Guidelines for De-identification of Personal Data

- Guide for De-identification Standards and Support/Management System -

June 30, 2016

Office for Government Policy Coordination | Ministry of Interior
Korea Communications Commission | Financial Services Commission
Ministry of Science, ICT and Future Planning | Ministry of Health and Welfare

◇ Emergence of new data technologies such as big data and Internet of Things (IoT), and the convergence technology has provided an opportunity to Korea, an IT power house, to make a new leap forward. Despite the fact, concerns over personal data infringement, which can be possibly resulted from the use of new data technologies, present us a challenge to strike a balance between development of new industries and personal data protection.

◇ In this regard, the guidelines for De-identification of personal data was published under the joint leadership of the Korean government including the Office for Government Policy Coordination, the Ministry of Interior, the Korea Communications Commission, the Financial Services Commission, the Ministry of Science, ICT and Future Planning, and the Ministry of Health and Welfare to provide clear standards on de-identification of personal data and scopes on utilizing de-identified data for the safe use of big data within the current legal boundary of personal data protection. By doing so, we aim to reduce uncertainties that businesses have in order to encourage corporate investment and industry development as well as strive our best to protect personal data.

◇ Furthermore, any de-identified data which has undergone appropriate de-identification measures in accordance with the guidelines is presumed non personal data, thus can be used for a big data analysis or other purposes.

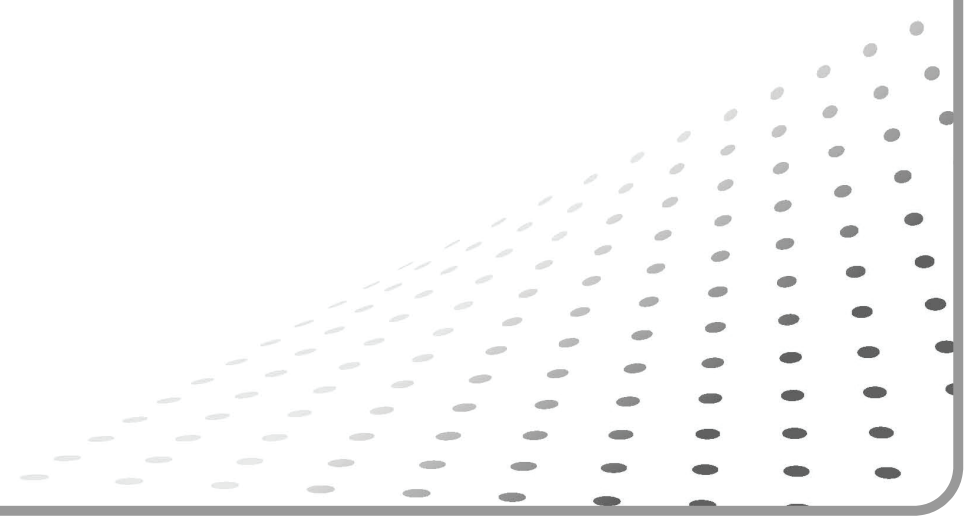
◇ One thing to note about is that re-identification of de-identified data is not entirely impossible due to technology development and increasing the amount of data. Therefore, businesses should take safety measures in terms of management and technology to prevent re-identification for the safe use of de-identified data.

June 30, 2016

Interdepartmental Joint Announcement

Table of Contents

Ⅰ. Background



Ⅰ Background

◈ Leading countries in IT such as the US and UK are implementing policies to promote the data industry as demand for the use of data increases due to development of IT convergence technologies such as big data and IoT.

◈ The Korean government aims to lay the foundation for the safe use of big data and strengthen personal data protection by providing detailed standards, procedures, and methods of de-identification for big data utilization.

1 Increased Value of Data in Its Usage Thanks to Expansion of the Government 3.0* and Big Data Utilization

○ Providing and sharing of public data are essential to achieve transparent and efficient government operation while big data utilization is absolutely necessary to execute science policies and provide customized services.

* *In 2013, the central government and municipalities put forward the Government 3.0 as a new governance paradigm aimed at creating new markets and jobs by making government data available to public.*

○ In particular, creation of new services through big data analysis and IoT technologies, and increasing value of big data in its usage are highly expected.

2 Continuous Social Demand for Enhanced personal data Protection

○ The Korean society has raised its voice for a need to enhance personal data protection policies as personal data leakage, regardless of its size, continuously breaks out.

○ Also, a risk of possible personal data infringement is on the rise along with development of new industries and technologies that require utilization of various data.

3 Active Response to Political Changes of the World that Seeks Harmony between Protection and Utilization

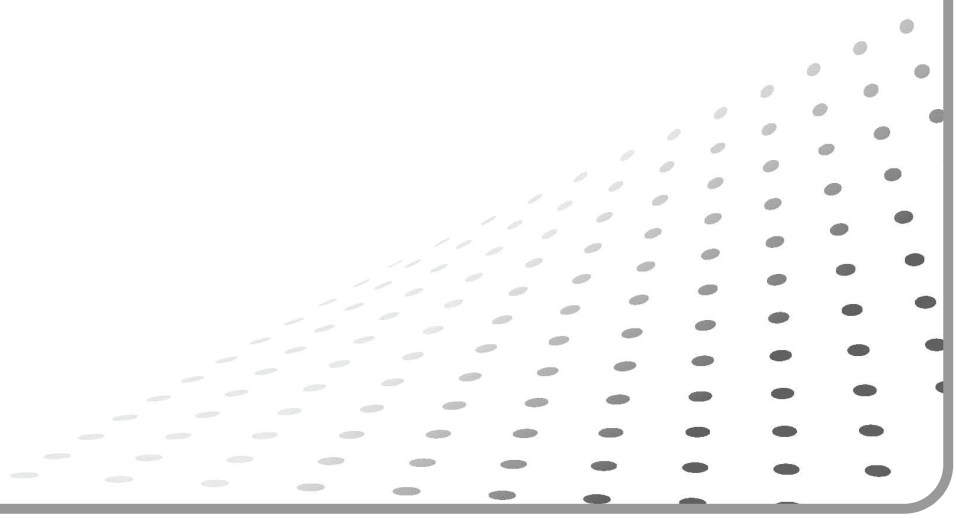
○ Leading countries in IT such as the US and UK are implementing policies to promote the data industry while minimizing the possibility of personal data infringement.

○ There is a need to come up with safety measures to prevent infringement of privacy and provide a detailed guidelines to well utilize de-identified data at the industrial level.

II. Standards of De-identification

1. Overview

2. Standards for De-identification at Each Step



Ⅱ Standards of De-identification

1 Overview

○ The guidelines provide standards for businesses which intend to use or provide de-identified personal data.

※ personal data shall be managed under a applicable law when collected and used in accordance with related laws namely the Statistics Act.

○ Measures to take at each step

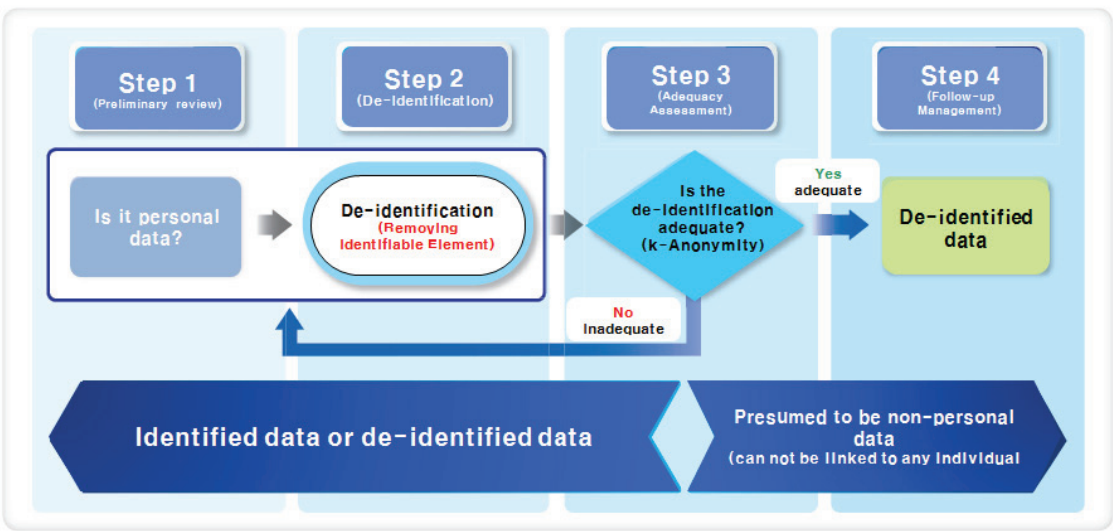
① **(Preliminary Review)** Verify whether specific data is personal data or not. data that is confirmed to be non-personal data can be used as freely as possible without any legal restrictions.

② **(De-identification)** Take measures to make individuals unidentifiable by means of, for example, deleting or replacing all or some of the elements from dataset.

③ **(Adequacy Assessment)** Assess whether an individual is identifiable by easily combining de-identified data with other data by the “De-identification Adequacy Assessment Panel” (the Assessment Panel).

④ **(Follow-up Management)** Take necessary measures such as safety measures for de-identified data and monitoring likelihood of re-identification to prevent re-identification in the process of using de-identified data.

<Procedure for De-identification and Follow-up Management>



2 Standards for De-identification at Each Step

2-① Preliminary Review : Review whether specific data is personal data or not.

○ Businesses which intend to process data for a number of purposes including big data analysis shall refer to the below standards to verify whether specific data is personal data or not.

○ If the specific data is clearly confirmed to be non-personal data, it can be used for big data analysis without additional measures.

⇨ Otherwise, taking a next step is necessary.

<Reference> Judgment criteria for personal information

A. The concept of personal data provided in related laws such as the Personal Data Protection Act is as follows. If not applicable, specific data is not personal data.

B. personal data means ⓐ data that can ⓑ identify a ⓒ living ⓓ person and includes ⓔ data that can be easily combined with other data when given data is not enough to identify a specific person.

ⓐ There is no special limitation on data regarding its type, form, characteristics, and format.

ⓑ data that is unable to identify a specific person is not personal data.

* A subject who "identifies" an individual is a personal data controller (including third party who is provided data). data that can't be used to identify an individual from the perspective of personal data controller is not personal data.

ⓒ data about a dead person, corporate, group or object is not personal data as it should be about a living person.

ⓓ data should be about a person, so statistical value of a group consists of multiple individual is not personal data.

ⓔ 'easy combination with other data' means that it should be possible to obtain other information to be combined and there is a high possibility of combining with other information.

* In other words, a concept of easy combination with other data does not include data that can't be legally collected and requires irrational amount of time or costs for collection.

□ De-identification Standards for Identifiers

○ As a rule, an identifier* included in datasets should be deleted.

* An "identifier" is data such as a value or a name uniquely assigned to an individual or a thing related to the individual.

○ However, an identifier that is absolutely needed for a purpose of the use of data can be utilized after de-identification.

<Example> Identifiers

○ Unique identifier (resident registration number, passport number, alien registration number, and driver's license number)

○ Name (in Chinese characters, English name, and pen name)

○ Detailed address (house numbers and street addresses)

○ Date data : Date of birth (lunar/solar calendar), anniversary date (wedding, first birthday, 60th birthday, etc.), date of obtaining a certificate, etc.

○ Phone number (mobile, home, office, fax, etc.)

○ Medical record number, national health insurance number, welfare recipient number

○ Bank account number, credit card number

○ Certificates and driver's license number

○ License plate number, and registration number and serial number of different equipment

○ Photos (still picture, video, CCTV videos, etc.)

○ Biometric data (fingerprint, voice, iris, etc.)

○ E-mail address, IP address, Mac address, homepage URL, etc.

○ Identification code (ID, employee number, customer number, etc.)

○ Other unique identification number: Military service number, business registration number, etc.

※ Refer to the "HIPAA Privacy Rule" issued by the U.S. Department of Health and Human Services

☐ **De-identification Standards for Attribute Values**

○ As a rule, an attribute value included in datasets should be also removed if it is irrelevant to a purpose of the use of data.

- De-identification measures such as pseudonymization and aggregation should be taken if an attribute value related to the use of data has identifiable elements.

○ Attribute values like rare disease or unique professional experiences need strict de-identification as they are highly likely to identify an individual under certain circumstances.

<Example> Attribute Values

Individual Characteristics	○ Gender, age, nationality, hometown, address, postal code, military service record, marital status, religion, hobbies, club/group affiliation, etc. ○ Smoking, drinking, vegetarian diet, matter of interests, etc.
Physical Characteristics	○ Blood type, height, weight, waist, blood pressure, eye color, etc. ○ Physical examination results, disability type, disability grade, etc. ○ Disease name or code, administration code, medical record, etc.
Credit Characteristics	○ Tax payment, credit rate, donation, etc. ○ Health insurance payment, income level, medical service recipient, etc.
Career Characteristics	○ Name of school, major, grade, academic performance, and academic background, etc. ○ Professional experiences, occupation, job category, company, department, position, previous jobs, etc.
Electronic Characteristics	○ Cookie data, date and time of login, date and time of visit, record of using services, access log, etc. ○ Internet access record, record of using mobile phones, GPS data, etc.
Family Characteristics	○ Family data (spouse, children, parents, and siblings), legal representative data, etc.

□ De-identification Standards for Identifiers

○ As a rule, an identifier* included in datasets should be deleted.

* An "identifier" is data such as a value or a name uniquely assigned to an individual or a thing related to the individual.

○ However, an identifier that is absolutely needed for a purpose of the use of data can be utilized after de-identification.

<Example> Identifiers

○ Unique identifier (resident registration number, passport number, alien registration number, and driver's license number)

○ Name (in Chinese characters, English name, and pen name)

○ Detailed address (house numbers and street addresses)

○ Date data : Date of birth (lunar/solar calendar), anniversary date (wedding, first birthday, 60th birthday, etc.), date of obtaining a certificate, etc.

○ Phone number (mobile, home, office, fax, etc.)

○ Medical record number, national health insurance number, welfare recipient number

○ Bank account number, credit card number

○ Certificates and driver's license number

○ License plate number, and registration number and serial number of different equipment

○ Photos (still picture, video, CCTV videos, etc.)

○ Biometric data (fingerprint, voice, iris, etc.)

○ E-mail address, IP address, Mac address, homepage URL, etc.

○ Identification code (ID, employee number, customer number, etc.)

○ Other unique identification number: Military service number, business registration number, etc.

※ Refer to the "HIPAA Privacy Rule" issued by the U.S. Department of Health and Human Services

☐ **De-identification Standards for Attribute Values**

○ As a rule, an attribute value included in datasets should be also removed if it is irrelevant to a purpose of the use of data.

- De-identification measures such as pseudonymization and aggregation should be taken if an attribute value related to the use of data has identifiable elements.

○ Attribute values like rare disease or unique professional experiences need strict de-identification as they are highly likely to identify an individual under certain circumstances.

<Example> Attribute Values

Individual Characteristics	○ Gender, age, nationality, hometown, address, postal code, military service record, marital status, religion, hobbies, club/group affiliation, etc. ○ Smoking, drinking, vegetarian diet, matter of interests, etc.
Physical Characteristics	○ Blood type, height, weight, waist, blood pressure, eye color, etc. ○ Physical examination results, disability type, disability grade, etc. ○ Disease name or code, administration code, medical record, etc.
Credit Characteristics	○ Tax payment, credit rate, donation, etc. ○ Health insurance payment, income level, medical service recipient, etc.
Career Characteristics	○ Name of school, major, grade, academic performance, and academic background, etc. ○ Professional experiences, occupation, job category, company, department, position, previous jobs, etc.
Electronic Characteristics	○ Cookie data, date and time of login, date and time of visit, record of using services, access log, etc. ○ Internet access record, record of using mobile phones, GPS data, etc.
Family Characteristics	○ Family data (spouse, children, parents, and siblings), legal representative data, etc.

☐ Methods of De-identification

○ A range of methods including pseudonymization, aggregation, data suppression and data masking can be used individually or in combination.

※ Applying a pseudonymization method alone hardly seems as a sufficient de-identification.

○ There are various kinds of technology in place to realize each method. The most suitable methods and specific techniques need to be chosen and utilized based on the purpose of data usage, and strengths and weaknesses of each method.

⇨ Take a next step once de-identification is completed.

<Example> De-identification Methods

Methods	Example	Detailed Techniques
Pseudonymization	○ Hong Gildong, 35-year-old, living in Seoul, going to Korea University → Lim Ggeokjeong, in his 30s, living in Seoul, going to International University	① Heuristic Pseudonymization ② Encryption ③ Swapping
Aggregation	○ Lim: 180cm, Hong: 170cm, Lee: 160cm, Kim: 150cm → A total height of students major in physics: 660cm, Average height: 165cm	④ Aggregation ⑤ Micro Aggregation ⑥ Rounding ⑦ Rearrangement
Data Reduction	○ Resident registration number: 901206-1234567 → Born in the 1990s, male ○ Process date data related to a person in terms of a year → 1988. 12. 25 (Birthday) : 1988	⑧ Reducing Variables ⑨ Reducing Partial Variables ⑩ Reducing Records ⑪ Trivial Anonymization
Data Suppression	○ Hong Gildong, 35-year-old → Mr. Hong, in 30s～40s	⑫ Data Suppression ⑬ Random Rounding ⑭ Data Range ⑮ Controlled Rounding
Data Masking	○ Hong Gildong, 35-year-old, living in Seoul, going to Korea University → Hong ○○, 35-year-old, living in Seoul, going to △△ University	⑯ Adding Random Noise ⑰ Blank and Impute

2-③ Adequacy Assessment: Utilize a k-Anonymity Model

☐ **Need for Adequacy Assessment**

○ An individual could be identified by data combination or using various inference techniques when de-identification are not sufficient.

○ Thus, there is a need to form the Assessment Panel involving external professionals under the leadership of a privacy officer to strictly assess possibilities of identifying an individual.

○ The k-Anonymity Model will be used among other privacy protection models when assessing adequacy.

- k-Anonymity is a basic means of assessment. Additional assessment models (*l*-diversity and *t*-accessibility) can be applied if necessary.

< Case of Netflix (2006, US) >

○ Netflix, a company that provides online streaming media, held the Netflix Prize to increase the accuracy of algorithm that is used to recommend movies tailored to interests of their customers.

- Netflix provided a training dataset of 100,480,507 ratings that 480,189 users gave to 17,770 movies from Dec. 1999 to Dec. 2005.

※ Although names and other data that could identify users were removed, unique identifiers, movie reviews, and review dates were disclosed in order to connect with the outcome of data processing.

○ A group at the University of Texas identified certain individuals by combining a viewing history disclosed by Netflix and user reviews disclosed in the IMDb (Internet Movie Database).

※ IMDb posts IDs and evaluation scores on its website.

○ The 2nd Netflix Prize was cancelled as the US Fair Trade Commission (FTC) raised privacy concerns.

☐ **Adequacy Assessment Procedures**

① **(Preparing Basic Documents)** A personal data controller shall prepare basic documents needed for adequacy assessment such as a data statement, de-identification status, and the management level of user organizations*.

* 'user organization' means that an organization intends to utilize the de-identified data.

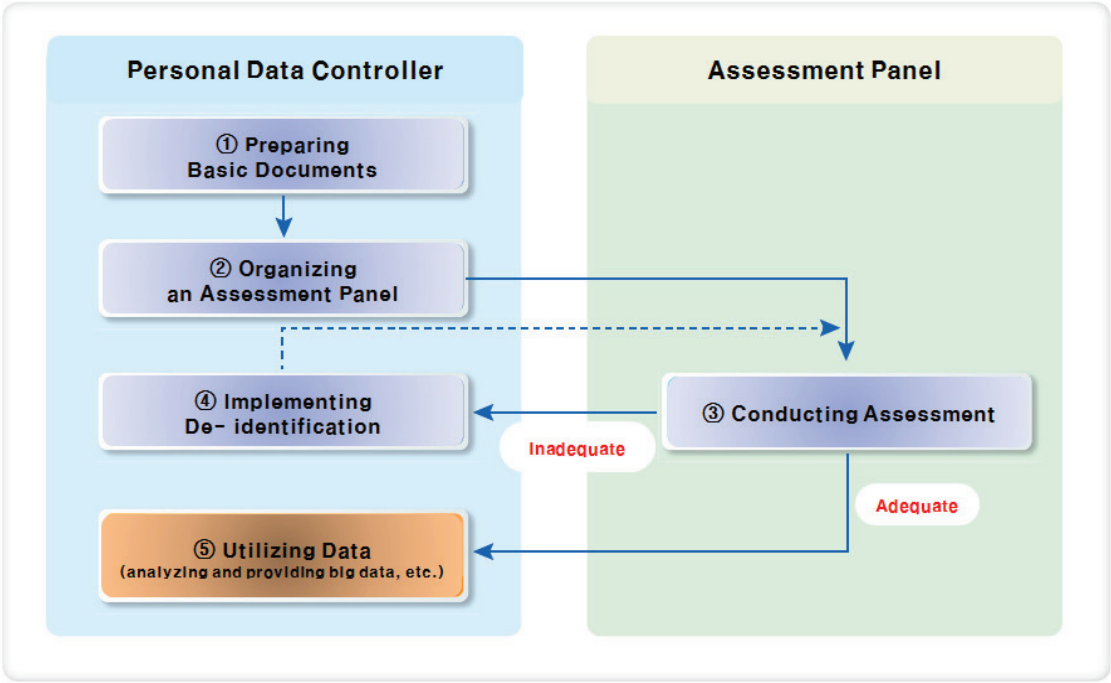
② **(Organizing an Assessment Panel)** A Privacy Officer shall organize the Assessment Panel comprised of more than three people (external professionals should be more than a majority).

③ **(Conducting Assessment)** The Assessment Panel shall assess adequacy of the level of de-identification measures using the basic documents prepared by the personal data controller and the k-anonymity model

④ **(Implementing Additional De-identification)** The personal data controller shall implement additional de-identification measures based on feedback from the Assessment Panel if the assessment result is inadequate.

⑤ **(Utilizing Data)** Data can be used or provided for big data analysis if the assessment of de-identification is adequate.

<Procedure for Assessing Adequacy of De-identification >



① Preparing Basic Documents

○ A personal data controller shall prepare basic documents needed for adequacy assessment such as data statement of an assessment subject, status of de-identification and the level of management of data user.

< Basic Documents Needed for Adequacy Assessment of De-identification >

Type	Basic Documents	Remarks
Data Statement	○ Data characteristics (size, generation and management environment, etc.), detailed specification, and an example of original data	Mandatory
	○ De-identified dataset, detailed specification	Mandatory
De-identification Status	○ Methods and detail techniques applied to de-identification	Mandatory
	○ k-Anonymity value of assessed datasets	Mandatory
Management Level of User Organization	○ Matters related to a purpose and method of utilization, period of use, and list of users who can access the de-identified data in organization	Mandatory
	○ Status of a series of measures for receiving and protecting data when de-identified data is provided	Mandatory
	○ A copy of contract or agreement related to data utilization and provision ※ Indicate reasons if there is no copy of contract	Mandatory
	○ Matters related to data (detailed contents) about personal data that is owned or can be owned by the organization that utilizes data	Optional
	○ A copy of certified certificate related to personal data protection or data security of the organization that utilizes data	Optional

② Organizing an Assessment Panel

○ Privacy Officer shall form the Assessment Panel comprised of more than three experts (more than a majority should be external professionals)

* Appoint more than one legal and de-identification expert from an expert pool that is operated by specialized agencies of each field when commissioning external professionals.

○ The Assessment Panel is comprised of members who do not have direct interests in the purpose of data utilization.

○ A head of the Assessment Panel is elected among the external professionals and shall be responsible for overall issues related to operation of the Assessment Panel.

○ Personal data controller shall host the Assessment Panel meeting at least twice including a kick-off meeting while adequacy assessment.

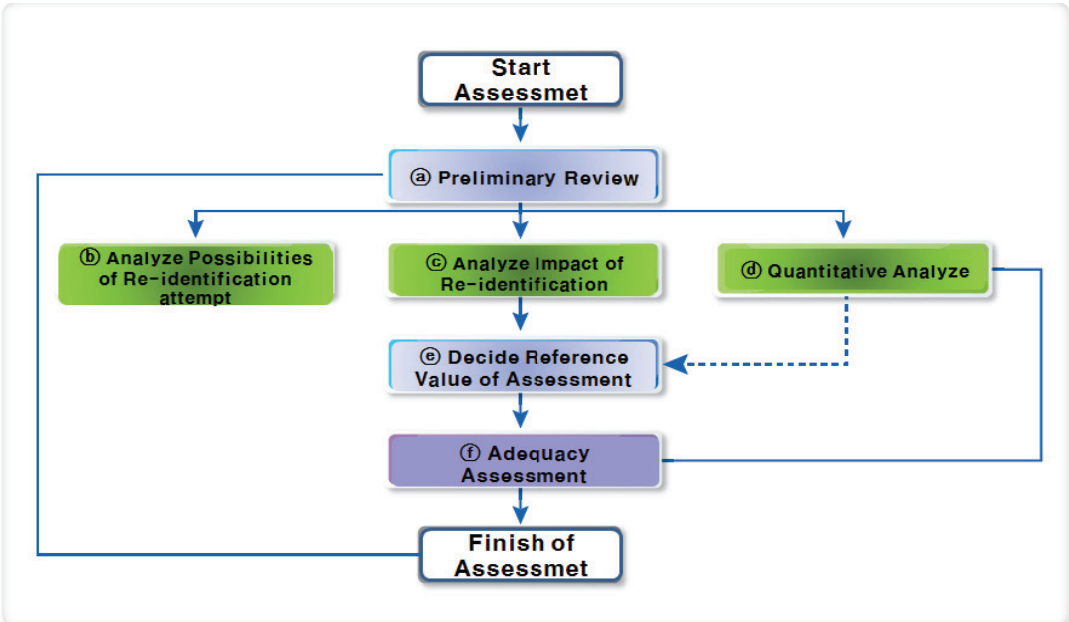
○ If needed, the Assessment Panel shall recommend establishing de-identification rules and supplementary measures.

- If necessary, suggest a re-assessment schedule on a yearly or three-year basis based on development of a computing environment, characteristics of data subject to utilization, and possibility of collecting combinable data.

③ Adequacy Assessment

○ The Assessment Panel assesses adequacy of de-identification based on basic documents and using the k-Anonymity model.

<Adequacy Assessment Execution Process>



※ Procedures were established based on the guidelines for Managing Disclosure of De-Identified Health Information (October 2010, Canada).

ⓐ **(Preliminary Review)** Review prepared basic documents and interview to check out whether personal identification elements are in dataset, whether the purpose of utilization and de-identification methods are appropriate.

ⓑ **(Analysis Possibility of Re-identification Attempt)** Analyze possibilities of re-identification attempt including intention, level of personal data protection and capability of the personal data controller who uses or receives the data.

ⓒ **(Analysis Re-identification Impact)** Assess possible impact on a data subject when data is intentionally or unintentionally re-identified.

ⓓ **(Quantitative Analysis)** Verify the accuracy of a k value provided by a personal data controller.

ⓔ **(Determination of Reference Value)** The Assessment Panel determines the assessment reference value (k-anonymity value) by taking into consideration possibilities of re-identification, impact of re-identification, results of quantitative analysis and purpose of data utilization comprehensively.

> **<Example: Safety Criteria of the Privacy Protection Technical Support Center of the US Department of Education>**
>
> ▷ "k=3" is the minimum level that guarantees safety.
>
> ▷ "5≦k≦10" means relatively high safety.
>
> ※ The k-Anonymity value is the recommended standard to legally allow providing of data.

ⓕ **(Adequacy Assessment)** Decide adequacy of de-identification by comparing calculated values resulted from the average reference value and quantitative analysis.

> ▷ Adequate if a k value from the result of quantitative analysis is 4 and an average reference value is 3
>
> ▷ Inadequate if a k value from the result of quantitative analysis is 4 and an average reference value is 6

- Adequate ⇨ Data can be utilized and provided.
- Inadequate ⇨ Additional de-identification measures and re-assessment are required.

④ Additional De-identification Measures

○ Personal data controller shall implement additional de-identification measures based on feedback from the Assessment Panel if the assessment result is inadequate.

○ The Assessment Panel shall proceed to re-assessment once personal data controller completes implementation of additional de-identification.

⑤ Data Utilization

○ Utilize the de-identified data in big data analysis or allow providing it to a third party if de-identification are assessed (reassessed) to be adequate.

○ In principle, providing or disclosing data to public or uncontracted data user is prohibited as there can be a high risk of re-identification.

※ Except in cases where data is provided pursuant to relevant laws such as the Act on Promotion of the Provision and Use of Public Data.

○ Immediately destroy data once the purpose of using data is fulfilled or it is no longer needed.

○ The follow-up management steps below should be observed in the process of utilizing data for effective usage in a form of de-identified data.

2-④ Follow-up Management

1 Security Measures for De-identified data

○ Safety measures must be implemented as there is a possibility of re-identification if de-identified data is leaked and combined with other data.

- **(Managerial Safety Measures)** Need to designate a person in charge of de-identified data files, prohibit sharing of data on de-identification, and destroy data once a purpose of usage is achieved.

- **(Technical Safety Measures)** Need to restrict access to de-identified data files, manage access records, and install and operate security programs.

< Managerial and Technical Safety Measures for De-identified data >

Type	Measures for De-identified data Protection
Managerial Safety Measures	① Designate a person in charge of managing de-identified data files ② Manage a logbook of de-identified data files ③ Prohibit sharing of data on de-identification measures between department (organization) managing original data and department (organization) managing de-identified data ④ Immediately destroy data if its purpose of usage is achieved ⑤ Establish a counteract plan to be ready for possible leakage of de-identified data
Technical Safety Measures	⑥ Manage access control and restrict access to de-identified data files ⑦ Manage records of access to system storing the de-identified data ⑧ Install and operate security programs to prevent malicious codes, etc.

○ Protective measures when de-identified data is leaked

- Analyze cause of leakage, and implement both managerial and technical safety measures to prevent additional leakage
- Withdraw and destroy leaked de-identified data

2 Monitoring Possibilities of Re-identification

○ Personal data controller which intend to use de-identified data or provide it to a third party shall regularly monitor possibilities of re-identification.

○ Personal data controller shall take additional de-identification measures if monitoring results match with any of the below circumstances.

Type	< Example > Monitoring Items of Possibilities of Re-identification
Changes in Internal Factors	○ When additional data that is collected or given which has possibilities of re-identification if linked with de-identified data
	○ When new data is generated by combining the generated data in the process of using data with de-identified data
	○ When a department using de-identified data requests that the de-identification level be lowered from the original level
	○ When newly or additionally established system cause critical changes to security system which manages and controls access to de-identified data
Changes in External Factors	○ When there is known that an instance of de-identification measures has been re-identified in a manner similar to that applied to the data in use
	○ When new technologies are introduced or disclosed that disable de-identification techniques and techniques applied to data in use
	○ When new data that can be combined with data in use is known to be generated or disclosed

○ Personal data controller who provided or entrusted de-identification data shall immediately notify a personal data controller who processes the data when possibilities of re-identification is found, and take necessary measures including request to stop processing of data, withdraw and destroy the data.

3 Requirements for Contract of Providing or Entrusting De-identified Data

○ Re-identification risk management shall be included in a contract when providing or entrusting de-identified data to a third party.

- (**Prohibition of Re-identification**) Stipulate that personal data controller which are given or consigned to process de-identification data shall be prohibited to re-identify the data by combining it with other data.

- (**Restriction on Re-provision or Re-entrustment**) Stipulate scopes on re-provision or re-entrustment in a contract when providing or entrusting processing of de-identified data.

- (**Notification on Risks of Re-identification**) Stipulate obligation to stop data processing and to inform the de-identified data provider or entrustor when the data is re-identified or the re-identification possibility becomes high.

< Example of Reflecting Special Terms in a Contract >

Article 00 (Prohibition of Re-identification)

① A shall safely use de-identified data provided by B for a purpose of XX, and shall not take any measures to re-identify specific individuals by using such de-identified data.

② A shall obtain a prior consent of B when providing or entrusting data provided by B to a third party, and A shall take necessary measures to prevent re-identification.

③ A shall immediately stop processing the data and notify B, a data provider, and cooperate if given data is re-identified or has high possibilities to be re-identified.

④ A shall take criminal and civil liability for all results of not complying with the preceding Paragraphs ① to ③.

※ A: a company that received de-identification data
B: a company that provided de-identified data

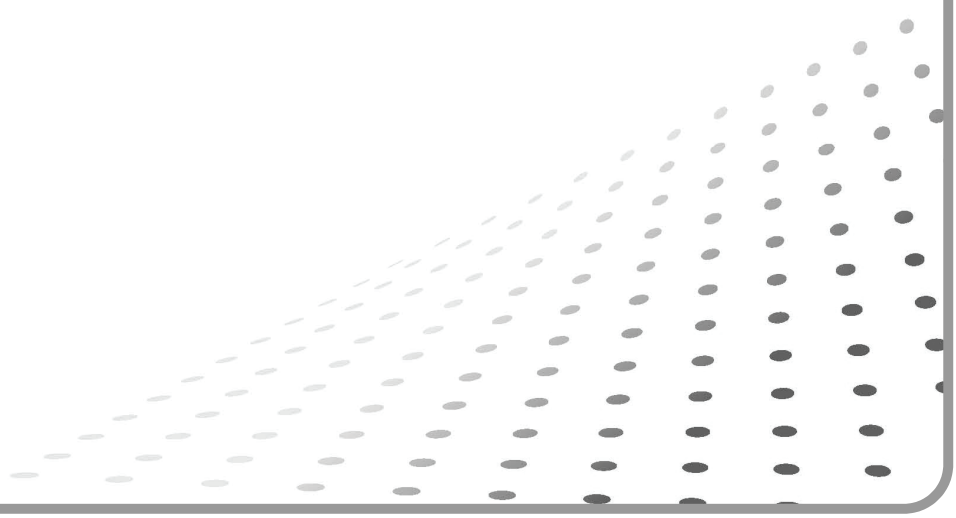
4 Countermeasures to Re-identification

○ If de-identified data is re-identified, processing of the data should be stopped, and necessary measures should be taken to prevent leakage of personal data.

○ Re-identified data shall be immediately destroyed. To use the re-identified data, whole de-identification process to the data should be taken again.

Ⅲ. Support and Management System

1. Support De-identification
2. Supporting the combination of datasets between Personal Data Controllers through Specialized Agencies
3. Legal sanctions against re-identification



Ⅲ Support and Management System

1 Support De-identification

☐ Needs for Support

○ There is a need to set a support system to safely utilize personal data through de-identification.

- Support adequacy assessment of personal data controller.

○ Support Small and Medium Sized enterprises(SMEs) and start-ups to utilize big data analysis by providing consultations and professional training needed for de-identification.

○ Actively counteract to risks of re-identification caused by emergence of convergence technologies such as the Artificial Intelligence.

☐ Support System and Its Details

◈ Specialized Agency in Each Sector

○ A specialized agency in each sector is designated and operated under the supervision of relevant government ministries.

※ Government ministries, in a form of official letter, designate and announce specialized agencies for each sector such as the Korea Internet & Security Agency, Korea Credit Information Services, the Financial Security Institute, the Social Security Information Service, and the National Information Society Agency. If necessary, the government ministries may additionally designate specialized agencies.

○ Roles of Specialized Agencies in Each Sector

- Organize and operate a pool of experts (de-identification experts, legal experts, etc.) to support making up the Assessment Panel at personal data controller.

- Recommend implementation of essential de-identification (k value of k-anonymity, etc.) tailored to each industry.

* Industries such as medicine, welfare, education, finance and credit, communications, distribution, and public.

- Check adequacy of de-identification measures of personal data controller.

◈ De-identification Support Center

○ Established the De-identification Support Center at the Korea Internet & Security Agency (KISA), a specialized agency for personal data protection.

○ Roles of the De-identification Support Center

- Draw operation guidelines for specialized agencies of each sector and monitor conformity with the guidelines.
- Operate working committees for specialized agencies in each sector.
- Manage and train experts who work as a member of the Assessment Panel in each sector, and provide consultations and training to SMEs and startups.
- Support updates and utilization of the guidelines.
- Research relevant policy and technology trends from home and abroad.

2 Support the Combination of Datasets between Personal Data Controllers through Specialized Agencies

☐ Need for Support

○ Datasets owned by different personal data controllers are needed to be combined for big data analysis. To do this, a key to create a dataset by matching different datasets is required.

- When creating the key, using direct identifiers that can identify an individual could possibly lead to violation of current laws. (Article 18 of the personal data Protection Act, Restrictions on the Use and Provision of personal data for any purposes other than the intended ones)

○ Therefore, in order to combine and analyze datasets, it is necessary to use 'Temporary Linking Key', which temporarily serves as a matching key only in the combining process.

○ Even if the combination using the temporary linking key is permitted, a support and management system is required to prevent the possibility of personal information infringement through indiscriminate combination.

☐ Support and Management System

○ Specialized agencies in each sector shall support combination of datasets between personal data controllers.

○ Criteria for Selecting Specialized Agencies

- Combining datasets between corporations in the same industry shall be supported by specialized agencies in the same sector.

- Combining datasets between personal data controllers in the different industry shall be supported by specialized agencies in their respective sectors.

- The Korea Internet & Security Agency (KISA) and the National Information Society Agency (NIA) shall support the industry where there is no specialized agency.

○ Key Roles and Responsibilities of Specialized Agencies

- Support combination of datasets between personal data controllers through using temporary linking key provided by them.
- Prohibited any attempts to identify individuals from the given datasets.
- Immediately destroy all datasets once combining datasets and providing data are completed

○ Each ministry shall draw and implement detailed guidelines regarding specialized agencies under their supervision.

□ Procedure and Requirements for Dataset Combination

◈ Combination Procedures

① Each personal data controller shall make temporary linking key(which is made from identifiers) by applying the same algorithm with other personal data controller, and carry out de-identification measures and adequacy assessment on the datasets.

※ Noise should be added to a temporary linking key when creating it, and when a temporary linking key is made from more than 2 identifiers it shall take necessary measures to prevent identifying a specific individual from illegal decryption by combining some identifiers or from combination with original data.

② Personal data controllers provide de-identified datasets to specialized agency and request its combination.

※ In this case, the specialized agencies can't identify a specific individual from the provided de-identified datasets.

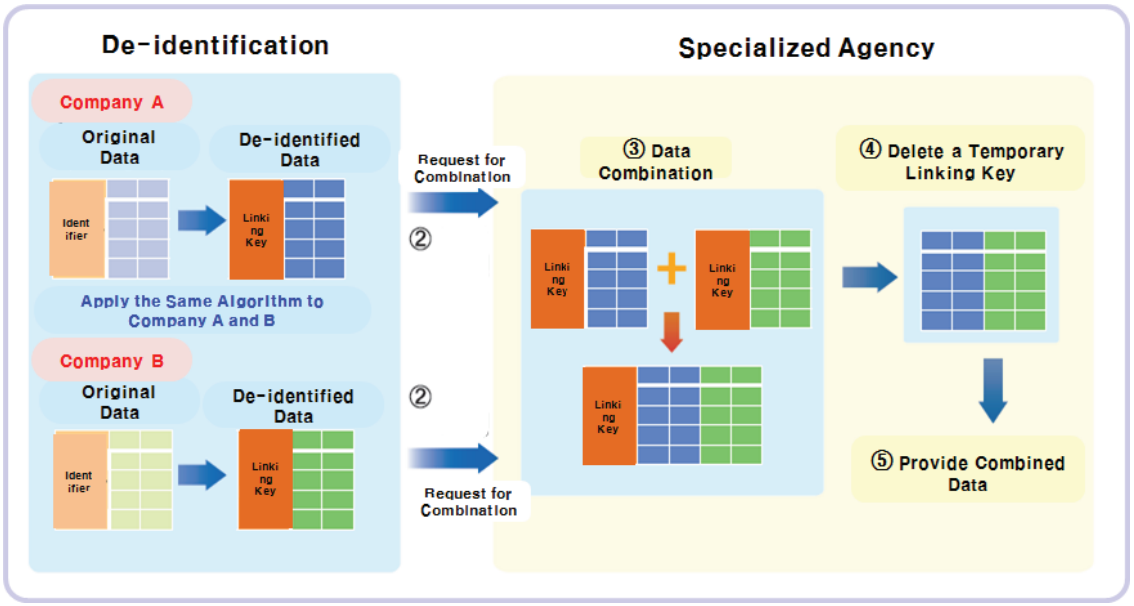
③ Specialized agency combines datasets using the temporary linking key.

④ And then, shall destroy the temporary linking key.

⑤ Specialized agency provides the combined dataset to personal data controller who requested dataset combination. (The specialized agencies shall destroy the combined dataset after providing it.)

※ personal data controllers are very difficult to re-identify a specific person from the combined dataset since there is no temporary linking key in the dataset.

<Procedures for Combining data Datasets between Personal Data Controllers through Temporary Linking Key>



◈ Matters to Consider When Combining Datasets

○ Personal data controller A and B shall not share algorithm information about temporary linking key with specialized agencies of each sector.

○ Using resident registration numbers is prohibited when creating a temporary linking key. (Article 24-2 of the personal data Protection Act, Restriction on the Management of Resident Registration Numbers)

○ A value of k-anonymity shall be calculated without temporary linking key.

* Inclusion of the temporary linking key will make the value of k-anonymity equal to 1, which leads result of adequacy assessment to ‘inadequate’.

○ Specialized agencies shall immediately destroy de-identified data when re-identification is detected in the process of combination.

○ A data user that receives the combined dataset shall conduct appropriate adequacy assessment before using the it.

3 Legal Sanctions against Re-identification

☐ Criminal Punishments

○ Where re-identified data is used or provided to a third party

(Example 1)
Where rendered personal data, as a result of sharing tips on de-identification techniques to researchers, is provided for a purpose other than the intented one while providing de-identified data.

(Example 2)
Where critical identifiable elements (name, date of birth, phone number, etc.) are provided to a third party in a form of easy re-identification like encrypting with open algorithms.

(Example 3)
Where de-identified data is preserved on purpose and used for various purposes like one-on-one marketing.

- Applicable to the use and provision of personal data for purposes other than the intended use (Violation of the Article 18.1 of the Personal Data Protection Act, Articles 24 and 24.2 of the Act on the Promotion of data and Communications Network Utilization and data Protection, Etc. (“Network Act”), and Articles 32 and 33 of the Credit data Use and Protection Act (“Credit data Act”).

- Subject to imprisonment not exceeding five years or fine not exceeding 50 million won(about 43,000$).

※ Service providers subject to the Network Act may be imposed of additional penalty surcharge less of 3% of their sales related to a violation.

☐ Administrative Measures

○ Where de-identified data is re-identified, but still in preservation without immediate destruction.

(Example)

Those who received de-identified data have re-identified a particular individual while doing their jobs, but still have the re-identified data instead of destroying it immediately.

- Applicable to collection of personal data without the consent of a data subject (Violation of the Article 15.1 of the Personal Data Protection Act, and Article 15.2 of the Credit data Act)

- Subject to a administrative fine not exceeding 50 million won(about 43,000$).

※ According to Article 22.1 of the Network Act, service providers subject to this Act may be imprisoned up to five years or imposed of fine not exceeding 50 million won(about 43,000$). They also may pay additional penalty surcharge less than 3% of sales related a violation.

Appendix Definitions of terms used in the guidelines

Selected terms used in the guidelines are defined below. However. some definition are sourced to such as Personal Data Protection Act.

◉ **Adequacy assessment** : a series of procedures that whether a de-identification of those who apply to assess is adequate to use of data or not by Adequacy Assessment Panel.

◉ **Adequacy assessment panel** : a group of experts in the area of de-identification and personal data acts to determine that a de-identification subject to assess is adequate or not, comprising of more than three people (external professionals should be more than a majority).

◉ **Attribute value** : characteristic data like sensitive values related to an individual and may identify a particular person if easily combined with other data.

◉ **Data subject** : the natural person who is identifiable by data processed hereby to become the subject of such data.

◉ **De-identification** : Measures to render data anonymous in such a way that an individual is no longer identifiable by fully or partially eliminating, replacing, deleting, or other means of identifiable elements from a datasets.

◉ **De-identified data** : a dataset that has been appropriately de-identified in accordance with the this guidelines.

◉ **Identifier :** data such as a value or a name uniquely assigned to an individual or a thing related to the individual.

◉ **Government 3.0** : a new Korea governance paradigm which the central government and municipalities put forward in 2013 aimed at creating new markets and jobs by making government data available to public.

◉ **Personal data** : any information which relates to a living natural person who can be identified or identifiable from those data including name, resident registration number and image, etc. (including the information that does not, on its own, permit direct identification of a specific individual, but that does identify specific individual when it is easily combined with other information.).

◉ **Personal data controller** : a natural or legal person, public institution, organization, etc. that directly or indirectly processes personal data to operate personal data files as part of tasks,

- **"public institution"** means the institution stated in the following Items; and
  a. The administrative bodies of the National Assembly, the Court, the Constitutional Court and the National Election Commission, the central administrative departments (including those agencies the President and the Prime Minister) and their agencies, and local governments; and
  b. Other national institutions and public entities which are provided by the Presidential Decree.

◉ P**ersonal data manager :** a person such as executive officers, employees, temporary agency workers, part-time workers, etc. who manage personal data under personal data controller' instructions and supervision.

◉ P**rivacy officer** : a person designated by personal data controller to take overall responsibility for the management of personal data.

◉ **Re-identification** : general terms for any process of identifying a particular individual from a de-identified data.

◉ **Requester** : natural or legal person, public authority, agency, organization or any other body to whom data are provided or disclosed.

◉ **User organization** : natural or legal person, public authority, agency, organization or any other body who intends to utilize the de-identified data.

◉ **Data user** : same as 'User organization'

◉ **Data provider** : natural or legal person, public authority, agency, organization or any other body who provides the de-identified data.

◉ **Temporary linking key** : a key created with the same algorithm by personal data controller that intend to combine their dataset each other. In this case, user organization should take necessary measures to protect re-identification.

◉ **User organization** : an organization that intend to utilize de-identified data after taking de-identification.