Computers, Materials & Continua
DOI: 10.32604/cmc.2023.031491
Article

Data De-identification Framework

Junhyoung Oh[1] and Kyungho Lee[2,*]

[1]Center for Information Security Technologies, Korea University, Seoul, 02841, Korea
[2]School of Cybersecurity, Korea University, Seoul, 02841, Korea
*Corresponding Author: Kyungho Lee. Email: kevinlee@korea.ac.kr
Received: 19 April 2022; Accepted: 12 July 2022

Abstract: As technology develops, the amount of information being used has increased a lot. Every company learns big data to provide customized services with its customers. Accordingly, collecting and analyzing data of the data subject has become one of the core competencies of the companies. However, when collecting and using it, the authority of the data subject may be violated. The data often identifies its subject by itself, and even if it is not a personal information that infringes on an individual's authority, the moment it is connected, it becomes important and sensitive personal information that we have never thought of. Therefore, recent privacy regulations such as GDPR (General Data Protection Regulation) are changing to guarantee more rights of the data subjects. To use data effectively without infringing on the rights of the data subject, the concept of de-identification has been created. Researchers and companies can make personal information less identifiable through appropriate de-identification/pseudonymization and use the data for the purpose of statistical research. De-identification/pseudonymization techniques have been studied a lot, but it is difficult for companies and researchers to know how to de-identify/pseudonymize data. It is difficult to clearly understand how and to what extent each organization should take de-identification measures. Currently, each organization does not systematically analyze and conduct the situation but only takes minimal action while looking at the guidelines distributed by each country. We solved this problem from the perspective of risk management. Several steps are required to secure the dataset starting from pre-processing to releasing the dataset. We can analyze the dataset, analyze the risk, evaluate the risk, and treat the risk appropriately. The outcomes of each step can then be used to take appropriate action on the dataset to eliminate or reduce its risk. Then, we can release the dataset under its own purpose. These series of processes were reconstructed to fit the current situation by analyzing various standards such as ISO/IEC (International Organization for Standardization/International Electrotechnical Commission) 20889, NIST IR (National Institute of Standards and Technology Interagency Reports) 8053, NIST SP (National Institute of Standards and Technology Special Publications) 800-188, and ITU-T (International Telecommunications Union-Telecommunication) X.1148. We

propose an integrated framework based on situational awareness model and risk management model. We found that this framework can be specialized for multiple domains, and it is useful because it is based on a variety of case and utility-based ROI calculations.

Keywords: Privacy; de-identification; anonymization; pseudonymization; information security

1 Introduction

As intelligent information technology develops and data technology advances, the amount of personal information collected for the use of various services has become vast, and the information processing procedure is becoming more complex. As the technology level is advanced, the amount of information used has increased, and the connection of information has also increased, and it has been developed into the concept of superintelligence [1]. Superintelligence uses the learning of big data to provide customized services to consumers. To this end, it is necessary to collect and analyze data generated from various devices. Therefore, a hyper-connected intelligence information society that creates new values by analyzing vast amounts of data has arrived [2].

While using vast amounts of data, data analysis technology is also being applied to the industry in real time. Each service can use machine learning technology to self-learn from big data and derive meaningful information from accumulated knowledge. It learns from analysis and provides useful judgments and predictions to users based on what it has learned. Eventually, the data analysis process has become a key element in the data economy. In the past, data and information which creates socio-economic value by reconstructing the data were distinguished, but in the era of big data, the information is often the data itself. In other words, the data is already in a state of the information that has socio-economic value, and information and data are often mixed because it is common to recombine the information to create new information.

Data may be identifiable on its own, and even if it is not a personal information that infringes on the rights of an individual, it becomes important, sensitive, or completely unthinkable personal information as soon as it is connected to each other. In a society with highly complex and connected systems, accidents and dangers become commonplace, and avoiding them becomes virtually impossible [3]. With the advent of the big data society and the concentration of data that stays in each region in the center, not only the power of data but also the size of the risk and the probability of the occurrence of the risk in case of personal information leakage has increased [4–6]. In general, in the relationship of personal information transactions where consumers provide personal information to data platform companies and companies provide related services to consumers, consumers and companies trade risks related to the personal information infringement through mutual contracts. At this time, companies collect and combine information in order to earn the profits of using personal information, and consumers accept risks and trade their information to get benefits even if they take a certain amount of risk. However, the use of personal information and the protection of the data subject's rights are contradictory.

Organizations have many restrictions on the use of personal information while protecting personal information. Government information can be shared to increase transparency, provide new resources to private enterprises, and become an overall effective government. Private companies can benefit from shared information in the form of openness and increased citizen participation, and even benefit from the sale of personal information or the results of analysis. If personal information contains identifiable

information such as name, e-mail address, geographical location information, or photo, there may be a conflict between the use of personal information and the goal of protecting privacy. Through de-identification, it is possible to remove sensitive personal information related to privacy that identifies an individual while leaving other useful information. Therefore, de-identification can be a solution that makes data useful while protecting personal information. If we perform an appropriate level of de-identification when collecting data and processing the data on the server, we can prevent re-identification when the data becomes public.

Proper de-identification can benefit both the data subject and the company. A properly de-identified dataset not only guarantees the privacy of the data subject, but also has sufficient utility to provide the services to the data subject. However, conducting a de-identification at an appropriate level is a very difficult challenge. Most companies do not have an in-depth understanding of de-identification, so they simply follow the guidelines suggested by each country and only cares about not to create legal issues. In addition, since various standards suggest non-identification models and frameworks that fit each point of view, it is difficult to perform a series of processes from the overall point of view unless the relevant data is utilized. Therefore, we try to solve this problem by analyzing the data and presenting an integrated framework. We organized various standards such as NIST IR 8053, NIST SP 800-188, ISO/IEC 20889, and analyzed regulations such as GDPR and HIPAA (Health Insurance Portability and Accountability Act) to create an overall framework. This framework has a series of flows for analyzing, evaluating, and solving privacy risks for de-identification based on risk management. To this end, the situational awareness model was utilized to make the quality of the framework better. Also, the usefulness was improved by mapping each part of the framework with the steps in the data life cycle and the stakeholders involved. In addition, this framework presents various elements needed when conducting risk assessment for personal information protection and de-identification. Each element has been established through extensive research in the past, but there are still areas that have not been properly studied. Through this framework, if researchers study the relevant parts properly, knowledge about de-identification technology will be advanced further.

2 Background

Personal information is information about an individual, which means information that can identify an individual. There are slight differences between countries as to the extent to which personal information is recognized. Moreover, due to the rapidly changing internal and external environment today, the definition method and scope of personal information is gradually changing. And personal information is increasingly subdivided as time goes by. The most representative example is PII(Personally Identifiable Information), which means identifiable information. Although there are various definitions of the term Personally Identifiable Information in the guidelines of various laws and institutions, this term usually refers to information that contains a specific identifier for an individual.

Various terms derived from personal information are used in various literatures. A representative example of the term is ‘Personally Identifiable Information’. This term is used to indicate personal information that identifies people. However, even though it cannot directly identify a person, there are some pieces of information that can identify people when several pieces of information are gathered. That’s why we called the information ‘Potentially Personally Identifiable Information’. These terms differ from country to country. In some places, ‘Personally Identifiable Information’ is called a direct identifier, and ‘Potentially Personally Identifiable Information’ is called an indirect identifier. In addition, this information is classified in more detail, and sometimes the terms are mixed with various similar terms. And it is difficult to clearly distinguish the information that directly affects identification

because it varies depending on the situation. For these reasons, it is difficult to accurately understand the terms unless we are an expert in this field.

2.1 Concept of De-identification/Anonymization/Pseudonymization

Terms related to de-identification as well as terms for personal information are ambiguous. In the past, the term de-identification was used, but over time, the terms anonymization and pseudonymization have been used a lot. Of course, the terminology slightly varies from country to country. Many people do not know the clear difference between de-identification, anonymization, and pseudonymization. Furthermore, the relationship between the de-identification technique and pseudonymization technique, or the de-identified dataset and pseudonymized data is not clear. Judging from the term itself, it seems that the de-identification technique should be used to create a de-identified dataset, and the pseudonymization technique should be used to create pseudonymized data. However, the de-identification technique should be fully utilized to create pseudonymized data. From the point of view of the process, pseudonymized data is at a specific point in the process of de-identification, and de-identified data becomes de-identified when an appropriate level of de-identification proceeds.

The United States has mainly used the term 'de-identification' through major legislation such as HIPAA. Representatively, data from the NIST (National Institute of Standards and Technology), which is an organization in charge of standardization in the United States, uses the concept of de-identification, which means that the data user can identify the data subject by removing or replacing the identification information. It is defined as 'processing that makes it difficult to do' [7]. It is rare that both de-identification and anonymization terms appear in the US (United States) government or public institution data. It is not clear whether there is a specific reason for the US to choose de-identification rather than anonymization as the term. In the US Department of Education material that introduces the terminology, both de-identification and anonymization terms appear. De-identification is the process of removing or obscuring any personally identifiable information [8]. Anonymization is distinguished by the process of data de-identification which produces de-identified data, so it is unclear what practical differences exist.

NIST IR 8053 [9] describes de-identification in two expressions. The first one is 'a tool that organizations can use to remove personal information from the information they collect, use, record and share with other organizations', and the second one is 'de-identification is not a single technique, and various types of personal information'. It is a collection of approaches, algorithms, and tools that can be applied to various effects. In addition, in NIST IR 8053, pseudonymization is expressed as a kind of specific transformation that replaces a name and other information that directly identifies an individual with a pseudonym. If all of the direct identifiers are systematically aliased, information belonging to an individual can be linked across a plurality of personal information records or information systems through the alias processing. If the organization that performed the pseudonymization process has a table that links the original identity with the pseudonym, or if the variable is revealed using an algorithm that can find it, the pseudonymization process can be easily reversed. In many cases, pseudonymization can be used to re-identify data subjects by reverting the pseudonym at a point in the future. If the mapping between the direct identifier and the pseudonym is preserved or can be reproduced, the pseudonymized dataset can be returned. For example, an identifier may be encrypted using a secret key to generate a pseudonym. When the key is decrypted, the pseudonymization process is reversed, and the original identifier is generated. If re-identification is not prohibited in the agreement for data usage and if a pseudonymized set of information is disclosed, the recipient can reverse the pseudonym and attempt re-identification using the identifying information.

The Privacy Working Group (the European Commission's working group) said, 'Personal data that has been pseudonymized cannot be identified with information that has been anonymized. This is because the pseudonymized personal information identifies individual data subjects and allows them to be linked across various information sets.'

In ISO/IEC 20889 [10], de-identification is expressed as 'process of removing the association between a set of identifying attributes and the data principal'. Data Principal means the natural person to whom the personal data relates to. And pseudonymization is expressed as 'de-identification technique that replaces an identifier (or identifiers) for a data principal with a pseudonym in order to hide the identity of that data principal'.

2.2 Stakeholder

The GDPR is a strong regulation on privacy and provides detailed explanations for each role of a stake holder [11]. Stakeholders related to privacy and de-identification are well represented in the GDPR. The data controller defines how personal data is processed (handled) and what purpose it has [12]. It is also responsible for ensuring that external contractors comply with the regulations. Data Processor means an internal group that keeps and processes records of personal data, or an outsourcing company acting on behalf of some or all of these activities. GDPR holds data processors accountable for any data breach or non-compliance. Even if it is entirely the fault of an external data processor, both the processor (data processor) and the enterprise, such as cloud providers, may face a fine. The GDPR requires controllers and DPOs (Data Protection Officers) to designate a DPO to oversee their data security strategy and GDPR compliance [13]. Data Protection Officers are responsible for processing of the personal data of the organization. Companies should appoint one DPO if they handle or store a lot of data about EU (European Union) citizens, if they are processing or storing special personal data, if they regularly monitor data objects, or if they are public authorities. Some public authorities, such as law enforcement agencies, may have exceptions to the DPO appointment requirements.

2.3 Data Life Cycle

Data life cycle refers to all stages from data creation to destruction [14]. Data can be managed safely and efficiently through appropriate actions at each stage. Data de-identification technique can be applied to data only at a specific stage, but various analyzes to perform risk management must be performed appropriately for each stage of the data life cycle [15]. Typically, an organization conducts de-identification for the purpose of protecting privacy and security. This clause defines a data life cycle and describes when to consider a de-identification process based on this data life cycle model. The data life cycle concept is used to select appropriate controls based on analysis of the possibility of re-identification [15]. The data life cycle model of ITU-T X.1148 [16] consists of data collection, data destruction, data management, data use, and data release. In the data collection phase, data is collected from the data subject. The dataset produced as a result of this data collection may include PII (Personally Identifiable Information). PII is any data that could potentially be used to identify a particular person. De-identification creates a new dataset from which all PII has been removed. It is recommended that de-identified datasets are internally used by an organization instead of the original dataset possible in anywhere. In the data management phase, to avoid archiving identifier, de-identification should be applied after data transformation and before data retention. Organizations are recommended to consider the potential for re-identification and set clear access controls, maximum retention limits, and data removal policies that maximally reduce the possibility for connection between de-identified data. Organizations are recommended to consider anonymization techniques such as data aggregation wherever the intended purpose of use allows. Data destruction can be made

at any phase, i.e., data collection, data management, data use, and data release. Data should be destroyed with verified measures to avoid recovery of data. Especially, destruction of data should be considered upon detection of possibility of re-identification. In data use phase, if PII is needed within an organization for data management, the data is recommended to be de-identified prior to being released as a dataset for data sharing. Data can be shared with third parties who are bound by additional administrative controls such as 'data-sharing' agreements. De-identified datasets may also be subject to data release. De-identified data release is classified into three models: public, semi-public or non-public. The amount of de-identification required may vary depending on the release model selected.

2.4 Risk Management

Risk Management refers to a series of activities aimed at maintaining an acceptable level of risk to an organization's assets [17]. To do so, it analyzes the risks of the assets, analyzes the cost-effectiveness to protect them from these risks, and devises protection measures. The overall process of risk management consists of establishing a strategy and plan, analyzing the elements constituting the risk, evaluating the risk based on the analysis, selecting information protection measures, and establishing a plan to implement it. Risk management may be divided into risk assessment [18] and risk mitigation [19] and may be subdivided into risk analysis and risk evaluation. In general, risk management refers to a series of processes that identify risks, analyze risks, determine the size of the risks, and identify countermeasures. Initial risk analysis was applied to physical threats such as fires and accidents to calculate the potential loss from a threat occurred. By applying this in the insurance industry, the concept of risk was established, and as it is applied to the information processing field, the demand for risk management has increased. In risk analysis, it is important not only to minimize the risk by simply applying safeguards, but also to effectively prevent the risk by measuring the value and loss of the asset. If safeguard is applied without systematic risk analysis, the priorities for vulnerabilities and threats cannot be set and unnecessary or excessive investment may occur. Risk refers to the possibility of loss due to the occurrence of unwanted events. In general, assets are set up to calculate risk. In the de-identification framework of this study, an asset can be a collection of personal information. And a threat is a source or agent of unwanted events that can cause loss of assets. Threats are defined in different contexts, and this study is based on scenarios. Vulnerability is defined as a potential property of an asset that is the target of a threat. In this study, various models such as K-anonymity, L-diversity, and T-closeness are used to quantitatively express the corresponding vulnerability.

3 Data De-identification Framework

This framework has created a flow based on the situational awareness model. In terms of content, this framework is based on a risk management framework. Therefore, we propose a series of steps to lower the risk until the data is secured and released. First, it begins with classifying identifiers as part of dataset analysis. Each identifier has different importance in terms of privacy. Some identifiers are used to directly identify a person, some are used indirectly, and the risk may increase if several identifiers of low importance interact each other. Considering these various situations, systematic classification of identifiers is essential. This classification is based on the content of ISO/IEC 20889. And classification for sensitive attribute is also necessary. As each law defines sensitive attribute, it varies according to cultural and historical viewpoints. In addition, since privacy preferences for sensitive attributes are different for every person, it is required to make a comprehensive judgment considering the context. After analyzing the dataset, it is necessary to analyze the risk. Various methods can be used for risk

analysis, but the basic methods presented in this framework are case study analysis and scenario-based analysis. If we analyze various re-identification cases and scenarios that have occurred so far, we can predict future attacks and easily find out what is lacking in the current state based on the case. In addition, attacks can be systematically organized by various attack types suggested by ISO/IEC 20889 and the models that quantitatively evaluate the attacks as threats and vulnerabilities. After analyzing the risk, it is supposed to assess the risk. Risk assessment can basically calculate ROI (Return On Investment) based on utility and cost, and the cost can be calculated based on fines imposed when legal issues arise. In addition, a large institution can check the overall process through self-adequacy evaluation, and evaluate it with the help of a professional institution. Then, various actions can be taken to mitigate or eliminate the risk. Considering the characteristics of each action, it is possible to appropriately select the action that best suits the situation. And through the feedback process, we can go back to the previous step and repeat it. A decision can be made to release the dataset when the risk is sufficiently mitigated and removed. An overview of this framework is depicted in Fig. 1.

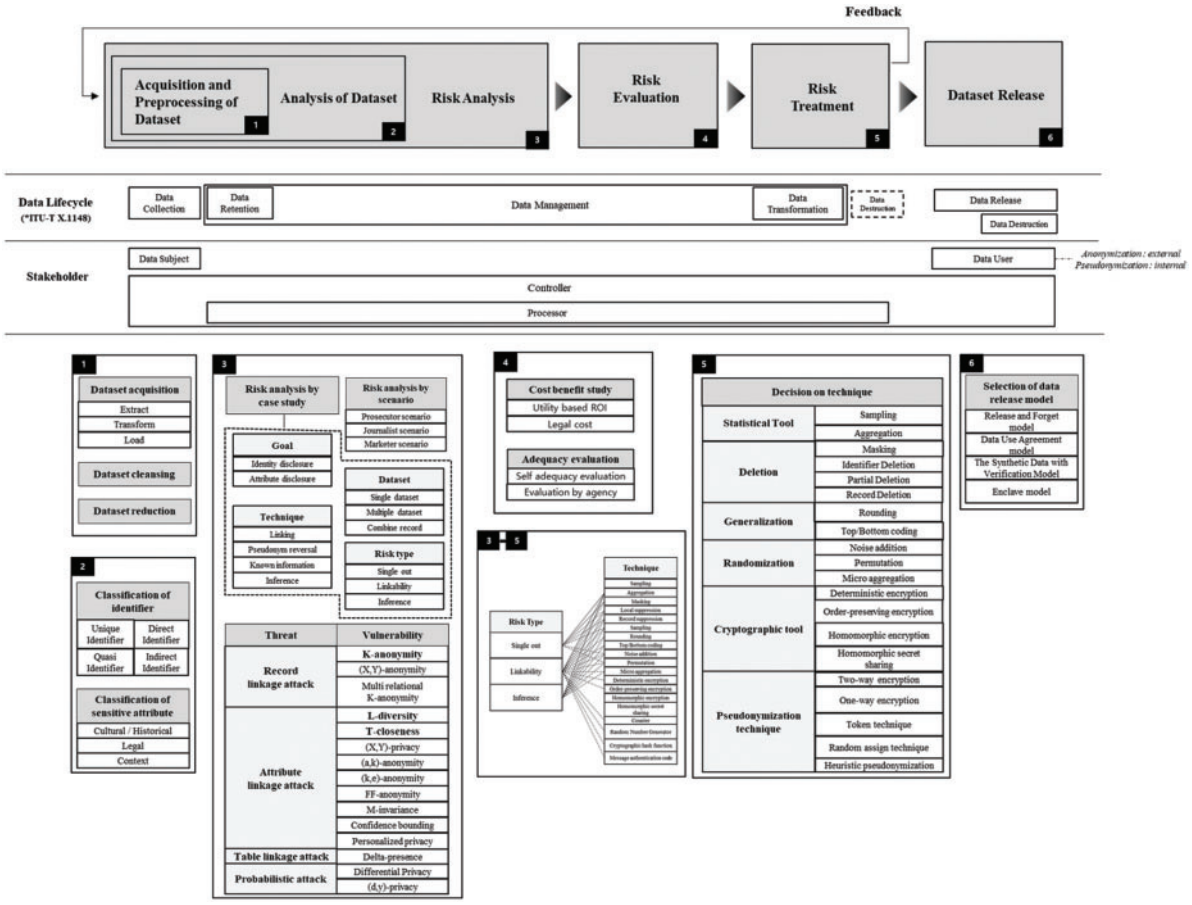


Figure 1: Data de-identification framework

Each element of the life cycle presented in ITU-T X.1148 is mapped to the current flow. Data collection and data retention are included in Acquisition and Preprocessing of Dataset, and a series of processes from dataset analysis to risk treatment are included in data management. In addition, data transformation is a risk treatment, and some data destruction is also included in it. Most of the

contents of Data Release and Data Destruction correspond to Dataset Release. From the stakeholder's point of view, this includes Data subject, Data controller, Data Processor, and Data User. Data Subject is involved when dataset is released and the data user can use it. In addition, the series of processes for managing risk can be seen as the Data Processor performing the entrusted work under the responsibility of the Data Controller.

3.1 Acquisition and Preprocessing of Dataset

Acquisition and Preprocessing of Dataset part is depicted in Fig. 2. ETL (Extraction, Transformation, Loading) is a framework related to data movement and transformation procedures. The main task is to extract and convert data from various data sources and load it into ODS (Operational Data Store), DW (Data Warehouse), and DM (Data Mart). Extraction is the operation of acquiring data from one or more data sources. Transformation is data cleansing, format transformation and standardization, integration, or the application of business rules built into multiple applications. Load is an operation that loads the data that has been processed in the transformation phase into a specific target system.

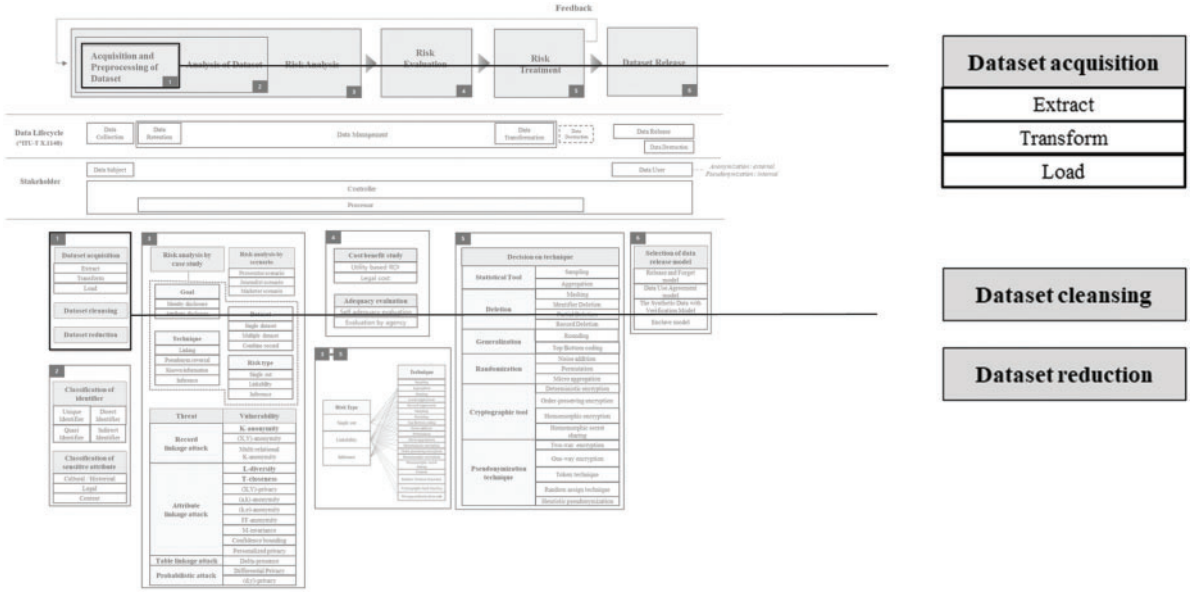


Figure 2: Acquisition and preprocessing of dataset

A certain degree of processing is required to analyze the collected and extracted data. There may be missing data, an incorrect value, or the unit of data in the middle. Besides, we need to figure out whether the data is in a good format for processing and whether it needs to be processed again or not. For example, in the case of categorical data in letters, the letters are changed to numbers to facilitate the processing. Or it normalizes the distribution of numerical data. Most of the data has a different purpose when it is first obtained. If we want to use this data for other purposes later, we have to process the data. There are various types of data preprocessing, including data filtering to select the necessary data and transformation to change the format of the data. And when we need to combine data from multiple sources, we conduct data integration. Data cleansing is the process of filtering out errors and cleansing the data to make it easier to analyze.

There are principles and techniques for data reduction. The principle of data reduction is to acquire information only to the extent of cancellation when acquiring information from the data

subject. There are eight principles in Privacy by design [20], one of which is minimization. According to this minimization principle, only a non-excessive amount of data should be collected. This principle is important because they greatly affect the trust of the data subject.

The data reduction technique means reducing the size of data while having the same amount of information. PCA (Principal Components Analysis) is a representative example of extracting new values with characteristics of existing data as a method of data reduction. PCA is a technique that preserves the variance of data as much as possible, finds a new basis (axis) that is orthogonal to each other, and transforms samples from a high-dimensional space into a low-dimensional space with non-linear association [21]. As a method of data reduction, it is possible to create new variables that can represent multiple data.

3.2 Analysis of Dataset

Analysis of Dataset part is depicted in Fig. 3.

3.2.1 Identifier

Identifiers presented in ISO/IEC 20889 [10] are largely classified in terms of operating environments and datasets. Classification by the operating environment is divided into Direct Identifier and Indirect Identifier. The operating environment includes de-identification, along with information owned by third parties or potential attackers residing in the public domain. It contains information owned by the entity processing the data. Direct Identifier is information that can be directly identified in a specific operating environment. Indirect Identifier is a property that enables unique identification of a data subject in a specific operating environment along with properties within or out of the dataset.

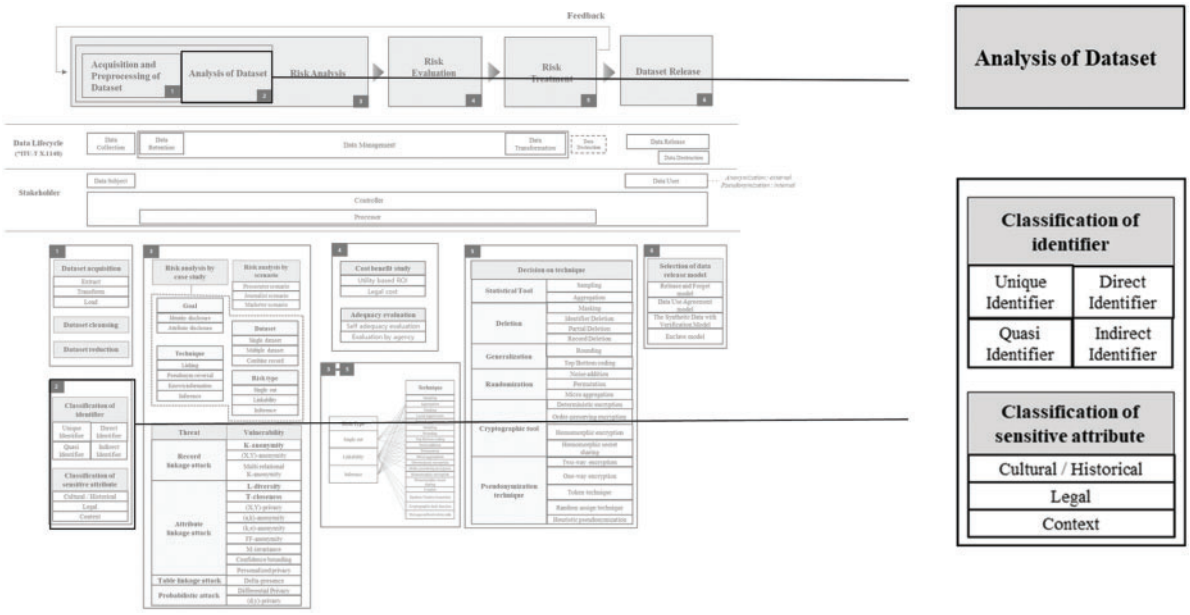


Figure 3: Analysis of dataset

Classification from the perspective of the dataset is divided into Unique Identifier and Quasi Identifier. Unique Identifier is an attribute that independently selects a data subject from a dataset.

Quasi Identifier is an attribute that selects a data subject considering other attributes included in the data set together. It is described in Tab. 1.

Table 1: Classification of identifier between ISO/IEC 20889 and NIST IR 8053

ISO/IEC 20889		NIST IR 8053
Dataset context	Unique identifier	Direct identifier
	Quasi-identifier	Indirect identifier
Operational context	Direct identifier	Direct identifier
	Indirect identifier	Indirect identifier

3.2.2 Sensitive Attribute

Sensitive attribute does not significantly affect the identification of an individual, but when an individual is identified due to direct and indirect identifiers, it refers to information that may infringe on an individual, and most of it is under the purpose of data analysis. Sensitive information often refers to legal sensitive attributes such as political views and sexual life. In addition, it contains sensitive information about stigmatism that can cause economic discrimination or shame and includes information that can be personally identifiable by itself because of its high specificity such as rare diseases. It also contains various information that can be sensitively accepted depending on the situation such as income and credit rating. Sensitive attribute has many contents defined in various laws such as GDPR.

Definition of sensitive data is data consisting of racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic data, biometric data, data concerning health, and data concerning a natural person's sex life or sexual orientation. And sensitive attributes vary greatly depending on cultural and historical circumstances. The protection of privacy regarding the disclosure of private life in the United States can be summarized as follows. There are no federal laws for protection of privacy in the United States. The collection and disclosure of personal information by private entities has not been dealt with in general, but has been dealt with by individual sectors. The areas where federal laws and regulations protect personal information are consumer credit, telecommunications, federal government agency records, educational records, bank records, customer financial information, etc. [22]. This legislation in the United States does not protect personal information in principle but only implies that it protects personal information in specially regulated areas, so most of the personal information is not included in these individual areas. It means that personal information is not protected by law. While the US legislation is in the balance of commercial interests and privacy interests in individual sectors, the European Union follows a focused approach and general law on consumer and privacy. Soon, Europe regards privacy as a basic right, and although basic rights are not superior to other rights, privacy is considered superior to economic rights.

People have different privacy preferences according to various environments. Therefore, rather than defining the sensitive attribute collectively in the law, it is appropriate to set a sensitive attribute under the context. The followings are studies on privacy preferences that vary among the environments. Naeini et al. [23] constituted a number of scenarios that can occur, such as the data type, the location where the data is collected, the data collection device, the purpose of the data collection, and the retention period after data collection. Among the scenarios, 380 feasible scenarios were selected and

divided into 39 vignettes to investigate the user's privacy concerns. As a result, it identified with which scenarios users felt comfortable or uncomfortable, and in which cases they felt more sensitive to data sharing. Zheng et al. [24] investigated privacy concerns related to various IoT (Internet of Things) devices used in smart homes. As a result, it was possible to identify the difference between users' privacy concerns and privacy preferences affected by the brand of IoT device manufacturers and perceived benefit by data sharing. Kim et al. [25] studied perceived privacy risk in various IoT environments such as smart medical care, smart home, and smart transportation. Users provided personal information for better personalized service in many environments, but in the smart healthcare field, they thought that the information was sensitive and tried to provide less personal information.

3.3 Risk Analysis

Risk Analysis part is depicted in Fig. 4.

3.3.1 Risk Analysis by Case Study

In order to systematically analyze the risk, it is necessary to analyze the events that have occurred before. In other words, it is possible to clearly understand the case study to prevent future events and interpret the current risk. There are various cases related to re-identification, and a framework or model that systematically organizes the cases has not been established yet. Several cases were analyzed and organized. And, assuming that these cases are used to prepare for future attacks, we have made a framework which is the easiest to apply with the advice of various experts. Case analysis consists of Goal, Technique, Dataset, and Risk type. The goals of re-identification attacks are largely divided into two.

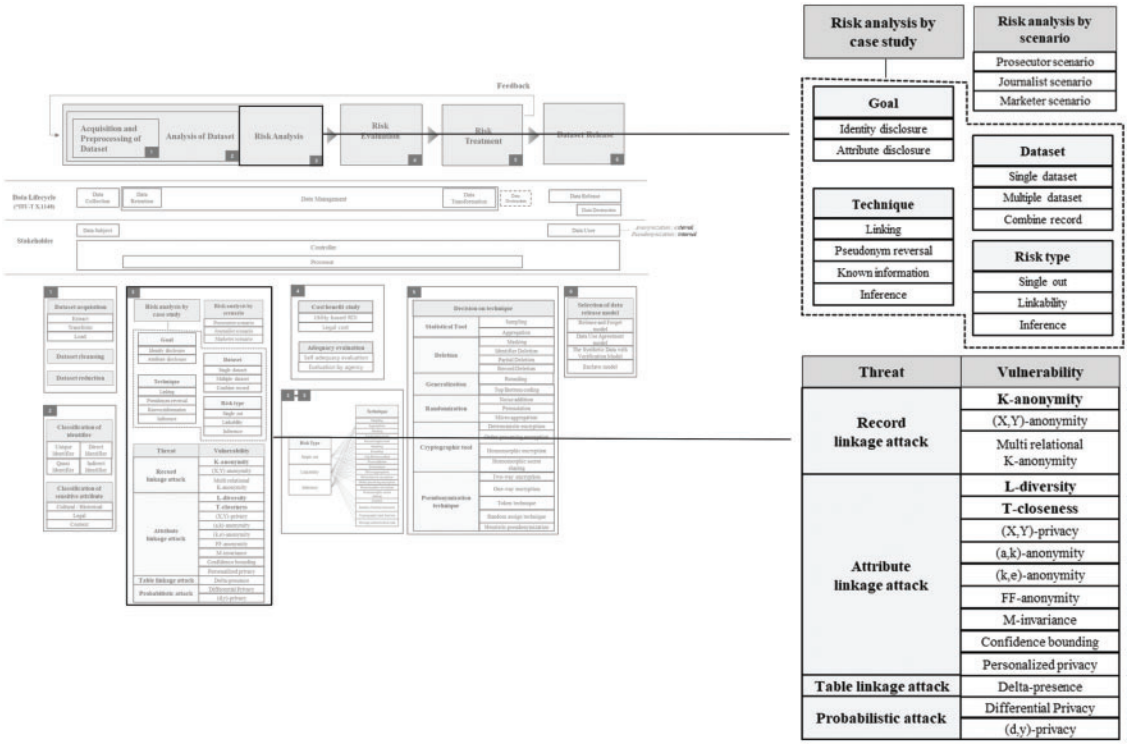


Figure 4: Risk analysis

- *Identity disclosure: In case of revealing the identity of a specific person in the dataset.*
- *Attribute disclosure: When specific information of a specific person (generally celebrity) is found in the dataset.*

Re-identification attacks basically use datasets. It can be classified as follows depending on whether it is one dataset, multiple datasets, or combining specific information into one dataset.

- *Single dataset: When attacking with only the public dataset.*
- *Multiple dataset: When attacking using multiple public datasets or datasets collected through a different path rather than the public dataset.*
- *Combined record: When one dataset is used and the information on a specific target from SNS (Social Network Service) or neighbors is combined to attack it.*

Re-identification attacks generally have the four methods as follows.

- *Linking: The most common way, an attack method that links the same record, etc.*
- *Pseudonym reversal: A method of finding the original data by inferring the key value for pseudonymized data.*
- *Known information: A method of verifying identity or finding specific attributes by applying already known information to the dataset.*
- *Inference: A method of probabilistically inferring who a specific individual is through the properties and values of specific information.*

This risk type is based on the privacy risk type of ISO/IEC 20889.

- *Single out: The degree to which records belonging to the data subject can be identified in the data set by observing a set of known characteristics to uniquely identify a given data subject.*
- *Linkability: Even if one piece of information is individualized so that a specific individual is not known, the degree to which it is possible to identify the information of a specific individual by linking the same value with other information.*
- *Inference: The degree to which a specific individual can be inferred through the properties and values of specific information (uncommon values, etc.) even if a specific individual cannot be distinguished by single out.*

The Governor Weld case was identified by Sweeny's study [26]. According to a 2000 study by L. Sweeney, processing the 1990 census data showed that 87.1% of people in the United States could be identified by their zip code, date of birth, and gender information. In addition, 53% of the US residents can be identified by their city, date of birth, and gender. In the state of Massachusetts, a summary of hospital access records of state government officials was disclosed for research purpose. Since this dataset has become a basic de-identification operation, identifiable information (names, addresses, etc.) has been removed. However, the quasi-identifiers in this dataset were not removed. Since the identifiers of this dataset were deleted, re-identification was almost impossible with this dataset alone. However, in the state of Massachusetts, voters' lists were publicly sold, including names, gender, date of birth, and zip code. Sweeny, a graduate student at the time, used information already known about Governor Weld, such as the fact that Governor Weld resided in Cambridge, Massachusetts with a population of 54,000, and the zip code of the governor's residence because the Governor is a public figure. Attempts to re-identify the data were successful. This is because only six Cambridge residents in the data obtained had the same birthday as him, only three of them were male, and only the governor lived in that zip code. In the Governor Weld case, when re-identification was performed using a dataset,

the identity of governor weld was revealed, and it was an attack using multiple datasets. And in this case, it was re-identified using the linking technique, and there was a risk of single out and linkability.

The American Online case arose in 2006 when AOL (America Online) announced a new initiative called "AOL Research" [27]. For academic research, America Online posted 20 million search queries searched by AOL's search engine over a three-month period by 650,000 users. This dataset has been de-identified, such as pseudonymizing user names and IP (Internet Protocol) addresses, and suppression of explicit identification information such as IP addresses. However, since no techniques such as noise addition were included in the pseudonymization, all queries made by the same person had the same pseudonymized value. Re-identification attacks took place using this fact. New York Times reporters Michael Barbarno and Tom Zeller asked about 4,417,749 users who made a search query such as 'a landscaper in Lilburn, Georgia', 'a few people whose last name is Arnold', and 'a house sold in the shadow lake district of Gwinnet county, Georgia'. They could determine the identity of the user by obtaining clues from it. They were able to identify this user as the widow of Thelma Arnold, 62, from Linburn, Georgia. In the American Online case, when re-identification was performed, the identity of the widow named Thelma Arnold was revealed, and various user attributes were also revealed. In this case, only one dataset published by AOL Research was used, and risk of single out and linkability occurred by performing re-identification using the linking technique.

The Netflix case occurred when Netflix released the target movie, the score, and the date [28]. From December 1999 to December 2005, 500,000 users released 100 million viewing history data that gave ratings to movies. This dataset anonymized records by deleting identifiers such as user names and assigning unique user identification numbers. However, for commercial research purpose, the user's evaluation of the movie, the evaluation date, etc. were disclosed. However, Narayanan and Shmatikov attempted to compare Netflix information with information available through the IMDb (Internet Movie Database), a movie-related website that allows users to rate movies. Unlike Netflix, IMDb publishes movie ratings publicly, much like Amazon publicly publishes user ratings for books. From this dataset, they were able to determine that the two users were also included in Netflix's database with a high statistical certainty. This was because inference was possible by linking similar records between Netflix and IMDb through a connection through movies commonly included in both datasets. By giving ratings to some movies recorded on IMDb, the user unintentionally revealed all the movies he had seen. This was possible because the corresponding IMDb data and the data of the Netflix Prize can be linked and analyzed together. In the end, the 2nd Netflix Prize was canceled after the US Federal Trade Commission pointed out privacy concerns. The Netflix case revealed the identity of a specific person by combining the datasets of IMDb and Netflix Prize using the linking technique and inference technique in this process. Accordingly, in this case, there was a risk of linkability and inference.

The New York taxi case occurred in 2014 when the New York City Taxi Limousine Commission released an information set containing all 2013 New York City taxi ride records (a total of 173 million) [29]. This information did not include the name of the taxi driver or passenger, but contained a 32-digit alphanumeric code that could be easily converted to the taxi's license plate number. An intern at Neustar found out on the Internet that photos of celebrities getting in and out of a taxi with a clear view of the cab license plate can be found on the Internet. The intern used this information to find out two taxi destinations, taxi fares, tips, and more out of 173 million taxi rides. An attacker obtained a photo of a taxi ride of a certain celebrity from social media, and then compared the taxi license plate with the reversal-pseudonymized data to find out the travel route of the celebrity. New York taxi case is an attack to find out the moving route of a specific celebrity. The attacker applied the pseudonym

reversal technique to the public dataset and created a risk of single out by linking the record on the SNS. Re-identification cases are described in Tab. 2.

Table 2: Re-identification cases

Case	Goal	Dataset	Technique	Risk type
Governor weld case	Identity disclosure	Multiple dataset	Linking	Single out/ Linkability
American online case	Identity/Attribute disclosure	Single dataset	Linking	Single out/ Linkability
Netflix case	Identity disclosure	Multiple dataset	Linking/ Inference	Linkability/ Inference
NewYork taxi case	Attribute disclosure	Combined record	Pseudonym reversal	Single out

3.3.2 Risk Analysis by Scenario

NIST IR 8053 [9] and NIST IR 800-188 present various scenarios for re-identification attacks. The scenarios include Prosecutor attack, Journalist attack, Marketer attack, indistinguishability attack, inference attack, etc. Details for each scenario are as follows. Prosecutor attack is an attack model that re-identifies records belonging to a specific data subject with prior knowledge. Journalist attack is an attack model that re-identifies the data subject of a specific record with prior knowledge. Marketer attack is an attack model that re-identifies the data subject with as many records as possible with prior knowledge. (In) distinguishability attack is an attack model that checks the existence of a specific subject in the dataset. Inference attack is an attack model that infers from sensitive information related to other attribute groups. PDPC (Personal Data Protection Commission) Report [30] presents several methods of calculating risk scores based on each scenario. The PDPC report suggests the following process. First, a risk threshold must be established. Reflecting the probability, this value ranges from 0 to 1. This reflects the level of risk the organization is willing to accept. The main factors influencing the value should include the data subject to be re-identified and the possible harm to the organization. However, consideration should also be given to what other controls have been implemented to mitigate other forms of risk other than anonymization. The greater the potential harm, the higher the risk threshold should be. There are no strict rules about the risk threshold we should use. When calculating the actual risk, this guide describes investigating "test risk", which attempts to determine if the other person knows a specific person in the dataset and which records in the dataset refer to that person. A simple rule of thumb for calculating the probability of re-identification for a single record in a dataset is to take the reciprocal of the size of the equality class of the record. To calculate the probability of re-identification of a record in the entire dataset, given that there are once again re-identification attempts, a conservative approach is to equate it with the maximum probability of re-identification of all records in the dataset.

3.3.3 Risk Analysis by Threat/Vulnerability

In this chapter, different types of attacks on datasets are set as threats. In addition, formal privacy measurement models are used to quantitatively grasp the level of contrast for the dataset against the threat. These contents are based on the annex of ISO/IEC 20889, which maps formal privacy

measurement models for each attack. Re-identification risks can be effectively measured using formal privacy measurement models. These models are mathematically verified and can be used appropriately for each case. Most of the models depend on the selection of parameters based on an empirical assessment of the used case, which includes both the re-identification risks of the specific system and the characteristics of the specific dataset. System re-identification risks depend on security and other technical and operational measures that are in place. The characteristics of a specific dataset include the likelihood of an attempted re-identification attack on the dataset (which is driven by an appetite for potential incentives) and the perceived impact on the data principals and/or the organization if the attack is successful.

In a record linkage attack, the attacker matches records based on quasi-identifiers. K-anonymity [31], (X,Y)-anonymity [32], and multi relational K-anonymity model [33] effectively quantify the degree of protection of the dataset against the record linkage model. The most representative model in record linkage attack is K-anonymity.

In an attribute linkage attack, the attacker infers sensitive values from the released data (based on the sensitive values associated with a group that shares the same set of quasi-identifier values). L-diversity [34], T-closeness [35], (X,Y)-privacy [32], (a, k)-anonymity [36], (k, e)-anonymity [37], FF-anonymity [38], M-invariance [39], Confidence bounding [40], and Personalized privacy [41] effectively quantify the degree of protection of the dataset. The most popular models in attribute linkage attack are L-diversity and T-closeness.

In a table linkage attack, an attacker speculates whether there is a record of released data that may disclose sensitive information. The Delta-presence [42] effectively quantifies the protection of the dataset against table linkage attacks.

In a probabilistic attack, the attacker's probabilistic beliefs regarding sensitive information are altered based on the released data. (d, y)-privacy model [43] and different types of differential privacy models [44–52] effectively quantify the degree of protection of the dataset against probabilistic attacks. The differential privacy model is the most representative model of probabilistic attack. And, in annex of ISO 20889, it is well represented which de-identification technique is effective for each risk type. If we refer to this mapping data, we can apply an appropriate de-identification technique for each risk type. Mapping of risk type and de-identification technique is depicted in Fig. 5.

3.4 Risk Evaluation

Risk Evaluation part is depicted in Fig. 6. There are several methods such as cost benefit study and adequacy evaluation to evaluate risk in relation to de-identification. In general, detailed calculations are performed through cost benefit studies, and evaluations are conducted from an overall perspective through adequacy evaluation. The most representative method among cost benefit studies is to calculate ROI. ROI is an abbreviation for Return on Investment, which means the return on investment. In general, in the field of information security, ROSI (Return on Security Investment) is widely used. ROSI is calculated as the gain from the investment *vs.* the cost of the company's investment in security products. ROI calculations are also required for de-identification. The investment cost here means the utility of the dataset lowered by performing de-identification, and the investment performance means the level of de-identification in terms of the risk of re-identification. The level of de-identification can be measured using several techniques and elements from different chapters. The utility of the dataset can be measured by discernibility, granularity, equivalence class size, entropy, etc. The most representative example of these measures is entropy. Usually, the lower the probability, the more uncertain what information it is, and it is expressed as having more information and higher

entropy. Entropy is related to the amount of information. The entropy value can be calculated simply by applying the logarithm to the inclusion ratio of the record value and adding the values multiplied by the weight again.

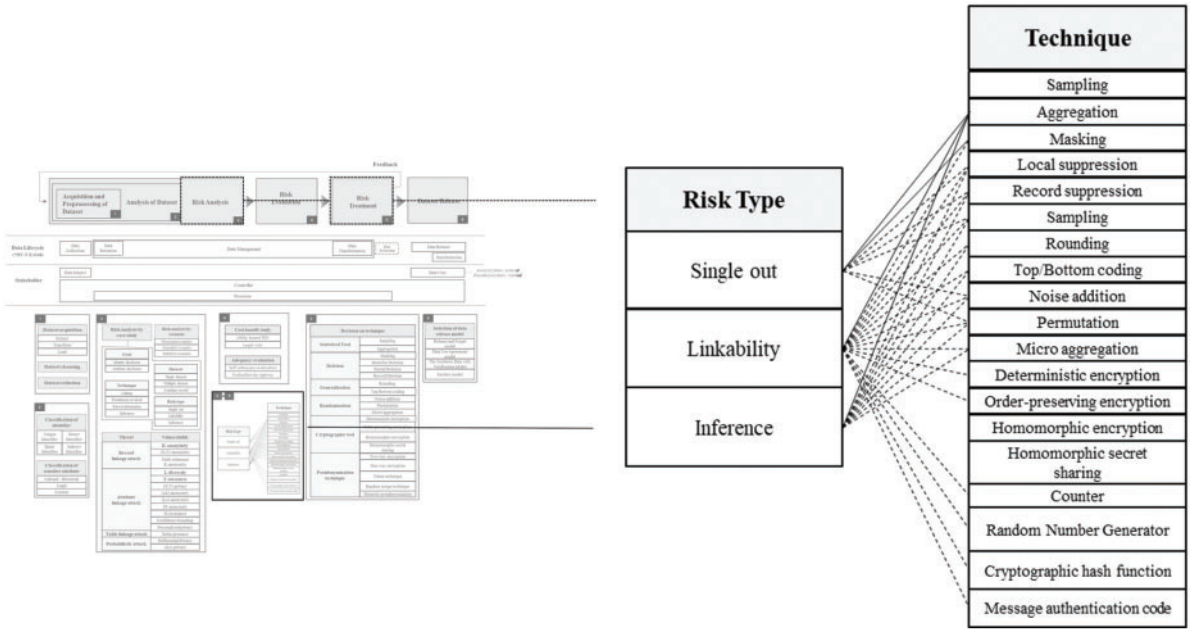


Figure 5: Mapping of risk type and de-identification technique [10]

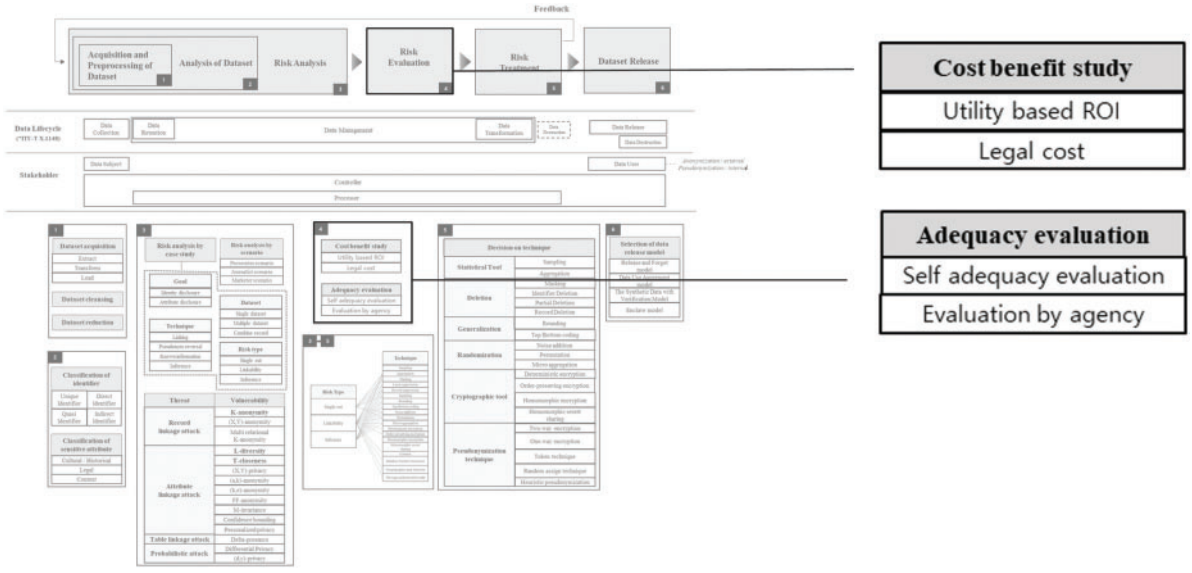


Figure 6: Risk evaluation

In order to determine the level of protection of non-identifiable information, evaluation in various aspects, such as re-identification intention and ability of the relevant user agency, ability to protect anonymous/pseudonymized data, and reliability of business performance, is required. It is necessary to

review the institution's re-identification intention and ability, and raise the level of pseudonymization if the re-identification intention and ability is highly evaluated. It should be reviewed whether users can re-identify information to obtain economic and uneconomic gains, or whether there is room for use of pseudonymized data to the extent that it goes beyond the purpose. It is necessary to review whether data users have specialized knowledge that can attempt re-identification or data that can be linked to pseudonymized data.

Legal cost is actually the most important part of this framework. In most cases, when de-identification is performed in general organizations, it is not performed for the purpose of protecting privacy, but for the purpose not to violate the law. In other words, when an organization performs de-identification, legal cost can have the greatest impact on decision making. In case of serious violation of GDPR, 4% of the company's global annual sales or €20 million, whichever is higher, is charged, and in the case of general violations, 2% or €10 million, whichever is higher. For instance, "Google" was fined by the France National Data Protection Commission in January 2019 for not properly obtaining consent from data subjects [53]. The decision was once again upheld at the French supreme administrative court following the appeal by Google LLC (Limited Liability Company) [54]. Therefore, the proportion of legal costs is very large. The definition of pseudonymization in GDPR is 'the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data is not attributed to an identified or identifiable natural person'. In Basic Pseudonymization, direct identifiers are converted and properly managed, Strong Pseudonymization lowers re-identification by converting a part of indirect identifiers in basic pseudonymization, and in the case of encryption, when the key value is removed to prevent decryption. In addition, various security and personal protection policies are applied. Since pseudonymized data is personally identifiable information, an individual must be notified when using it. Although pseudonymized information and personally identifiable information are basically treated the same, anonymized data is not personal information, so it can escape from these regulations. In order to use the data conveniently and usefully, it is recommended to use anonymized data, but the GDPR does not suggest what degree of de-identification can make anonymized data. Since anonymization can break as time goes by and technology advances, it should always be monitored and properly managed. Therefore, if we use this framework, it will be of great help in maintaining anonymization.

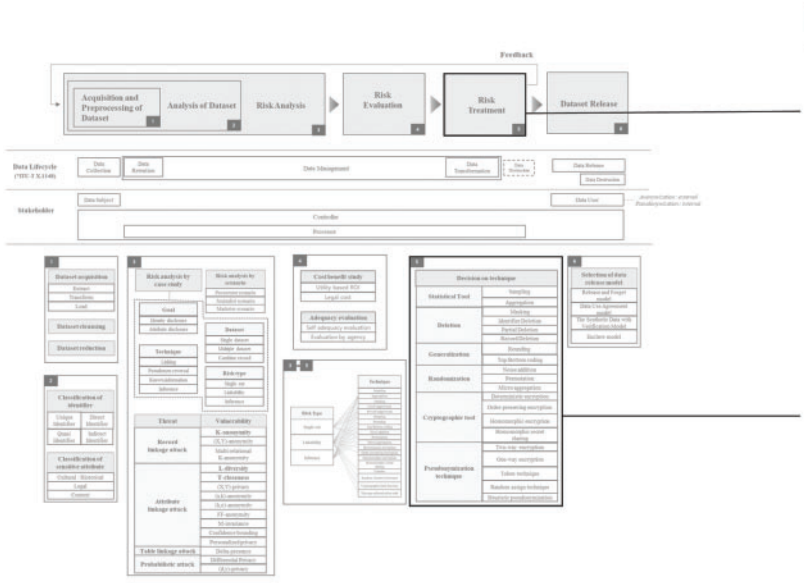
In addition, it is possible to analyze the level of protection and reliability of non-identifying information. It is necessary to review the level of protection of non-identifiable information and reliability of business performance, and increase the level of pseudonymization when the level of protection of non-identifiable information and reliability of business performance is evaluated to be low. The user organization may establish and operate a non-identifiable information management plan to protect non-identifiable information, and examine whether technical, administrative, and physical protection measures have been prepared, and whether or not there is a certification related to personal information protection. Considering these various conditions, it can be evaluated by requesting a specialized agency, and self-evaluation can be conducted.

3.5 Risk Treatment

Risk Treatment part is depicted in Fig. 7. The process of selecting the de-identification techniques needs to be tailored to each specific case of use. Although there is no best or standard way in which the selection process can be done in all cases, the factors are presented in a logical order that can be

used in practice as a part of a data processing system design. Not all the listed steps are relevant in every case.

In a specific data processing system, feasible technical and organizational measures are implemented to balance the need for preservation of data usefulness with the need for data de-identification. They provide an essential context for the choice of the degree to which the data is de-identified. In turn, this degree of de-identification can be achieved by tuning the selection of identifying attributes and the techniques as a part of the selection process. Performing data minimization, i.e., limiting the data to what is directly relevant to and necessary for accomplishing a specific purpose, at the earliest possible stage typically makes the task of data de-identification easier. In some cases, quantifiable guarantees against the risk of re-identification need to be achieved. This can be done by implementing one of the formal privacy measurement models. Suggested designs and implementations of such models tailored to different cases of use and objectives are described in the existing literature. They are not presented in this document.



Decision on technique	
Statistical Tool	Sampling
	Aggregation
Deletion	Masking
	Identifier Deletion
	Partial Deletion
	Record Deletion
Generalization	Rounding
	Top/Bottom coding
Randomization	Noise addition
	Permutation
	Micro aggregation
Cryptographic tool	Deterministic encryption
	Order-preserving encryption
	Homomorphic encryption
	Homomorphic secret sharing
Pseudonymization technique	Two-way encryption
	One-way encryption
	Token technique
	Random assign technique
	Heuristic pseudonymization

Figure 7: Risk treatment

We need risk treatment as an appropriate measure to eliminate or reduce a risk. The risk treatment in this study includes Statistical tool, Deletion, Generalization, Cryptographic tool, Pseudonymization technique, etc. We can respond to risks by using these techniques appropriately. Statistical tool is a technique that analyzes a part of the population through techniques such as generalization or random record extraction for a sample rather than analyzing the entire population. Statistical tools can achieve a similar result with analyzing the whole population at a lower cost. Therefore, it is important to obtain a representative sample of the population. Sample data has a relatively low risk of identification compared to population data. Deletion is a technique that removes a record, an attribute value, or a selected record from a dataset. This technique includes techniques such as masking, record deletion, and column deletion. Generalization, also called categorization, is a de-identification technique that replaces certain values with higher-level attributes. Techniques such as rounding, random rounding, control rounding, top and bottom coding, and local generalization are included. Cryptographic tool

is used to implement security measures that improve the efficiency of de-identification technology, and is an encryption technique that can be used as de-identification technology. Pseudonymization technique is a kind of non-identification technique in which the identifier of a data subject is replaced with an indirect identifier specially created for each data subject.

3.5.1 Statistical Tool

Statistical tools include sampling and aggregation. Sampling [55] is a technique that analyzes a part of the population through techniques such as generalization or random record extraction to a sample rather than the entire population for each data subject. Statistical tools can achieve results similar to analyzing the whole population at only a part of the cost of analyzing the whole data. Therefore, it is important to obtain a representative sample of the population, and sample data has a relatively low risk of identification compared to population data.

Aggregation [56] is a technique that statistically processes a specific column. All or part of the data is processed as an aggregation. In general, the aggregation method is usually processed with one of average, maximum, minimum, mode, and median values. The average value is a technique commonly used in statistics replacing the data with the average value of the whole.

3.5.2 Deletion Tool

Masking [57] is a technique that prevents identification by replacing some or all of the information with other letters. The biggest advantage of masking is that the range of application can be selected, so if masking is applied only to some parts, the usefulness of information on the unmasked part can be maintained. Data masking techniques can be classified into replacement, scrambling, encryption, data blurring, and deletion, depending on how they are changed.

Identifier deletion [10] is a method of deleting all identifiable elements. HIPAA's Safe Harbor method is a representative case that is actually being used. HIPAA recommends that the following 18 identifiers are defined and all of them are deleted.

Partial deletion [10] is a technique that lowers the identification of data by deleting a part of the column. There are a variety of methods, such as deleting part of the address or part of the date. When deleting, if the deletion range is too small, it may not be possible to lower the identification of the data, so it should be carefully deleted.

Record deletion [58] is a method of deleting the record for highly identifiable records. Data corresponding to outliers can adversely affect the analysis results, and the identifiability is very high. Eliminating these values can also help achieving the purpose of the analysis. However, only when the purpose of the analysis is to study general characteristics, and in the case of analysis on an unusual value, removing it may adversely affect the analysis.

3.5.3 Generalization

Generalization includes rounding [59], Top & Bottom Coding, etc. In general, it is often used when overall statistical information rather than detailed information is needed and can also be used as a real method of categorization. Random rounding is a rounding technique that allows us to freely specify the number of digits and the standard number of rounding. Control rounding is a rounding technique which the sum of the original row and column and the sum of the row and column are same after the rounding is applied.

In Top & Bottom Coding [10], information skewed at both ends of the data with the characteristic of normal distribution has a small number of distributions, so that individual identifiability can be obtained. It is a technique that lowers the individual's discrimination by categorizing it. This is a technique used for de-identification measures when the frequency of a specific value is very small or has a very peculiar value among data.

3.5.4 Randomization

There are noise addition, permutation, and micro aggregation in randomization. Noise addition [60] is a task of adding noise at random. When adding noise, the same noise should be applied to the related column.

Permutation [61] is a technique that randomly changes the order of data in a specific column. It is a technique that has a very high degree of damage to the data, so it requires a careful attention in selecting conditions for randomly changing the order. It should be applied only when the analysis purpose is not related to the values of other columns. In case of rearrangement under certain conditions, it is possible to minimize the impact on the analysis result while lowering the discrimination.

Micro aggregation [62] is a technique that partially applies aggregate processing according to specific conditions. A specific column in the homogeneous set is totaled or there is a value that is too peculiar to a specific condition, so the possibility of individual identification is high, but it is processed when the value is essential for analysis.

3.5.5 Cryptographic Tool

Cryptographic tools include Deterministic Encryption, Order-Preserving Encryption, Format-Preserving Encryption, Homomorphic Encryption, and Homomorphic secret sharing. Cryptographic Tool is a kind of de-identification technique.

Deterministic encryption [63] is an encryption technique in which the encrypted value is always generated as a constant value when the same value is encrypted with the same algorithm and the same key. This technique enables point lookup, equal join, grouping, and indexing of encrypted columns. And if the encrypted value set is small, it is possible to guess the information about the encrypted value.

Order-Preserving Encryption [64] is an encryption method in which the order of the original data and the order of encrypted password values are kept same. The order of the originals is maintained even in the encrypted state, which overcomes the decrease in search speed. In addition, the encryption strength is relatively low, and the order itself becomes important information, so it involves possibility of identification.

Format-Preserving Encryption [65] is an encryption method that converts encrypted data into a series of symbolic formats having the same format and length as the original data. It is a method of reducing the cost of data storage space due to encryption, and encryption can be applied without changing the system in a system with an existing process. Among pseudonymization techniques, form-preserving encryption is often used from token technique to token generation technique.

Homomorphic Encryption [66] was first proved its technical potential in 2009 by Gentry, an IBM (International Business Machine) researcher, after theoretical research began in the 1970s. It is an encryption method that can perform four arithmetic operations in an encrypted state. It can be used for various analysis by performing calculation processing in an encrypted state of original values. It

is an encryption method that can be used in an encrypted state and if necessary, decrypted in a safe environment to extract the original value.

Homomorphic secret sharing [67] is a technique that replaces an identifier or other characteristic information with two or more shares generated by a message sharing algorithm. It divides identifiers or other attribute values into multiple shares using mathematical operations and distributes them to share-holders. Although the performance overhead for calculation is relatively low, additional overhead is incurred when exchanging shares with the share owner, and significant performance costs may be incurred depending on the usage method.

3.5.6 Pseudonymization Technique

Pseudonymization techniques include two-way encryption, one-way encryption, token, random assignment, and heuristic pseudonymization.

Two-way encryption [68] is a pseudonymization technique in which data is encrypted and then the encrypted value is replaced with the original data. The pseudonymization technique using two-way encryption can be used as a pseudonymization that can be returned to the original when additional information is stored.

One-way encryption [69] cannot be reverted to the original even if additional information is stored. When it is necessary to combine it with previously pseudonymized data due to the need for time series analysis, or when additional information is retained, it can be combined and analyzed. When applying pseudonymization through one-way encryption, the use of salt is essential as a means to defend against rainbow table attacks as well as the safety of the algorithm used, and it is appropriate to give the salt length of at least 32 bytes.

Token [20] is used to protect personal information such as credit cards in the payment system of the financial sector. This is a technique that converts personal information into tokens generated by applying techniques such as random number or one-way encryption, and uses tokens where personal information should be used.

Random assignment [10] is a method of replacing values corresponding to identifiers with random values through a dictionary or the like, without a set rule. A table (mapping table) containing the mapping (or assignment) of the original identifier and the pseudonym may be created by creating a pseudonym independently for each identifier, but some techniques do not create such a table.

Heuristic pseudonymization [10] is a method of hiding detailed personal information by replacing values corresponding to identifiers with certain rules or processing them under human judgment. Since all data is processed in the same way without considering the distribution of identifiers or prior analysis of the collected data, users can easily understand and use them, but if the rules are relatively exposed, the possibility of individual identification is very high. In recent years, heuristic pseudonymization is a trend of using a sophisticated pseudonymization method through analysis of the distribution of identifiers or data to solve this problem.

3.6 Dataset Release

The reason for performing de-identification is to use the dataset safely and usefully. If, in the previous steps, an appropriate level of de-identification was performed while maintaining the utility of the dataset in an optimal manner according to the situation, this dataset should be utilized. In order to use this dataset, release is eventually required, and the release model varies depending on the level and purpose of the dataset's de-identification.

The model for making a dataset public depends on the degree of de-identification. Basically, the model for releasing dataset is divided into public model, semi-public model, and private release model. Anonymized datasets that have been sufficiently de-identified can be publicly disclosed because re-identification is impossible. When de-identification is minimal and separate rules are required, the private release model should be used. An example of a representative model is the enclave model. When disclosing pseudonymized data, a semi-public model or a private release model can be used.

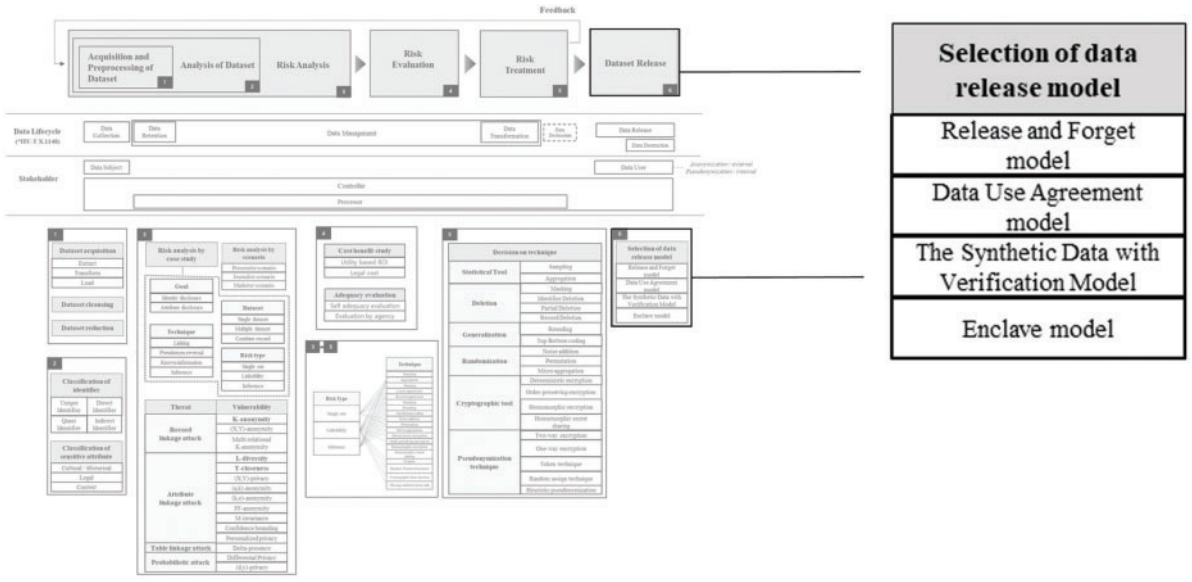


Figure 8: Dataset release

NIST SP 800-188 [62] presents the following four models. Release and Forget models correspond to public models. The DUA (Data Usage Agreement) model belongs to the semi-public model or the private release model depending on the degree of restriction. The Synthetic Data with Verification model corresponds to a public model or a semi-public model according to the threshold value of differential privacy. And the enclave model is one of the representative private release models.

- Release and Forget model: This model is a fully open model that can be released and forgotten as the term is. This applies to data that has been completely de-identified, since in order to release and forget, we must not worry about the risk of re-identification. Anonymized data (de-identified data) refers to a dataset in a state that is not re-identified even if it is made public through the Internet, etc. and people use it freely. Therefore, it is a model that freely publishes such datasets in the desired format.
- Data Usage Agreement (DUA) Model: A dataset that is not anonymized (de-identified), but has performed an appropriate level of de-identification for data usefulness, is exposed to the risk of re-identification. When sharing and using such datasets, the most appropriate model is the data usage agreement model. This model assumes that certain conditions are observed when using the dataset, and the dataset is delivered only to a specific organization. The most commonly used conditions include the content not to attempt re-identification and the content not to be additionally shared with other organizations.
- Synthetic data with validation model: This model is a model that creates a new dataset to prevent re-identification while maintaining the usefulness of the dataset. It is to create a new dataset

while maintaining the distribution of the various statistics and figures in the dataset from an overall perspective. Only highly trained and qualified researchers can create this type of dataset.

- The Enclave model: This model performs minimal de-identification on the dataset and stores the data only on a secure server without delivering the dataset. And organizations that want to conduct research using the dataset cannot see the dataset, but can only see the results of analyzing the dataset through the code of a set format. If this model is used, appropriate statistical data and results can be obtained, but a lot of preparation is required in advance. Without sufficient prior preparation, it is difficult to utilize the dataset, and it may not be possible to utilize the dataset rather than using the released dataset with the relatively low-utility release and forget model.

4 Discussion

In this paper, a framework including the overall process for de-identification is presented. Since this framework is presented by analyzing and organizing a vast amount of data including various laws and standards, this framework can be used effectively in terms of risk management. However, it is not clear whether this framework will be used properly. This is because when de-identification is performed in general organizations, it is not performed for the purpose of protecting privacy, but in most cases, it is performed for the purpose of not violating the law. From a broader perspective, using this framework helps protecting privacy while not breaking the law, but each organization wants to take minimal action. Therefore, there are many cases of mechanically de-identifying each identifier by referring to the guidelines published by the country to which each organization belongs. This type of de-identification is not wrong. However, technology continues to develop, and the organic connection among identifiers is not properly studied. And by referring to various existing cases, new re-identification attacks are continuously emerging. In such a situation, the guidelines issued by the state considering only general circumstances may be out of date or not be able to assume specific situations. Therefore, in order to de-identify in the most effective way, each organization analyzes its own systems and procedures, and performs de-identification measures optimized for the context of the dataset. This framework is effective to perform de-identification in this way. This framework will be greatly helpful for solutions that are currently being developed or will be developed in the future. Currently, various de-identification solutions such as AMP (Approximate Minima Perturbation), ARX Data Anonymization Tool, and Chorus have been developed. Each solution has its own set of strengths, but in the end, they can be effective when the user sets the rules in an appropriate way. If this de-identification framework can be applied to each solution to manage risk by user-friendly operation, the performance will be much better.

This framework can be specialized for multiple domains. For example, in the medical domain, the analysis dataset part can focus on the content that distinguishes sensitive attributes and the part that discloses the dataset. Medical information basically includes a number of sensitive attributes, and it is necessary to configure sensitive attributes suitable for each treatment and operation through various methods presented in this framework. In addition, medical data is not generally disclosed to the public, and an enclave model is used for cohort studies. Several models in this framework can make the use of datasets more useful. And for the financial domain, we can focus on the safe usage of data. In order to properly de-identify financial data, we can focus on utility-based ROI calculations in the risk evaluation part and de-identification treatments according to each risk type that are linked in case studies. Currently, financial data is de-identified in a manner determined for each feature, but in the future, de-identification should be performed according to the current situation based on a case study.

5 Conclusion

With the development of data collection and processing technology, big data analysis has become an essential element for companies to secure a competitive advantage. Many companies set their top priority to collect and accumulate data in large quantities, and use advanced algorithms to attract citizens to make decisions to the direction they want. However, national standards and laws are being developed to guarantee the rights of data subjects. The rights of the data subject are important as a concept of data sovereignty, not allowing companies to use big data freely. Therefore, de-identified/pseudonymized data is needed to provide advanced services while securing data subject rights.

It is difficult to know how and in what specific ways each researcher and institution must de-identify their data. While there are many studies on de-identification techniques and pseudonymization techniques, there has been no work to systematize the latest knowledge by collecting these techniques. Each knowledge was divided into paper, law, report, and standard. Therefore, in this paper, a systematic framework is presented by collecting techniques related to de-identification. This model enables companies to systematically analyze their data when they de-identify it and receive the de-identified data based on a re-identification scenario. Each company can make an optimized decision for the situation based on the results of the analysis. In addition, it can be used as an authentication system for non-identification adequacy evaluation. In the future, it is necessary to collect many samples based on this model and develop an optimized model for each industry.

Efforts to protect personal information are essential to vitalizing the data economy. If only the use of personal information is emphasized to revitalize the data economy, it is possible to face extreme opposition by forcing the sacrifice of privacy for the data subject. In the end, the data economy, which is the core of the 4th Industrial Revolution, will not be activated and national competitiveness will decline. Therefore, personal information de-identification measures that support the use of data by converting personal information into pseudonymized data and anonymous information will play a role as an axis of vitalization of the data economy. To this end, technology development for securing the stability and usefulness of data must be followed. In addition, with the development of technology, social awareness and consensus on the safe usage of data must be accompanied. Personal information de-identification measures technology to revitalize the data economy requires data experts who have the capability to use the technology and understand data along with the development of the technology. In the process of performing adequacy assessment through experts, it is necessary to be able to determine personal information de-identification measures techniques and privacy protection models. In addition, it must be able to demonstrate statistical analysis ability according to data transformation. To this end, a de-identification framework is essential in the future society. Through the de-identification framework, organizations can more efficiently and systematically de-identify personal information. In that case, the likelihood of paying fines for violating legal regulations will be decreased, and there will be less chances of re-identification incidents becoming an issue. In addition, it is expected that data sharing will become more active as people becomes more positive about data sharing.

Acknowledgement: We deeply acknowledge Korea University supporting this study.

Funding Statement: This work was supported by a Korea University Grant.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] M. Brundage, "Taking superintelligence seriously: Superintelligence: Paths, dangers, strategies by nick bostrom (Oxford university press, 2014)," *Futures*, vol. 72, pp. 32–35, 2015.

[2] J. C. Ramo, *The Seventh Sense: Power, Fortune, and Survival in the age of Networks*, NW, USA: Little, Brown, 2016.

[3] L. Cummings, "Normal accidents: Living with high-risk technologies," *Administrative Science Quarterly*, vol. 29, no. 4, pp. 630–632, 1984.

[4] S. M. He, W. N. Zeng, K. Xie, H. M. Yang, M. Y. Lai *et al.,* "PPNC: Privacy preserving scheme for random linear network coding in smart grid," *KSII Transactions on Internet & Information Systems*, vol. 11, no. 3, pp. 1510–1532, 2017.

[5] K. Gu, W. J. Jia and J. M. Zhang, "Identity-based multi-proxy signature scheme in the standard model," *Fundamenta Informaticae*, vol. 150, no. 2, pp. 179–210, 2017.

[6] Z. Xu, C. Xu, J. Xu and X. Meng, "A computationally efficient authentication and key agreement scheme for multi-server switching in WBAN," *International Journal of Sensor Networks*, vol. 35, no. 3, pp. 143–160, 2021.

[7] NIST, *De-identification of Personal Information*, 2015. [Online]. Available: http://dx.doi.org/10.6028/NIST.IR.8053.

[8] C. G. Miller, J. Krasnow and L. H. Schwartz, *Medical Imaging in Clinical Trials*, London, England: Springer, 2014. [Online]. Available: https://link.springer.com/book/10.1007/978-1-84882-710-3#bibliographic-information.

[9] K. Ito, J. Kogure, T. Shimoyama and H. Tsuda, "De-identification and encryption technologies to protect personal information," *Fujitsu Sci. Tech. J.*, vol. 52, no. 3, pp. 28–36, 2016.

[10] ISO/IEC, *Privacy Enhancing Data De-identification Terminology and Classification of Techniques*, 2018. [Online]. Available: https://www.iso.org/standard/69373.html.

[11] S. Zinsmaier, H. Langweg and M. Waldvogel, "A practical approach to stakeholder-driven determination of security requirements based on the GDPR and common criteria," in *6th Int. Conf. on Information Systems Security and Privacy*, Valletta, Malta, pp. 473–480, 2020.

[12] C. Tziogas and N. Tsolakis, "The dawn of GDPR: Implications for the digital business landscape," in *Springer Proc. in Business and Economics*, Switzerland: Springer, pp. 623–627, 2019.

[13] P. Cheimonidis, "The responsibilities of the DPO according to the GDPR," M.S. thesis, School of Science and Technology, International Hellenic University, Thessaloniki, Greece, 2019.

[14] S. T. Liaw, J. G. N. Guo, S. Ansari, J. Jonnagaddala, M. A. Godinho *et al.,* "Quality assessment of real-world data repositories across the data life cycle: A literature review," *J. Am. Med. Inform. Assoc.*, vol. 28, no. 7, pp. 1591–1599, 2021.

[15] H. Khaloufi, K. Abouelmehdi, A. Beni-hssane and M. Saadi, "Security model for big healthcare data lifecycle," in *The 8th Int. Conf. on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2018)*, Leuven, Belgium, pp. 294–301, 2018.

[16] ITU-T, *Framework of De-identification Process for Telecommunication Service Providers*, 2020. [Online]. Available: https://handle.itu.int/11.1002/1000/14249.

[17] T. Aven and O. Renn, "Risk management," in *Risk management and governance: Concepts, guidelines and applications*, Berlin, Germany: Springer, pp. 121–158, 2010.

[18] E. M. Faustman and G. S. Omenn, "Risk assessment," in *Casarett and Doull's Toxicology: The Basic Science of Poisons*, NY, USA: United States Environmental Protection Agency, pp. 107–128, 2008.

[19] V. Page, M. Dixon and I. Choudhury, "Security risk mitigation for information systems," *BT Technology Journal*, vol. 25, no. 1, pp. 118–127, 2007.

[20] E. Androulaki, J. Camenisch, A. D. Caro, M. Dubovitskaya, K. Elkhiyaoui *et al.,* "Privacy-preserving auditable token payments in a permissioned blockchain system," in *Proc. of the 2nd ACM Conf. on Advances in Financial Technologies*, NY, USA, pp. 255–267, 2020.

[21] S. Wold, K. Esbensen and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[22] D. Ness, "Information overload: Why omnipresent technology and the rise of big data shouldn't spell the end for privacy as we know it," *Cardozo Arts & Entertainment Law Journal*, vol. 31, pp. 925, 2012.
[23] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer *et al.,* "Privacy expectations and preferences in an iot world," in *Thirteenth Symp. on Usable Privacy and Security ({SOUPS} 2017)*, CA, USA, pp. 399–412, 2017.
[24] S. Zheng, N. Apthorpe, M. Chetty and N. Feamster, "User perceptions of smart home iot privacy," in *Proc. of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–20, 2018.
[25] D. Kim, K. Park, Y. Park and J. -H. Ahn, "Willingness to provide personal information: Perspective of privacy calculus in iot services," *Computers in Human Behavior*, vol. 92, pp. 273–281, 2019.
[26] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, no. 2000, pp. 1–34, 2000.
[27] K. El Emam, E. Jonker, L. Arbuckle and B. Malin, "A systematic review of re-identification attacks on health data," *PloS One*, vol. 6, no. 12, pp. e28071, 2011.
[28] A. Narayanan and V. Shmatikov, "How to break anonymity of the netflix prize dataset," arXiv preprint, 2006. [Online]. Available: https://arxiv.org/abs/cs/0610105.
[29] X. Qian, "Big data analytics with nyc taxicab data," Ph.D. dissertation, Purdue University, US, 2014.
[30] H. Y. Youm, "An overview of de-identification techniques and their standardization directions," *IEICE TRANSACTIONS on Information and Systems*, vol. 103, no. 7, pp. 1448–1461, 2020.
[31] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: K-anonymity and its enforcement through generalization and suppression," *Data Privacy Lab*, 1998. [Online]. Available: https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf.
[32] K. Wang and B. C. Fung, "Anonymizing sequential releases," in *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, PA, USA, pp. 414–423, 2006.
[33] M. E. Nergiz, C. Clifton and A. E. Nergiz, "Multirelational kanonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 8, pp. 1104–1117, 2008.
[34] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "L-diversity: Privacy beyond K-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–es, 2007.
[35] N. Li, T. Li and S. Venkatasubramanian, "T-closeness: Privacy beyond kanonymity and L-diversity," in *2007 IEEE 23rd Int. Conf. on Data Engineering*, Birmingham, UK, pp. 106–115, 2007.
[36] R. C. -W. Wong, J. Li, A. W. -C. Fu and K. Wang, "(A, k)-anonymity: An enhanced K-anonymity model for privacy preserving data publishing," in *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, PA, USA, pp. 754–759, 2006.
[37] Q. Zhang, N. Koudas, D. Srivastava and T. Yu, "Aggregate query answering on anonymized tables," in *2007 IEEE 23rd Int. Conf. on Data Engineering*, Birmingham, UK, pp. 116–125, 2007.
[38] K. Wang, Y. Xu, A. W. Fu and R. C. Wong, "FF-anonymity: When quasiidentifiers are missing," in *2009 IEEE 25th Int. Conf. on Data Engineering*, Shanghai, China, pp. 1136–1139, 2009.
[39] X. Xiao and Y. Tao, "M-Invariance: Towards privacy preserving republication of dynamic datasets," in *Proc. of the 2007 ACM SIGMOD Int. Conf. on Management of Data*, Beijing, China, pp. 689–700, 2007.
[40] K. Wang, B. C. Fung and S. Y. Philip, "Handicapping attacker's confidence: An alternative to K-anonymization," *Knowledge and Information Systems*, vol. 11, no. 3, pp. 345–368, 2007.
[41] X. Xiao and Y. Tao, "Personalized privacy preservation," in *Proc. of the 2006 ACM SIGMOD Int. Conf. on Management of Data*, Chicago IL USA, pp. 229–240, 2006.
[42] M. E. Nergiz, M. Atzori and C. Clifton, "Hiding the presence of individuals from shared databases," in *Proc. of the 2007 ACM SIGMOD Int. Conf. on Management of Data*, Beijing, China, pp. 665–676, 2007.
[43] V. Rastogi, D. Suciu and S. Hong, "The boundary between privacy and utility in data publishing," in *Proc. of the 33rd Int. Conf. on Very Large Data Bases*, Vienna, Austria, pp. 531–542, 2007.
[44] I. Mironov, "On significance of the least significant bits for differential privacy," in *Proc. of the 2012 ACM Conf. on Computer and Communications Security*, NC, USA, pp. 650–661, 2012.

[45] U. Erlingsson, V. Pihur and A. Korolova, "Rappor: Randomized ag-gregatable privacy-preserving ordinal response," in *Proc. of the 2014 ACM SIGSAC Conf. on Computer and Communications Security*, Arizona, USA, pp. 1054–1067, 2014.
[46] R. Hall, A. Rinaldo and L. Wasserman, "Random differential privacy," arXiv preprint, 2011. [Online]. Available: https://arxiv.org/abs/1112.2680.
[47] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke and L. Vilhuber, "Privacy: Theory meets practice on the map," in *2008 IEEE 24th Int. Conf. on Data Engineering*, Cancun, Mexico, pp. 277–286, 2008.
[48] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," arXiv preprint, 2016. [Online]. Available: https://arxiv.org/abs/1603.01887.
[49] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conf.*, Tel Aviv, Israel, pp. 635–658, 2016.
[50] P. Kairouz, S. Oh and P. Viswanath, "Secure multi-party differential privacy," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2008–2016, 2015.
[51] I. Mironov, O. Pandey, O. Reingold and S. Vadhan, "Computational differential privacy," in *Annual Int. Cryptology Conf.*, California, USA, pp. 126–142, 2009.
[52] K. Gu, L. H. Yang and B. Yin, "Location data record privacy protection based on differential privacy mechanism," *Information Technology and Control*, vol. 47, no. 4, pp. 639–654, 2018.
[53] O. Tambou, "Lessons from the first post-GDPR fines of the CNIL against google LLC," *Eur. Data Prot. L. Rev.*, 2019. [Online]. Available: https://doi.org/10.21552/edpl/2019/1/13.
[54] J. F. Carrez, A. Linden, D. Castera, M. H. Mitzavile, M. Ronal *et al.,* "Deliberation of the restricted committee SAN-2019-001 of 21 January 2019 pronouncing a financial sanction against GOOGLE LLC.," 2019. [Online]. Available: https://www.cnil.fr/sites/default/files/atoms/files/san-2019-001.pdf.
[55] K. Chaudhuri and N. Mishra, "When random sampling preserves privacy," in *Annual Int. Cryptology Conf.*, CA, USA, pp. 198–213, 2006.
[56] E. Shi, T. H. Chan, E. Rieffel, R. Chow and D. Song, "Privacy-preserving aggregation of time-series data," in *Proc. of Network and Distributed System Security Symp.*, CA, USA, vol. 2, pp. 1–17, 2011.
[57] K. Wada and K. Sakurama, "Privacy masking for distributed optimization and its application to demand response in power grids," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 6, pp. 5118–5128, 2017.
[58] B. J. Keele, "Privacy by deletion: The need for a global data deletion principle," *Indiana Journal of Global Legal Studies*, vol. 16, no. 1, pp. 363–384, 2009.
[59] L. Cox and L. Ernst, "Controlled rounding," *INFOR: Information Systems and Operational Research*, vol. 20, no. 4, pp. 423–432, 1982.
[60] K. Mivule, "Utilizing noise addition for data privacy, an overview," arXiv preprint, 2013. [Online]. Available: https://arxiv.org/abs/1309.3958.
[61] X. He, Y. Xiao, Y. Li, Q. Wang, W. Wang *et al.,* "Permutation anonymization: Improving anatomy for privacy preservation in data publication," in *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, CA, USA, pp. 111–123, 2011.
[62] J. Domingo-Ferrer, "Microaggregation for database and location privacy," in *Int. Workshop on Next Generation Information Technologies and Systems*, Kebbutz Sehfayim, Israel, pp. 106–116, 2006.
[63] M. Bellare, M. Fischlin, A. O'Neill and T. Ristenpart, "Deterministic encryption: Definitional equivalences and constructions without random oracles," in *Annual Int. Cryptology Conf.*, CA, USA, pp. 360–378, 2008.
[64] V. Kolesnikov and A. Shikfa, "On the limits of privacy provided by orderpreserving encryption," *Bell Labs Technical Journal*, vol. 17, no. 3, pp. 135–146, 2012.
[65] M. Bellare, T. Ristenpart, P. Rogaway and T. Stegers, "Format-preserving encryption," in *Int. Workshop on Selected Areas in Cryptography*, Alberta, Canada, pp. 295–312, 2009.
[66] F. D. Garcia and B. Jacobs, "Privacy-friendly energy-metering via homomorphic encryption," in *Int. Workshop on Security and Trust Management*, Athens, Greece, pp. 226–238, 2010.
[67] R. W. Lai, G. Malavolta and D. Schroder, "Homomorphic secret sharing¨ for low degree polynomials," in *Int. Conf. on the Theory and Application of Cryptology and Information Security*, Brisbane, Australia, pp. 279–309, 2018.

[68] P. Kukade, R. Tale, S. Thakre, A. Sonwane and R. Jain, "A two-way encryption for privacy preservation of outsourced transaction database for association rule mining," *Int. J. Sci. Res. Sci. Technol.*, vol. 4, pp. 276–285, 2018.

[69] M. Bellare, A. Boldyreva, A. Desai and D. Pointcheval, "Key-privacy in public-key encryption," in *Int. Conf. on the Theory and Application of Cryptology and Information Security*, Gold Coast, Australia, pp. 566–582, 2001.