the only person from the base in this case, then you can deduce sensitive information about him.

Even though these techniques are not fail-proof¹², they insure the minimum anonymity and are practical to set up. We did not talk about differential confidentiality because it involves a lot of computing and cannot fit our context, as of today.

Now that we are aware of the techniques, we will see in the next section the overall process to undertake anonymization.



Information

The G29 has evaluated strengths and weaknesses of each technique towards the 3 risks: singling out, linkability, inference. The result is presented in the following table.

	<i>Is Singling out still a risk?</i>	<i>Is Linkability still a risk?</i>	<i>Is Inference still a risk?</i>
<i>Pseudonymization</i>	Yes	Yes	Yes
<i>Noise Addition</i>	Yes	May not	May not
<i>Substitution</i>	Yes	Yes	May not
<i>Aggregation or K-anonymity</i>	No	Yes	May not
<i>L-diversity</i>	No	Yes	May not
<i>Differential privacy</i>	May not	May not	May not
<i>Hashing/Tokenization</i>	Yes	Yes	May not

¹² See G29 Matrix of strengths and weaknesses of anonymization techniques

Section

4

List of personal data and sensitive data and preferred anonymization techniques to apply

In this section, we will give you some samples of applying anonymization techniques to data, Identifiers, Quasi-Identifier and even sensitive data.

This list is far from complete but it may help you find ways facing a data to understand which is the best.

We will also talk about another technique (not stated earlier): How about creating features? Indeed, specific dates can be quite “identifiable” data. Instead of keeping a start date and an end date, how about creating the feature nbofday, which will tell you how long it took between opening and closing an action/contract/claim....

We will also consider that you have no data which relates to origins, ethnics, religion, criminal record, health data. Remember that there are usually strict rules for that kind of data and you have to see with your DPO before considering storing them.

Remember also that some data which seems non-direct or even quasi-direct can store a lot of sensitive and personal data. This is the case with comments, unstructured data like text files, emails.... Be sure to assess those also before leaving them aside.

personal data definition	type of personal data (identifier/quasi-identifier)	preferred anonymization/pseudonymization technique	other potential technique	Comments
AXA unique identifier	Identifier	Hashing	basic data masking	Hashing is preferred to basic data masking to ensure the linkage with other databases
full name, maiden name, surname, first name....	Identifier	basic data masking	randomization	
Customer National ID number	Identifier	basic data masking		
Customer Social Security number	Identifier	basic data masking		
gender, sex, title	Quasi-Identifier	None	generalization	Usually this field is a defined list of value.

Occupation Name	Quasi-Identifier	generalization		If this field is not a defined list of value, create one rather based on type of Occupation. Moreover, if you have a field comments to specify an occupation which does not belong to the list, you should apply basic data masking to this field
Highest attained level of education	Quasi-Identifier	generalization		Usually already based on a list of defined values
marital status	Quasi-Identifier	generalization		Usually already based on a list of defined values
Customer full residential address	Identifier	basic data masking	generalization	A main address could be transformed into a square based on National Statistical Data, or other non-discriminant squares (IRIS)
zip code	Quasi-Identifier	generalization		have the zip code transformed into a department number. e.g: 92960 -> 92XXX
City	Quasi-Identifier	generalization	basic data masking	have the city code transformed into a department number. Colombes -> Haut de Seine or 92
country of residence	Quasi-identifier	generalization		usually based on a list of defined values. If your diversity of customer is 99% in one country and 1% around the world, you could envisage to turn your list into 2 variables, one for the "country", the other corresponding to "rest of the world"
department	Quasi-Identifier	generalization		should be disperse enough
Customer full date of birth	Quasi-Identifier	generalization	basic data masking	A date of birth is quite touchy. Generalization should be applied. You could keep the year of the date of birth, in many cases it should be ok but be careful on the diversity you have. For every type of date, you could envisage discretization. E.g. date of birth: 26/01/1980 with generalization you could have: 1980, with discretization, you could have : [1980-1990]
Age	Quasi-Identifier	discretization (generalization)		you should have range of ages in which you would assign according to the real age.
Customer email address	Identifier	basic data masking	randomization	
mobile, work, home phone number	Identifier	basic data masking	randomization	

Customer driving license details (obtention date, type, etc.)	Quasi-Identifier	basic data masking		
Customer IBAN details	Identifier	basic data masking	generalization	If some fields are important like the type of bank, maybe you could keep this information but be very careful and check with DPO because not only this is an identifier data but also a sensitive one
Customer wealth	Quasi-Identifier	discretization (generalization)		sensitive data
Details on Customer personal loans	Quasi-Identifier	discretization (generalization)		sensitive data
Balance held in deposit accounts / cash like assets	Quasi-Identifier	discretization (generalization)		sensitive data
Value of home	Quasi-Identifier	discretization (generalization)		sensitive data
Outstanding mortgage balance	Quasi-Identifier	discretization (generalization)		sensitive data
Employment status / occupation	Quasi-Identifier	generalization		Usually already based on a list of defined values. Consider if there is a comment added to this field, if yes please apply basic data masking on the comment
Current health status	Quasi-Identifier	basic data masking	discretization (generalization)	sensitive data
income can be from labor, property ownership,	Quasi-Identifier	discretization (generalization)		sensitive data
Balance in tax qualified savings plans	Quasi-Identifier	generalization		sensitive data
Details on Customer personal revolving credit	Quasi-Identifier	discretization (generalization)		sensitive data
Details on Customer savings	May be unstructured	discretization (generalization)		sensitive data
Details on Customer credit card	Identifier	basic data masking		highly sensitive data
Customer location	Identifier	basic data masking	generalization	If you are working with GPS coordinates, please consider generalizing with square
Standardized VIN (identifier + interpretable for features)	Identifier	basic data masking		

Vehicle license plate identification	Identifier	basic data masking		
Policy first registered Date/ Duration	Quasi-Identifier	generalization		try keeping only the year, or consider ranges
Policy Terminated Date	Quasi-Identifier	generalization		try keeping only the year or consider ranges
Balance of Loan Accounts	Quasi-Identifier	discretization (generalization)		sensitive data
Date of fist contract underwritten with AXA	Quasi-Identifier	generalization		try keeping only the year and consider ranges
Web browsing activity	Quasi-Identifier	basic data masking	generalization	sensitive data, with short retention dates usually
National Produced Number of distributor	Identifier?	basic data masking	None (depending on DPO)	when you talk about companies the National Identification Number is not consider as a personal data since it concerns a moral person. However, be very careful because for some small companies the manager's info is very touchy. So go back to your DPO to define with him the rules to apply
Distributor Name	Identifier	basic data masking		
Address of distributor in contact w/ Customer	Identifier?	generalization		Same thing, if the address is a company address then no problem but it could be a personal address as well.
Customer Socio-economic Classification level	Quasi-Identifier	generalization		Usually based on a list of values.
Full list of email communications w/ AXA	Identifier/Quasi-Identifier	basic data masking		sensitive data. Be careful with the content of the documents. If it is only a reference then hash it, if it is more see what technique could be applied on unstructured data
Full list of documents exchanged w/ AXA	Identifier/Quasi-Identifier	basic data masking		sensitive data. Be careful with the content of the documents. If it is only a reference then hash it, if it is more see what technique could be applied on unstructured data
Full list of email web chats and IM w/ AXA	Identifier/Quasi-Identifier	basic data masking		sensitive data. Be careful with the content of the documents. If it is only a reference then hash it, if it is more see what technique could be applied on unstructured data
Web browsing activity	Identifier/Quasi-Identifier	basic data masking		sensitive data. Be careful with the content of the documents. If it is only a reference then hash it, if it is more see what technique could be

				applied on unstructured data, there are many comments within web browsing.
Social network activity	Identifier/Quasi-Identifier	basic data masking		sensitive data. Be careful with the content of the documents. If it is only a reference then hash it, if it is more see what technique could be applied on unstructured data, there are many comments within social activity.
ID of underwritten policy	Identifier	Hashing	basic data masking	Hashing is preferred to basic data masking to ensure the linkage with other databases
effect date of policy	Quasi-Identifier	generalization		consider keeping the dates into a year or even create another field which will enclosed the number of days the policy has been opened.
end date for policy	Quasi-Identifier	generalization		
In case Customer added or removed coverage features	Quasi-Identifier	basic data masking	generalization	keep track of the nb of times the coverage has been changed rather than the date of last change. Otherwise you can also keep the year.
Gross underwritten premium	Quasi-Identifier	discretization (generalization)		sensitive data
In case Customer did not pay entirely	Quasi-Identifier	discretization (generalization)		sensitive data
ID of recorded claim	Identifier	Hashing	basic data masking	Hashing is preferred to basic data masking to ensure the linkage with other databases
Date of FNOL	Quasi-Identifier	generalization		
Claim label	Quasi-Identifier	generalization		Usually a short description. Try considering types of claims rather than hand-written description
Date of last change occurred on claim	Quasi-Identifier	generalization		consider adding nb of changes in the claim rather than a date
Date of closure (if closed)	Quasi-Identifier	generalization		consider adding nb of days till closure rather than a date
Current estimation of final cost		generic		
Claim being investigated		basic data masking		sensitive data. Make sure the individual is not recognizable at all within your data because the rules of fraud are sometimes quite strong
Suspected fraud on claim		basic data masking		sensitive data. Make sure the individual is not recognizable at all within your data because the rules of fraud are sometimes quite strong
Date and time of accident	Quasi-Identifier	generalization		

Location of accident	Quasi-Identifier	generalization		try generalizing with a square
(Textual) description of accident	Quasi-Identifier	basic data masking		if you can ensure that the text is not sensitive then keep it but if you can do without please apply basic data masking
Reported cause of accident	Quasi-Identifier	generalization		should be based on a list of causes. If comment is added to the cause, remove it. This list could help you mask the claim label field
Contact details people involved in the accident	Identifier	basic data masking		
If police reported on the accident, copy of the report	Identifier	basic data masking		
If accident report did not involve the police	Quasi-Identifier	basic data masking	discretization (generalization)	consider keeping only discretized data referring to your accident report nothing else.
Driver when the accident occurred	Identifier	basic data masking		
Breathalyzer value if measured		discretization (generalization)		sensitive
Breathalyzer delay after accident if measured		discretization (generalization)		sensitive
Reported damage type (bodily injuries, car damage, etc.)		generalization		Usually based on a defined list of values
Reported bodily injuries		basic data masking		try anonymizing this field, sensitive data
Initial assessment of claim valuation		discretization (generalization)		sensitive
In case an expert is ordered, expert ID	Identifier	basic data masking		
In case an expert is ordered, fees paid		discretization (generalization)		sensitive data
In case an expert is ordered, date of expertise	Quasi-Identifier	discretization (generalization)		sensitive data
Total cost of repair		discretization (generalization)		sensitive
if claim with bodily injuries		Hashing		not a sensitive data nor a quasi-identifier (except if the hospital was used once)
if claim with bodily injuries	Identifier	basic data masking		
Part of injury costs covering hospital fees		discretization (generalization)		sensitive data

Part of injury costs covering medical equipment		discretization (generalization)		sensitive data
bodyshop ID		Hashing		not a sensitive data nor a quasi-identifier (except if the hospital was used once)
bodyshop address	Quasi-Identifier		new feature: square location of the address	However, this is an address for a company so it could be kept.
bodyshop IBAN	Identifier	basic data masking		sensitive data
bodyshop legal status	Quasi-Identifier	generalization		usually a list of known data
bodyshop financial health	Quasi-Identifier	basic data masking		sensitive data. This information has a retention date in itself. It is not because you had a rough year 5 years ago that it has to have effects on you all the time. Basic definition of right to be forgotten.
number of bodyshop employees	Quasi-Identifier	discretization (generalization)		try considering discretization.
Repair order ID if given	Identifier	Hashing		
Repair payment amounts	Quasi-Identifier	discretization (generalization)		
Repair date	Quasi-Identifier	generalization		try considering new feature: nb of days to repair

Section 5

Various ways to create an anonymized dataset

Anonymizing a record is one of the solution to avoid deleting it. There are two possibilities to store the anonymized data: either within the same dataset, or within a dedicated dataset. This section will give an overview of these two possibilities and the impacts it can have towards the data structure in place and the use cases who are using these data.

CAVEAT: the following approaches will be challenged in the next coming months to industrial market best practices.

1. Anonymizing data within the same dataset: anonymization at a record level

This approach is currently investigated by AXA Switzerland.

Let's imagine an existing customer dataset. This dataset is in place for many years now. One or more entries are to be deleted because these customers have not had any contact and no open contracts for many years (past retention date).

Instead of deleting this entry, you will replace the existing entry by an anonymized dataset.

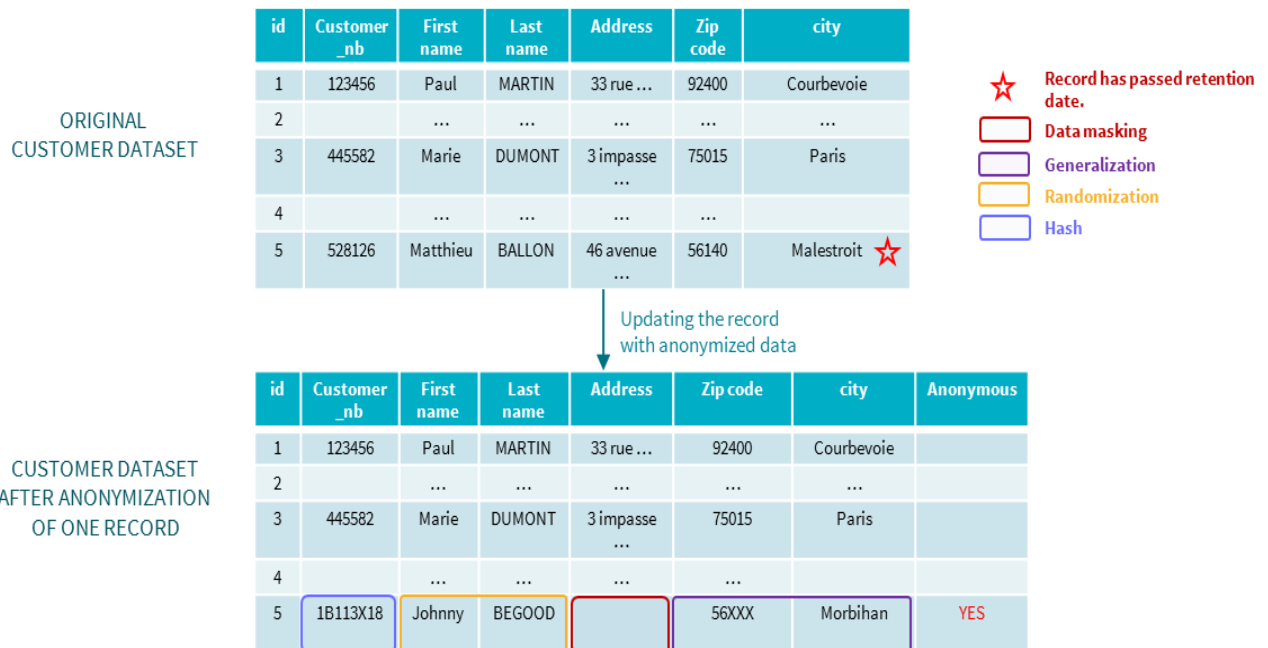


Figure 13 Anonymization within the same dataset

Many anonymization techniques have been applied, the customer cannot be re-identified. The anonymization is quite perfect (within AXA's data sources). However, there are some impacts on the

table structure. The first impact is that you have to add a flag to the existing table to indicate that the data is anonymized (in our example, the flag is called Anonymous). If you do not do so, and you have a use case assessing the quality of the address then you will have a problem of accuracy on record number 5. Indeed, the address of the customer does not exist anymore, precisely because of anonymization.

A. Impacts

The previous example is one of the many impacts you may encounter when applying this technique.

Impacts on :	How	Sample
Data structure	Add a field to identify anonymous data	
Use Cases	The added anonymous field must be taken into account by the existing Use Cases	Quality of the address use case
Field type	According to the anonymization technique the resulting field may be of a different type and thus can not fit in the existing table	If customer_nb has been set to a numeric type, the hash result may not fit: 1B113K18 is definitely not numeric
Business/technical constraint	When you deal with a table, there can be constraint on field applied to it. Those constraints may have to be removed to accept anonymized data	If zip code only accepts number the generalization may not be acceptable "56XXX". You could only keep "56". However, if you add a constraint saying that the zip code should have 5 digits, this constraint does not apply anymore
Data meaning	If you apply generalization, then the meaning of the data contained inside a field may change	The field city of the anonymized data is not a city anymore but a department. If you looked at it on a map of France's cities, you could not see it anymore.
Data Structure	According to the generalization you want to proceed with, you may add a new field to contain the real meaning of the data	To be more consistent, a data field called department-should be added to
Use cases	Each change in data structure may have an impact on Use Cases.	A department field has been added, however the Use Case X runs without naming the various useful fields. Use Case X may crash because it awaits 7 (or 8 fields with anonymous) but finds 9

Assessing the anonymization effectiveness	One of the key aspect of the anonymization effectiveness is to prove there is no singling out. If you have only few anonymized data in your dataset then the effectiveness can not be assessed right away	
--	---	--

B. Pros and Cons of this technique

PROS	CONS
No duplication of Data	Impacts on the table structure
No need to deal with legacy data/ No changes in entry flows	No dissociation between statistical use cases and operational use cases. Not the same access rights
No other actions needed when records are updated	Potential impact on use cases
	Difficulty to assess the effectiveness of an anonymization

2. Anonymizing data within a dedicated dataset: anonymization at a dataset level

Now, the view is different from before. Out of the first step of the anonymization process, you understood that there is a need for an anonymized view because you have some use cases dedicated.

The paradigm now is different because you know that there is a need for a specific anonymized view that can be specified with your requesters.

Let us assume that the table/dataset seen before is the dataset that needs anonymizing. You will then create a dedicated anonymized view containing all the data of the original view and you will apply your anonymizing method to insert the anonymized record into this new anonymized dataset. What is important to notice is that since it is a new dataset, the structure can differ according to your needs.

ORIGINAL CUSTOMER DATASET							DEDICATED CUSTOMER DATASET AFTER ANONYMIZATION						
id	Customer_nb	First name	Last name	Address	Zip code	city	id	Customer_nb	First name	Last name	Address	Department code	Department
1	123456	Paul	MARTIN	33 rue ...	92400	Courbevoie	1	4458FTRX	Paul	DUPONT		92	Hauts de seine
2		2	
3	445582	Marie	DUMONT	3 impasse ...	75015	Paris	3	13358ZH2	Delphis	ARTEM		75	Paris
4		4	
5	528126	Matthieu	BALLON	46 avenue ...	56140	Malestroit ★	5	1B113X18	Johnny	BEGOOD		56	Morbihan

Data masking

Generalization

Randomization

Hash

★

Record has passed retention date.

Figure 14 A dedicated Anonymized Dataset

As you can see, the address can be suppressed from the anonymized dataset since the data masking with blank has occurred. Moreover, the zip code and city have been replaced by department number and department name. You can then apply rules, field types dedicated to this dataset. In one data processing, you can deal with the legacy.

Now let us imagine that you have a customer entry that needs to be deleted because of the GDPR right to be forgotten rule (retention period is over). The resulting operation will be to delete the entry in the original table but nothing happens on the dedicated anonymized dataset.

id	Customer_nb	First name	Last name	Address	Zip code	city	id	Customer_nb	First name	Last name	Department code	Department
1	123456	Paul	MARTIN	33 rue ...	92400	Courbevoie	1	4458FTRX	Paul	DUPONT	92	Hauts de seine
2		2	
3	445582	Marie	DUMONT	3 impasse ...	75015	Paris	3	13358ZH2	Delphis	ARTEM	75	Paris
4		4	
5	528126	Matthieu	BALLON	46 avenue ...	56140	Molécotroit	5	1B113X18	Johnny	BEGOOD	56	Morbihan

4 records left

Still 5 records

Figure 15 Applying the GDPR deletion on both current/anonymized datasets

The deletion process can be applied only on the first table “operational view” the anonymized dataset will not be touched.

A. Impacts

The main impacts of such a technique is on entry data flows. Indeed, once created the dedicated dataset should be kept up to date as well as the first table.

- Any new entry in the operational view has to be reported into the anonymized view
- Any new update on operational records has to be reported into the anonymized view.

For those reasons, it is imperative that you can create a link between operational view and anonymized view. This link can be done by using the hash technique on the primary key of the operational table.

In our example, we assumed that our primary key was the customer number. Now let us imagine that Paul MARTIN, Customer_nb=123456 has moved from Courbevoie to Paris. How can I report this update towards my anonymized table/dataset.

Since the hashing technique is deterministic, it means that if I apply the hash technique to the customer_nb the result will always be 4458FTRX if the source is 123456.

I then have a direct link in order to update my anonymized line by applying the previous anonymization methods.

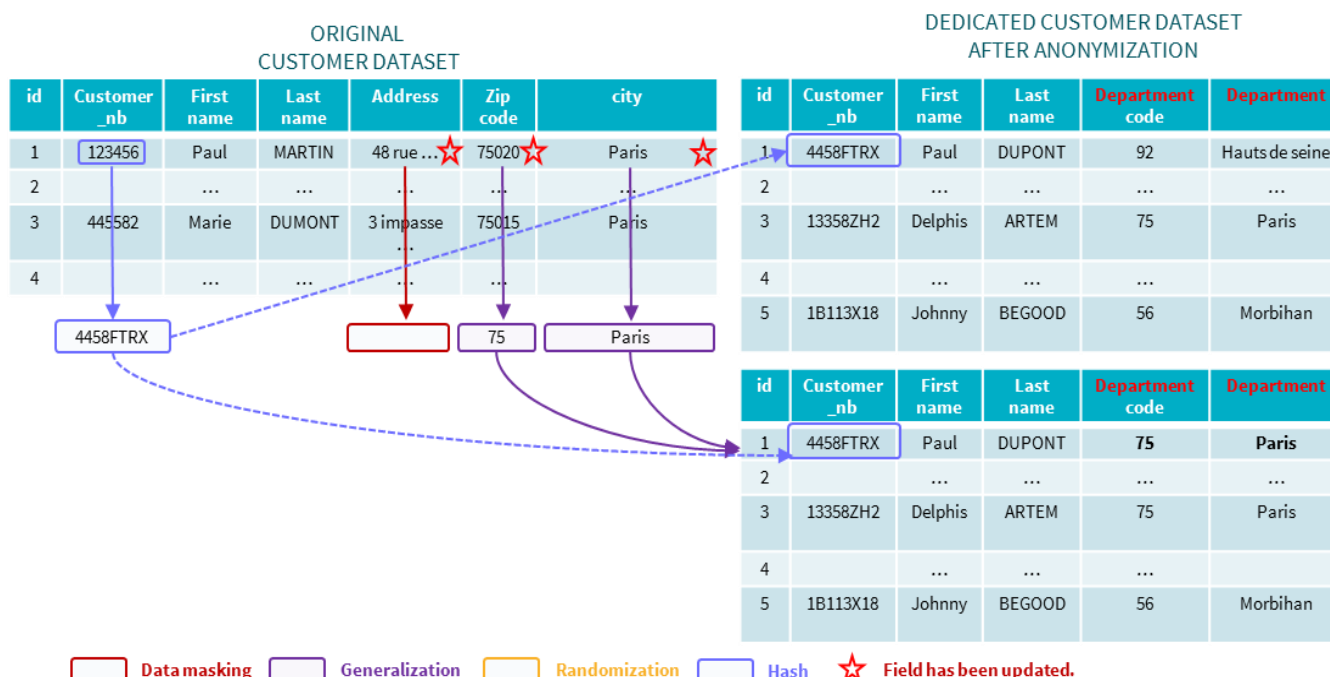


Figure 16 Replicating an update to the anonymized view

In the previous sample, for explaining purpose, we showed a flow of update in the anonymized view that was asynchronous to the update in the operational table or rather triggered by an update in the operational table.

The updating flows for both tables can be parallel since the primary is the same in both table. The second flow should just apply anonymization techniques before updating/inserting in the anonymized table.

Then you might tell me, if I can link both tables by the primary key, I can do it also in my use cases. The answer is yes. However, you will not be able to retrieve the original data concerning the deleted entry. Johnny BEGOOD will always stay in the Morbihan because if the customer was forgotten from our operational database he will never update his address. By the way do you remember the original name of Johnny BEGOOD.

Impacts on :	How
Data flows	New update/insert data flow needing to take into account data anonymization techniques

B. Pros and Cons of this technique

PROS	CONS
Dissociation between statistical use cases and operational use cases. Differentiation possible on access rights	Duplication of data, however duplicated data are anonymous so no need to perform deletes
No impacts on aggregates if done on anonymized views	Need to deal with legacy data/ Entry flows should be added ¹³

¹³ All the existing flows need to be reviewed in order to deal with anonymized data. And sometimes the number of flows are very high

NO IMPACT on existing Use Cases	Maintenance in case of new fields or updated fields is much higher because of data duplication
Assessing the effectiveness of the anonymization can be performed on your legacy (meaning a lot of data)	

Section

6

Main Takeaways

Anonymization (as well as Pseudonymization) are **methods** that use many techniques on data to avoid identification of a subject. Encryption and deletion are only **techniques** used to achieve this objective.

When you talk about Anonymization, make sure your DPO is ok with the level of anonymity you want to achieve.

As a rough definition, Identifier must be hidden (basic data masking) whilst Quasi-Identifiers should be transformed to be less accurate (generalization technique)

You can either keep a set of anonymized data and a set of non-anonymized data, or mix anonymized data with non-anonymized data.

You should envisage doing regular checks on your anonymized data to verify your level of anonymity (same thing as a business quality rule)

Anonymization can be hard to apply and have a high cost, so make sure you have a business case pertaining to it. Moreover, it is not easy for the business to understand why anonymization is needed. Usually they would rather extend the retention period than apply anonymization on their data.

As for the tool to use, the technics shown are usually available on the market except for feature creation.

So, as a conclusion, understand your data and how it contributes to your use cases before trying to anonymize it!