Main Contents of Guidelines for Pseudonymizing Unstructured Data

개인정보보호위원회
Personal Information Protection Commission

Main Contents of Gidelines for Pseudonymizing Unstructured Data

1

Background

■ With advances in AI technology and computing resources, demand for data use has shifted from the traditional structured data(figures) to unstructured data (images, videos, audio recordings, and text documents).

* Unstructured data such as images, videos, audio recordings, and texts account for up to 90% of the world's data. (IDC, 2023)

Reference. Cases of pseudonymization and use of unstructured data

- (Images/videos) For research and development of an AI model that can diagnose(assist) certain diseases, MRI/CT/X-ray images/videos taken by hospitals were pseudonymized and used to train the model.
- (Images/videos) To develop a smart CCTV that can detect and alert illegally placed banners, CCTV recordings of public places were provided by local governments which were pseudonymized before being used for an AI model development
- (Audio recordings/text documents) To develop an AI voice generator that can provide consultation/ response to customer complaints, audio recordings of calls and consultation records were pseudonymized before being used to train the AI model.

■ However, the existing guideline was based on structured data, failing to accommodate unique attributes of unstructured data and technological advances in AI.

Reference. Difference between structured and unstructured data

Structured Data	Unstructured Data
(Definition) Data that exist in a structured format based on a fixed set of rules *e.g. tables composed of rows and columns*	(Definition) Data that exist without a structed format based on a fixed set of rules *e.g. photos, videos, voice calls, transcripts, reports, e-mail content, etc.*
(Attributes) Data processing method such as computing & analyzing as well as data pseudonymization technology/method is relatively simple	(Attributes) Data processing method and pseudonymization technology/method are complex and diverse depending on the purpose and environment of a research

➡ With AIs being normalized in this increasingly digitalized era, establishing a guideline for pseudonymizing unstructured data can reduce uncertainty for businesses and researchers and also support technological innovation.

2

Characteristics and considerations for Pseudonymization Process/Usage of Unstructured Data

■ (Difficulty in determining identifiers) Whether a piece of information is an identifier is relative to context and can be determined differently depending on the purpose or environment of processing.

e.g.

- A single scan of Cranial CT has a low risk of identifying an individual but an accumulation of multiple CT scans taken from different angles make it possible to reconstruct a person's facial structure, increasing the identification risk.
- CCTV clip taken from a distance that is difficult to distinguisha person's facial features usually has a low risk of identifying an individual. However, if the person has a unique feature such as a significant scar, tattoo, or hairstyle, the risk of identifying that individual becomes high.

■ (Limitations of Pseudonymizing technology) The technology to flawlessly detect and process any and all information that can identify an individual does not exist.

e.g.

- With regards to image/video data, there are cases where the faces could not be detected depending on resolution, angle of light, and how big they appear →Recent AI technology has a detection accuracy of 90−98%[1)]
- Text that says "I live in a blue building in front of exit 1 of Gangnam station, first floor" is not perceived as an address and left untouched or a brand name such as "Kim GangnamGimbap" is identified as a person's name and unnecessarily removed.

1) Kaur, J., Singh, W., "Tools, Techniques, datasets and application areas for object detection in an image: a review", Multimed Tools Appl 81 (2022)

■ (Risk of Re−identification Attack) With advances in AI and data restoration technology, the risk of re−identification has increased even without connecting/ combining the data with another set of information.

e.g.

- There is technology[2] that enables restoration of a speaker's original voice that does not require knowledge of the voice change algorithm
- Technology is being researched[3] so that pictures mosaiced with AI can be reversed close to its original version regardless of the mosaic pattern used for pseudonymization.

➡ When pseudonymizing/using unstructured data, the**identification risk should be lowered by considering the data processing context, limitations of pseudonymizing technology, re−identification risk, etc.**

2) Deng, Jiangyi et al, "Catch You and I Can: Revealing Source Voiceprint Against Voice Conversion", ArXivabs /2302.12434 (2023)

3) R. Dahl, M. Norouzi and J. Shlens, "Pixel Recursive Super Resolution", 2017 IEEE International Conference on Computer Vision (ICCV) (2017)

3

Basic Principles of Pseudonymization of Unstructured Data

1) Purpose, environment, and sensitivity of data should be considered comprehensively when determining identification risks and deciding on a suitable processing method/level.

- Minimizing damage to raw data to suit research purposes and at the same time applying a variety of safety measures including management/environmental control can be done.
  - Retaining data elements that are essential to achieving the purpose of a research while upscaling the level of pseudonymization for the rest of the information or utilizing sufficient safety measure including restriction tointroducingexternal information or software.

2) In order to address theconstraints of pseudonymizations technology, associated risks should be reviewed at the early stages (while planning for the research and technological development) and take safety measures accordingly.

- In order to complement shortcomings of pseudonymization technology, the following measures are advised.

① Document &retain evidence supporting adequacy/reliability of the pseudonymization technology

② After pseudonymization technology is applied, conduct internal inspection on the results

③ When conducting review of pseudonymization adequacy, include①&② as part of that assessment(It is advisable that half or more of the persons involved in the review be external experts)

- To prevent identified personal data breach risks in advance, it is necessary to enhance internal control of all the organizations involved in the use of pseudonymized information.
- Once the purpose of pseudonymization is achieved, pseudonymized information is to be promptly disposed in order to minimize risk after the fact.

3) In response to advancement in data restoration technology, control measures such as access control to the relevant system/software should be established with regard to using pseudonymized unstructured data.

* Separate storage of data that can be used to restore original data, restriction of access to restoration software, etc.

- Since it is impossible to flawlessly remove any and all risks than can appear in the course of AI development/usage, there should be constant monitoring on possible violation of rights/interests of the data subjects.

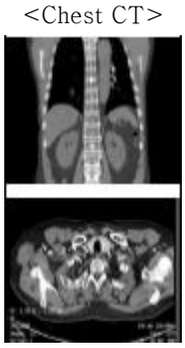
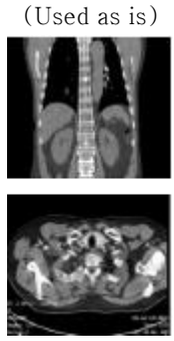
▶ Pseudonymization Scenarios of Major Unstructured Data

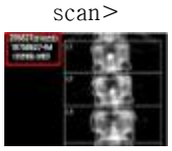
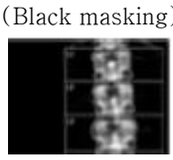
※ Pseudonymization scenarios provided below are based on actual cases where unstructured data was pseudonymized for use. They have been edited through discussionsamong businesses /organizations/experts andare purely for reference purposes. Method/level of pseudonymization can be applied freely depending on the field/situation per decision of the processors and review committees.

[Case1] Developing an AI model to diagnose breast cancer/decreased bone density

A case where CT scans (video/images) and pathology record (text) of breast cancer patients accumulated by a university hospital was used for a research to develop an AI that can diagnose breast cancer and reduced bone density

☞ Data processing environment safely controlled to remove identification risk and took measures such as banning restoration software which made it possible to use CT scans as is

<Chest CT> 	Review of identification risks	– Chest CT alone has **little to no risk of identification** – The research **used 200 CT scans per person** which could be used to reconstruct bodily figure in 3D through restoration technology, and for a handful of patients who have unique figure or scars **could be identified** albeit a low risk – **Closed research and analysis environment based on a cloud system was used** and unauthorized data/programs were strictly forbidden in the facility thereby making it **impossible to apply 3D reconstruction technology** * Data was stored in the cloud server which could be accessed from an analysis center where only authorized personnel could connect.	(Used as is) 
	Data processing method	➤ Although there is risk of identification through 3D restoration, the environmental control made it unlikely, and so **the scans were used as is without pseudonymization**	

<table>
<tr><td colspan="4">☞ Meta data within the image that can be used as an identifier removed before use</td></tr>
<tr><td rowspan="2"><Patient info within a CT scan>
  </td><td>Review of identification risks</td><td>– There is identification risk in the event patient information marked within the images is combined with additional information and analyzed.
* DICOM header information (patient no., date of birth, gender) marking
– the marked information was not necessary for the research</td><td rowspan="2">(Black masking)
  </td></tr>
<tr><td>Data processing method</td><td>➤ Patient information was removed through black masking method.</td></tr>
<tr><td colspan="4">☞ Unstructured data was converted to structured data format</td></tr>
<tr><td rowspan="2"><Text on a pathology record>
  </td><td>Review of identification risks</td><td>– Pathology record contained identifiers in raw form which was unnecessary for the research and had a high risk of identifying individuals</td><td rowspan="2">(Used after conversion to structured data)
  </td></tr>
<tr><td>Data processing method</td><td>➤ Data was converted to structured data through natural language processing technology before being used and any information with the risk of being an identifier was pseudonymized
➤ Because neither natural language processing technology nor pseudonymization technology are 100% accurate, additional inspection was conducted for the entiredatasetafter the conversion.</td></tr>
</table>

[Case2] Developing an AI model to diagnose dental caries

A case where pictures taken during dental check-ups(images) at a university hospital were provided to a company after pseudonymization to develop an AI model that can diagnose dental caries such as cavities and periodontitis.

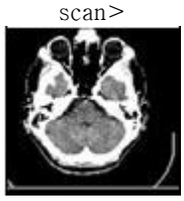
☞ Portions unnecessary for the research was blurred and metadata removed before use

<dental picture> 	Review of identification risks	– Dental pictures alone pose almost no risk of being identifiers – Only the cavity portion is required for the research – Metadata associated to the pictures (name, age, etc.) carry a risk of being identifiers	(Cavity: used as is Non-cavity: blurred) 
	Data processing method	➤ Cavity portion of was used as is while blurring the rest of the pictures – The level of blurring was set in consideration of current restoration technology capabilities and data processing environment (access to other information/restoration technology) ➤ Meta data was not required for the research and therefore removed	

[Case3] Developing an AI model to diagnose facial fracture

A case where facial CT scans were pseudonymized for a joint project between a university hospital and a private company to develop an AI model that can diagnose facial bone fractures

☞ Portions unnecessary for the research was masked to lower the risk of identification

<Patient info within a CT scan> 	Review of identification risks	– CT scans alone pose almost no risk of being identifiers. – Three-dimensionalreconstruction is a possibility and for a very small number of patients who have unique features or a well-known face can be exposed to a risk of being identified albeit low. – Although the research used massive videos/images making three-dimensional reconstruction possible, the risk of 3D reconstruction attacks can be lowered by masking some of the edges – The occipital area (the back of the head) was not required for the research	(Facial area: used as is Occipital: masked) 
	Data processing method	➤ The facial area needed for the research was used as is while unnecessary occipital area was masked to reduced the risk of 3D reconstruction	

[Case4] Developing an AI model that can determine abnormal situation while self−driving

A case where car driving videos accumulated by a research institute was pseudonymized before providing them to a company to develop an AI model that can perceive an abnormal situation while self−driving.

＊A pedestrian unexpectedly entering a roadway, another vehicle suddenly cutting in, jaywalking, etc.

☞Portions unnecessary for the research was masked for use

<faces/license plates> 	Review of identification risks	− Risk of identification when a person's face is clearly visible or a vehicle's license plate is exposed which can lead to inferring who the passenger is. − The research only requires detection of overall figure and movement of people/vehicles, making it immaterial to mask faces/license plates	(Masked) 
	Data processing method	➤ Faces/license plates were masked to the extent they were unidentifiable to a computer before use	

[Case5] Developing an AI model to monitor high−occupancy lanes on highways

A case where CCTV snapshots of highway traffic collected by a local government was pseudonymized before being provided to a company to develop an AI that can monitor cars that have violated the high−occupancy lane requirement

＊When a vehicle that has an occupancy of less than three passengers used the lane

☞Portions unnecessary for the research was blurredbefore use

<Picture of a passing vehicle>  <Unique vehicle>  	Review of identification risks	− Risk of identification when a passenger's face is clearly photographed or a vehicle has a−typical features − The research does not require identification/distinction but only need to tell (1) whether a figure is a person (2) How many passengers are on board − Enable the AI to determine whether a figure is a passenger but blur the image to avoid identification	(①location /scope of passengers identified) 
	Data processing method	➤ Images of vehicles with a−typical features removed from the database. ➤ Data pseudonymized by differentiated blurring levels(level 1−10) so as to achieve AI accuracy while avoiding the risk of identification. ＊During adequacy review, the likelihood of achieving the intended purpose and the risk of identification are examined	(②Blurred) 

[Case6] Developing an AI model that can chat in Korean

A case where a company specializing in AI chatbot pseudonymized day-to-day texts they collected through an app to develop an AI chatbot that can chat in Korean

☞ Data that can lead to identification of an individual was strictly filtered and removed, and metadata was deleted			
<Dialogue text files> 	Review of identification risks	- In day-to-day conversation data (text messages), there is a lot of information with high risk of identification.	(metadata removed, identification information filtered/removed) 
	Data processing method	➤ Meta data (e.g. user ID) removed and replaced with random ID to eliminate association with a specific individual ➤ Information that has a risk of being an identifier were strictly filtered and pseudonymized (*replaced, removed) * e-mail addresses were replaced with the word"e-mail"	

☞ Measures taken so that pseudonymized information used to train the AI does not appear in AI responses			
<Risk of chatbot response> 	Review of identification risks	- In the event pseudonymized data used to train language models is uttered as AI chatbot response, there is ahigh risk of personal identification	(training DB and response DB separated) 
	Data processing method	➤ Measures were taken to separate the database for language training and for response so that sentences used to train the AI is not exposed. * Sufficientlyinspect the response database to make sure there is no information with identification risk	

[Case7] Developing an AI model that can generate hypothetical scenarios for training

A case where a company pseudonymized voice recorded data between employees and customers to develop an AI that can train employees at call centers to talk to customers

☞ Audio recordings were converted to written text (STT, Speech to Text) and pseudonymized before use

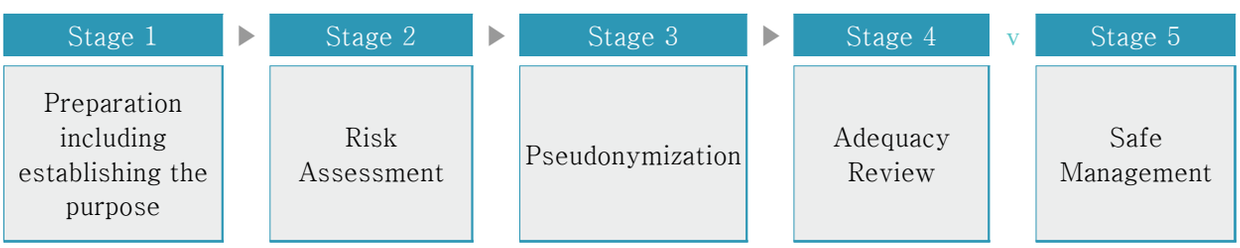
< Audio recordings> 	Review of identification risks	– Recordings of conversation with customers contain actual audio data of customers & employees (voice, tone, intonation, accent, etc.) and unrefined form of personal information. – In developing an AI that can generate hypothetical scenarios, what matters is understanding questions/responses and the flow of conversation depending on the purpose of each call and character of customers. Actual voice recordings are not a necessity.	(①Conversion to text)  (②Identification information replaced/removed) 
	Data processing method	➤ Recordings were converted to text data through STT technology after which information with identification risk was pseudonymized (replaced/removed) before use ➤ Since the accuracy of pseudonymization for text data is less than 100%, additional inspection of the entire data was to be done to remove any information with the risk of identification	

4

Things to Consider at Each Stage of Pseudonymization of Unstructured Data

■ With regards to pseudonymization of unstructured data, the pseudonymization process in chapter 2 of pseudonymization process guidelines are to be followed, but it is also recommended that inspections be conducted to assess identification risksand to consider additional safety measures.

◖ Stages for Pseudonymizing Personal Data ◗



1) Preparation Stage

The stage of setting up and reviewing the purpose of pseudonymization and selecting what data should be subject to the pseudonymization

- Identify pieces of information with the potential for personal identification in unstructured data, clarify the type and scope of information thereby selecting the target for pseudonymization.

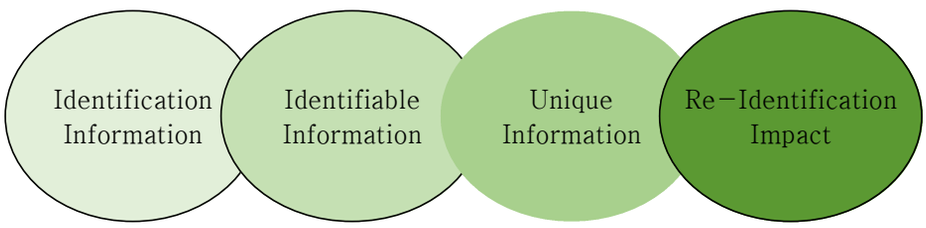
2) Risk Assessment Stage

The stage of assessingpseudonymization target and processing environment in order to determine the method and level of pseudonymization.

- Method and level of pseudonymization is determined by comprehensively reviewing “Identification riskinherent in the data” and “Identification risk in the processing environment”

(1) Identification risk assessment

- Review whether there is information that has a high probability of identifying an individual (identifiers, identifiable information), unique information, or information that can be highly affected by re-identification due to data characteristics.



– (Identification) Unlike structured data, making an absolute distinction between identification information and identifiable information is difficult. As such, determination on whether a piece of information can lead to identification has to be made in consideration of the data processing environment such as the purpose and method of processing.

* (Identification Information) Information directly associated to a specific individual and distinguishes that person from others

** (Identifiable Information) Information that cannot lead to identification on its own but possibility of identification rises when combined with other information subject to pseudonymization

e.g.

- When using CCTV footage that has caught a person's face on camera
  - ▸ (Case where identification possibility varies depending on the situation)
    - – If the location in the footage is evident and the face has been captured clearly, the likelihood of identification is high.
    - – If the location in the footage is unclear and the face has been captured very small or the resolution low, making it difficult to distinguish a particular individual, likelihood of identification is low.
  - ▸ (Case where identification varies depending on the purpose/method of processing)
    - – If a footage with a person's face is combined with other information such as physical features, gait, and direction of movement to be used for the purpose of analyzing the person's gender and pattern of behavior, likelihood of identification is high.
    - – If a footage with a person's face is used only to determine whether a subject is a person and used without combining facial information with other information, the likelihood of identification is relatively low.
- If a patient's MRI scan includes meta data and the meta data has identification information such as the patient's name and patient number, likelihood of identification is high

– (Unique Information) Assessment of identification risk for physical/external/behavioral features unique to a person or uniqueness of anentity/object associated to an individual

e.g.

- Case where there is possibility of identifying an individual due to unique physical/external features
  – (Image/video) If physical feature, body figure, hair style, tattoo, etc. is unique
  – (Audio Recording) If pronunciation(such as palatal stop) or tone (voice) is unique

- Case where there is possibility of identification due to unique behavioral features
  – (Image/video) If there is uniqueness in gait, gesture or behavior.
  – (Voice) If there is uniqueness in intonation (dialect, accent), repeated use of a word, or linguistic habit.
  – (Text) If there is uniqueness in repetitive use of a word, grammar, writing style, or linguistic habit.

- Case where there is possibility of identification due to a unique entity/object associated to an individual
  – (Image/video) If there is uniqueness in place of residence, type of car (such as rare supercar), clothing, or pets

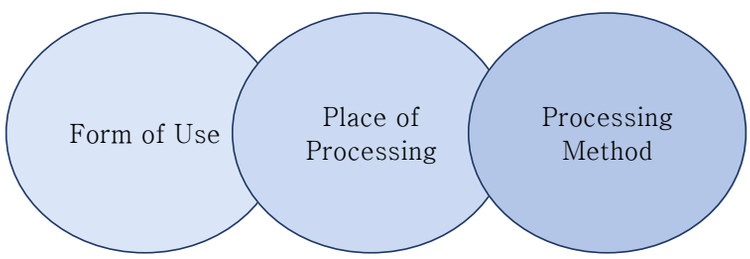
– (Impact in the event of re–identification) Assessment of whether there is any information that can greatly affect an information subject such as having a social repercussion in the event pseudonymized unstructured data is re–identified.

e.g.

- CCTV footage of being a victim to a crime, audio recordings that contain sensitive personal information, etc.

(2) Assessment of identification risk associated to the processing environment

- Assessment required for identification risk depending on the pseudonymization environment such as form of use, place of processing, and processing method of pseudonymization.



– (Form of use) Assessment of possible identification information in consideration of information the controller (or handler) has, accessible/obtainable information, and scope/type of use.

e.g.

- If a dataset of patients' X-ray is to be analyzed by a doctor includes the X-ray of his/her patients, the risk of the doctor deducing a specific patient rises based on his/her treatment information and experience.
- If metadata included in CT scans are removed, the risk of a person with knowledge of one of the patients deducing that individual is lowered.
- Based on how AIs are trained, if text information is absorbed and information with identification risk is generated/produced by such AIs, the risk of exposure to inference attack rises.
- If an AI only identifies and collects the general figure of people in CCTV footage to generate statistics, because there were no analysis on individual features, the risk of identifying a specific individual is low.

– (Place of processing) Assess whether pseudonymized information is processed in an environment or location where access to and acquisition of information other than the pseudonymized information and access to re-identification technology are restricted.

e.g.

- If there is technology that can restore pseudonymized(mosaiced) pictures and the pictures are processed in an environment where the technology can be applied, the risk of identification increases.
- If audio recordings have been modulated to avoid identification and the recordings are processed in an environment where other information cannot be obtained for compare/contract purposes, the risk of identification is lowered.

– (Processing method) If pseudonymized information can be linked/combined with other information or expected to be provided repeatedly, assess whether there are any items that have an increased probability of being identified.

e.g.

- When using 100 CT scan of head and neck taken at various angles for each patient, there is a risk that facial features can be restored in 3D using image restoration technology.
- If a research is conducted using just one CT scan per patient, image restoration technology cannot restore facial features, lowering risk of identification.

3) Pseudonymization Stage

The stage of carrying out actual pseudonymization based on risk assessment and pseudonymization plan for each item

- Distinction is made between items that does and does not require pseudonymization. For items that require pseudonymization, reasonable method and level of pseudonymization is determined in this stage.

  ➜ Of the unstructured data, items that are crucial to achieving the purpose of processing but has low risk of identifying an individual is to be used as is without pseudonymization.

➜ Of the unstructured data, items thatwould not hinder the purpose of processing even if they were pseudonymized but have a high risk of identification is to be pseudonymized before use.

e.g.

- When researching and developing an AI model to diagnose facial bone fracture using CT scans
  - ▸ Front of the head (facial area)
    : Because this is essential to diagnose facial fractures, was used as is.
  - ▸ Back of the brain(back of the head)
    : Because this part is not necessary to diagnose facial fractures, was used after masking.

- When developing an AI model to analyze/diagnose cavity using dental pictures
  - ▸ Dental part that is labeled and suspected of having cavities
    : Because this is essential to analyzing/diagnosing cavity, was used as is.
  - ▸ Healthy teeth and gum that is not suspected of having any cavity
    : Because this part is not required to analyze/diagnose cavities and there is identification risk associated to possibly unique structure or traces of previous treatments, was used after blurring.

- When applying pseudonymization technology for unstructured data, it is advised that relevance/reliability is assed and the basis* of the assessment is written and kept.

e.g.

- If the edges of CT scans were masked for pseudonymization, supporting evidencesuch as effectiveness of the solution, recognition rate/processing accuracy(error rate), etc.

- In order to complement technical limitation of pseudonymization of unstructured data and reduce residual risk, additional and internal inspection against the processed result is necessary.

– It is recommended that appropriate inspection method should be applied in consideration of the pseudonymization purpose, characteristics of the data, features of the applied technology, level of control at the processing environment, etc. all the while keeping record/saving corrective feedback incurred during the inspection process intended to lower risk and also get adequacy review*

e.g.

- Review of whether sufficient effort has been made to lower reasonable and predictable risk through inspections such as visual inspection of the entire data conducted by persons and sampled inspection that is statistically reliable.

4) Adequacy Review Stage

The stage for forming an adequacy assessment committee that includes external experts to review the suitability of the processing purpose, the adequacy of risk assessment results, the adequacy of pseudonymization results, feasibility of achieving the intended purpose, etc.

- Review whether pseudonymization was conducted using reasonable method/level in consideration of characteristics of the unstructured data, purpose and environment of processing, etc.
- Review the relevance and reliability of the technology applied to pseudonymize unstructured data and whether additional inspection has been conducted to sufficiently lower residual risk that may come from limitations of the applied technology.
- It should be noted that when pseudonymizing unstructured data, the following should be considered comprehensively: characteristics of such data, advancement of relevant technology and re-identification risk. Doing so requires expertise,

– and so the recommendation is that half or more of reviewers should be external experts to conduct an objective and professional inspection.

5) Safe Management Stage

The stage of monitoring and managing the possibility of re-identification in the course of utilizing pseudonymized data after the adequacy review

- Those who intend to train or research & develop an AI model based on the special cases for pseudonymized information are required to take sufficient measures to lower various risks that can occur before and after pseudonymization, in consideration of their AI technology and service.
- However, it should be noted that because it is impossible to completely eliminatevarious risks that may arise in the context of AI with existing technology,
  - the level of post-management implementationshould be determined based on the extent of efforts that had been made to minimize the residual risk
- Even after an AI service has been deployed, continued monitoring is required to keep track of whether identification risk and/or privacy violation risk have increased.
  - Also, upon detecting increased identification risk or case of privacy violation, the pseudonymization process should be suspended and the relevant risk removed.

5

Checklist to Assess Risk of Identification in Unstructured Data

■ The purpose of this checklist is to review risk of identification prior to setting up pseudonymization plan, and even if the review result is "Yes," data can be used as long as appropriate measure have been taken to lower the relevant risk.

<table>
<tr><th colspan="2">Category</th><th colspan="2">What to review for identification risks</th></tr>
<tr><td rowspan="5">Data</td><td rowspan="3">Identifiability</td><td colspan="2">Whether there is information that can lead to identification</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>① Is there information that has a high probability of identifying a specific individual in itself?
* (e.g.) △Case where the entire face is clearly visible (frontal face, profile, face before/after plastic surgery, a person's face reflected on a mirror or glass) △Case where the meta data included in an image or video shows highly identifiable information such as the persons' names, patient number, etc.</td><td>☐ Yes
☐ No</td></tr>
<tr><td>② Is there information that has a high probability of identifying a person when two or more pieces of information are combined?
* (e.g.) △Case where a cell phone photo indicates time, location, who took the photo △Case where a chest CT scan indicates the patient's information in the scan or the doctor's note on a patient's unique condition is written</td><td>☐ Yes
☐ No</td></tr>
<tr><td rowspan="3">Unique Information</td><td colspan="2">Whether there is information that can act as an identifier due to a person's unique features</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>③ Is there possibility of identification due to unique physical/external features?
*(e.g.: image/video data) Case where one has unique physical features, body figure, hair style, tattoo (on a specific body part), scar, etc.
*(e.g.: voiceaudio data) Case where one has unique pronunciation (such as palatal stop sounds), tone (voice), etc.</td><td>☐ Yes
☐ No</td></tr>
</table>

<table>
<tr><th colspan="2">Category</th><th colspan="2">What to review for identification risks</th></tr>
<tr><td rowspan="6"></td><td rowspan="3"></td><td>④ Is there possibility of identification due to unique behavioral aspect?

*(e.g.: image/video data) Case where there is uniqueness in gait, gesture or action
*(e.g.: audiodata) Case where there is uniqueness in accent (dialect, speech), repeated use of a word, linguistic habit, etc.
*(e.g.: text data) Case where there is uniqueness in repeated use of a word, grammar, writing style, linguistic habit, etc.</td><td>☐ Yes
☐ No</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>⑤ Is there possibility of identification due to a being/object associated to a person?

*(e.g.: image/video data) Case where there is uniqueness in a person's house, car(such as rare super cars), clothing, pet, etc.</td><td>☐ Yes
☐ No</td></tr>
<tr><td rowspan="3">Re-identification Impact</td><td colspan="2">Whether there is information that can cause serious damage or disadvantage to the information subject if re-identified</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>⑥ Is there any information that can make the information subject suffer damage or disadvantage due to social norm?</td><td>☐ Yes
☐ No</td></tr>
<tr><td rowspan="3">⑦ Is there any information that can cause significant damage or disadvantage to the information subject if re-identified?

*(e.g.: audio/text data) audio recording or consultation report on medical treatment/consultation that includes sensitive private information or a disease</td><td rowspan="3">☐ Yes
☐ No</td></tr>
<tr><td rowspan="3">Processing Environment</td><td rowspan="3">Use and Provision</td><td colspan="2">Whether there is identifiable information in consideration of how the pseudonymized information is used and the level of personal information protection of the organization using it.</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>⑧ Is there any identifiable information, in consideration of information available or accessible/obtainable to the processing entity as well as scope and type of use?

*(e.g.: image/video data) △Case where a doctor plans to analyze an X-ray data set that includes the X-ray of a patient the doctor had treated. A specific patient can be identified based on the doctor's treatment data and experience. △ Case where CT scans includes patients' meta data, allowing whoever has background information on a patient to identify that patient</td><td></td></tr>
</table>

<table>
<tr><th colspan="2">Category</th><th colspan="2">What to review for identification risks</th></tr>
<tr><td rowspan="6"></td><td rowspan="2"></td><td>⑨ Is additional information kept rather than deleting it?</td><td></td></tr>
<tr><td>⑩ When providing pseudonymized data to a party, what level of personal information protection does the recipientoffer and do they have reliable certification (ISMS, ISMS-P, ISO 27001, etc.)?</td><td></td></tr>
<tr><td rowspan="2">Processing Location</td><td colspan="2">Whether data pseudonymization is being process at a location that is safe in terms of management, technology,and physical location</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>⑪ Does the processing take place where other information can be accessed/obtained when processing pseudonymization?
*(e.g.:) △ Whether the location and network is accessible to anyone △ Whether the location and network is accessible only to those who have been authorized △ Whether the environment can restrict access to external information , data restoration technology, etc.</td><td></td></tr>
<tr><td rowspan="3">Combining with other Information</td><td colspan="2">Whether there is possibility of identification when pseudonymized information is combined with other information</td></tr>
<tr><td>Items to review</td><td>Review Result</td></tr>
<tr><td>⑫ Is there any plans to conduct analysis in connection to other data?</td><td></td></tr>
<tr><td></td><td>⑬ Is there information that can lead to identification by linking or combining other information available or accessible/obtainable to the processing entity?</td><td>☐ Yes
☐ No</td></tr>
</table>

6

Guidelines on Measures for Each Identification Risk

Category		Action Guideline
Data	Identifiability	● (Image/video data) – If biometric data that can be directly used to identify a person such as finger prints and retina have been taken in high resolution, such data is to be deleted in principle – If identifiable information such as a person's face, body part, car license plate, etc. is essential to achieve the purpose of processing, such information can be used after it has been safely pseudonymized. ＊Pseudonymization methods include image filtering (blurring, pixilation (mosaiced), masking (black box), image encryption, synthesizing facial image(such as k−same model), inpainting technique, etc. – In particular, blurring and pixilation can be reversed depending on the level of pseudonymization, and so additional information (image filtering algorithm, intensity of image filtering) has to be removed in principle – If there is uniqueness in body figure, clothing style, hair style, tattoo, etc. there is high probability of identifying a person. So, such data is to be deleted unless absolutely necessary for the purpose, and if such data is necessary, pseudonymization such as image filtering is required. ● (Voice data) – In a recording, information that can identify an individual (name, address, phone number, etc.) whether it be the speaker or the conversation partner have to be removed or replaced with generated information that can take its place ＊Special care must be taken to avoid identification of a person if a recording includes sensitive topics (e.g. sexual orientation, religion, disease) – In addition to the content of a recording, the voice of a speaker can be used in its own to identify a person, and so simple removal of personal information based on rules or

Category		Action Guideline
		pseudonymized using voice transformation, conversion, STT, etc. that are based on voice distortion. – In particular, when voices are distorted, it can be reversed if one knows the rules applied for distortion. So in principle, such additional information(voice distortion algorithm) should be removed. ● (Text data) – Information that can identify an individual (name, address, phone number, etc.) in text data such as free–form text, STT, caption for a video has to be removed or replaced with generated information than can take its place. * Pseudonymization methods include simple rule–basedremoval, replacement orscrubbing of personal information, regular expressions,annotation, etc.
	Unique Data	● Unique information may not lead to identifying an individual on its own, but due to its uniqueness (rarity), there is a high possibility of recognizing a person. Therefore, if it is not essential to achieving the purpose of use, it should be removed. Should it be essential, such information is to be safely pseudonymized before use.
	Impact when re–identified	● nformation that may have a significant impact such as discrimination based on social nor or violation of basic human rights, it can result in suffering of an individual as well as social repercussion when re–identified, unlike regular information. Therefore, such information is to be deleted unless absolutely necessary.
Processing Environment	Use & Provision	● f there is risk in use and provision data, the user and provider need to have a discussion on proving safety of the processing environment in order to lower risk. – As an example, there could be a case where identification risk is higher forathird–party recipient compared to when the provider uses the data internally due to the processing environment, resulting in an intense requirement for pseudonymization. If the user determines that the purpose of processing cannot be achieved with such high level of pseudonymization, the two parties can have a discussion on

Category		Action Guideline
		to lower the level of pseudonymization on the premise that the control on the processing environment will be strengthened – If the original data is retained even after pseudonymization, identification risk may increase by comparing with and making connections with the original data. Identification risk also increases if the original data includes meta data that can identify an individual and also if one has background information of the original data, requiring special attention.
	Processing Location	● If the level of pseudonymization of a dataset needs to be lowered for use, taking additional safety measures to the processing location in terms of physical, managerial, and technological aspect to lower the overall risk of identification. – In the event it is difficult to ensure relevance or reliability a pseudonymization technology, overall identification risk can be lowered by strengthening control of data itself. ● If the pseudonymized data is vulnerable to data restoration technology, an environment has to be set up to restrict access and use of external information and restoration technology(software).
	Combining with other Information	● If there is a plan to make connections and combine with other information, addition review should be conducted to see whether there is any information whose identification risk increases when combined with such other information. ● Review whether there is any information that can lead to identification through information available or accessible/obtainable to the processing entity – Since it is difficult for the data provider to determine on their own whether the party to receive and use a pseudonymized dataset has experience or knowledge from the past of handling similar information, it is necessary to check and review in advance the party to receive the pseudonymized dataset. In the event such review is difficult, the risk must be lowered by increasing the level of pseudonymization.

Main Contents of Guidelines for Pseudonymizing Unstructured Data

개인정보보호위원회
Personal Information Protection Commission