JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 27 No. 6, pp. 53-63, June 2022
https://doi.org/10.9708/jksci.2022.27.06.053

A study on Data Context-Based Risk Measurement Method for Pseudonymized Information Processing

Dong-Hyun Kim*

*General Researcher, Korea Internet & Security Agency, Naju, Korea

[Abstract]

Recently, as digital transformation due to the COVID-19 pandemic accelerates, data to improve individual quality of life is being used in large quantities, and more reinforced non-identification processing procedures are required to utilize the most valuable personal information among data. In Korea, procedures for de-identification measures are presented through amendments to laws and guidelines, but there is no methodology to measure the level of de-identification in the field due to ambiguous processing standards and subjective risk measurement methods. This paper compares and analyzes the current status of policy and guidelines related to de-identification measures proposed at home and abroad to derive complementary points, suggests a data context-based risk measurement method centered on pseudonymized information processing, and verifies its validity. As a result of verification through Delphi survey and focus group interview (FGI), it was confirmed that the need for the proposed methodology and the validity of the indicators were high.

▸ **Key words**: BigData, Personal Information, De-Identification, Pseudonymized information, Pseudonym Risk Assesment

[요 약]

최근 코로나19 팬데믹으로 인한 디지털 트랜스 포메이션이 가속화되면서 개개인의 삶의 질을 향상시키기 위한 데이터가 대량으로 활용되고 있으며 데이터 중 가장 가치 있는 개인정보를 활용하기 위한 보다 강화된 비식별 처리 절차가 요구되고 있다. 국내에서도 법률 개정과 가이드라인을 통해 비식별 조치를 위한 절차를 제시하고 있지만 모호한 처리 기준과 주관적인 위험도 측정 방식으로 인해 현업에서는 비식별정보의 처리 수준을 측정할 수 있는 방법론이 부재한 상황이다. 본 논문은 국내외에서 제시하고 있는 비식별 조치 관련 제도 및 지침 등에 대한 현황을 비교분석하여 보완점을 도출하고 이를 해결하기 위해 가명정보 처리 중심의 데이터 상황 기반 위험도 측정 방법을 제안하고 타당성을 검증하고자 한다. 델파이 조사 및 표적집단면접(FGI)을 통한 검증 결과 제안한 방법론에 대한 필요성과 지표들에 대한 타당성이 높은 것을 확인하였으며, 실무에서 이를 활용할 경우 가명정보의 위험성을 측정하는데 많은 도움이 될 것으로 사료된다. 또한 제안하는 방법론은 가명정보의 위험성을 계량적인 체크리스트 방식으로 측정할 수 있는 유일한 방법론이란 점에서 의의가 있다.

▸ **주제어**: 빅데이터, 개인정보, 비식별 조치, 가명정보, 가명처리 위험도 측정

• First Author: Dong-Hyun Kim, Corresponding Author: Dong-Hyun Kim
*Dong-Hyun Kim (kdonghyun@kisa.or.kr), Korea Internet & Security Agency
• Received: 2022. 05. 10, Revised: 2022. 06. 13, Accepted: 2022. 06. 13.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

4차 산업혁명 시대의 지능정보기술은 사람과 인공지능에 기반을 둔 정보와 사물 인터넷(IoT), 빅데이터, 클라우드 등과의 상호 연결을 통해 기술과 사회의 융합을 가속화하고 있다. 과거 기업들은 자사가 수집한 대량의 정보를 독자적으로만 이용할 수 있는 시대였다면 이제는 다른 산업의 정보와 융합을 통해 빅데이터를 생성하고 보다 창의적이고 획기적인 신산업을 발굴할 수 있는 시대가 등장한 것이다. '21세기 원유'라 불리는 빅데이터는 현재 정보화 사회에서 거부할 수 없는 거대한 트렌드로 우리 사회에 다가오고 있으며[1], 코로나 팬데믹으로 인한 디지털 트랜스포메이션이 가속화 되면서 데이터가 경제적 자산이 되고 가치창출의 원천이 되는 '데이터 경제[2]'의 시대가 시작되었다. 우리나라에서도 디지털 뉴딜 등의 데이터 경제 활성화 계획[3]을 적극 추진한 결과, `21년 스위스 경영대학원의 보고서에 따르면 Fig. 1과 같이 디지털 국가경쟁력은 `17년 대비 7단계, 빅데이터 활용과 분석지표는 30단계가 상승한 것으로 나타났다[4].

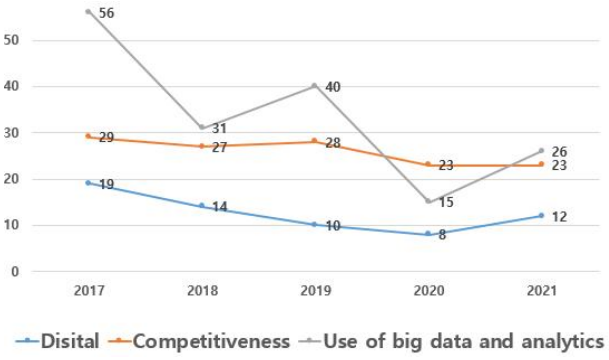

Fig. 1. IMD World Competitiveness Ranking

한편 일명 FFANG[1]이라 불리는 미국의 5대 IT기업은 이미 `20년 나스닥에서 차지하는 시가총액 비중이 40%를 육박하고 있다. 이렇게 선진국들이 개인정보를 데이터로써 활용을 촉진하는 가운데 우리나라도 현행 개인정보보호 법령의 틀 내에서 빅데이터가 안전하게 활용될 수 있도록 `16년 '개인정보 비식별 조치 가이드라인[5]'을 마련하였다. 그러나 비식별정보에 대한 법적 근거 및 명확한 정의 미흡 등의 문제로 `20년 개인정보 보호법을 개정하고 동년 9월 '가명정보 처리 가이드라인[6]'을 발간하였다. 위와 같은 제도 마련에도 불구하고 가이드라인에서는 일반적인 절차만 제시하고 있을 뿐 실무적으로 가명정보에 대한 위험성을 측정하거나 가명 처리 수준을 검토하기 위해 계량으로 측정하는 절차는 제시하고 있지 않다. 본 연구는 국내 개인정보 보호 법률에 근거하여 조직 내 가명처리를 수행하는 실무자들이 개인정보를 안전하게 활용할 수 있도록 데이터 상황에 기반을 둔 위험도를 측정하고 처리 수준을 정할 수 있도록 도움을 주고자 하는 것이 목적이다. 이를 위해 국내외 비식별 처리 지침[5][6][10-13]과 위험도 측정방법에 대한 문헌들[14-18][20]과의 비교분석을 통해 데이터 상황 기반의 정량적 위험도를 산출할 수 있는 방법론을 새롭게 제안하고 타당성 검증을 통해 향후 국내에서 추진하는 정책에 대한 개선안을 제시하고자 한다. 이러한 연구의 목표를 달성하기 위해 다음과 같은 요구사항들을 도출하였다.

첫째, 국내외에서 개인정보를 비식별 조치하여 활용하기 위한 구체적인 절차가 있는가?

둘째, 국내외에서 제시하는 절차에서 데이터의 상황을 기반으로 위험도를 측정하여 비식별 조치 수준을 결정하는 방법이 있는가?

셋째, 비식별 조치 수준을 정성이 아닌 정량으로 측정하여 판단할 수 있는 방법이 있는가?

위와 같은 요구사항들을 해결하기 위해 제2장에서는 국내외에서 제시하고 있는 정책과 비식별 조치를 위한 구체적인 절차(가이드 등)들에 대해 살펴보고, 제3장에서는 선행연구에 대한 비교분석을 바탕으로 국내 컴플라이언스에 적합한 데이터 상황 기반의 위험도 측정 방법을 새롭게 제안한 후, 제4장을 통해 제안한 방법에 대한 적합성과 타당성을 검증하여 종합적인 결과를 분석하고, 마지막 제5장을 통해 결론을 맺고자 한다.

II. Status of de-identification policies at home and abroad

그동안 개인정보 비식별 조치는 우리나라에 비해 국외에서 더욱 활발한 정책 및 가이드라인이 제시되어 왔다. 캐나다, 미국, 영국 등에서는 개인정보 관련 규정의 개정 또는 일반 보고서, 지침 등을 통해 새로운 비식별 조치 절차를 제시하고 있으며, 국제표준으로는 ISO/IEC 20889[9]에서 비식별 조치 기술에 대해 언급하고 있다. 본 연구에서는 다양한 국내외 문헌 중 비식별 조치에 대한 절차와 위험도를 측정하는 방법을 제시하는 문헌들을 대상으로 선행 연구를 실시하였다.

1) Facebook, Amazon, Apple, Netflix, Google

1. Guidelines for De-Identification of Personal Information

국내에서는 `16년 '개인정보 비식별 조치 가이드라인[5]'을 통해 개인정보를 비식별 조치하여 빅데이터로 활용하는 방안을 마련하였다. 동 가이드라인은 1단계(사전검토), 2단계(비식별 조치), 3단계(적정성 평가), 4단계(사후관리)의 절차로 구성되어 있으며, 가이드라인에 따라 적정하게 처리된 정보는 EU의 익명화(anonymization)와 사실상 같은 개념으로 활용 목적에 제한이 없으며, 불특정 다수에게도 공개할 수 있는 정보로 정의[7]하고 있다. 동 가이드라인의 특징으로는 k-익명성[8]이란 프라이버시 보호 모델을 3단계(적정성 평가)에서 필수적으로 적용을 하고, 외부전문가를 통해 평가를 수행 후 활용하여야 한다. 프라이버시 보호 모델은 비식별정보에 어느 정도의 추론 방지 기술과 수준이 적용되었는지 계량적으로 측정 가능한 방법론[9]이라 할 수 있다. 동 가이드라인을 통해 생성되는 비식별정보(익명정보)는 활용 목적에 제한이 없는 만큼 공개된 정보 등 다른 정보와의 결합 등을 통해 재식별이 될 수 있기 때문에 3단계(적정성 평가)는 재식별 방지를 위한 위험도 측정 방법 중 하나로 분류될 수 있다.

2. Guidelines for processing pseudonym information

`20년 8월, 데이터 경제 활성화 및 빅데이터 활용의 법적근거가 되는 개인정보 보호법이 개정되어 동년 9월, 실무에서 가명정보를 원활히 활용하기 위한 '가명정보 처리 가이드라인[6]'이 마련되었다. 동 가이드라인은 1단계(사전준비), 2단계(가명처리), 3단계(적정성 검토 및 추가가명처리), 4단계(활용 및 사후관리)의 절차로 구성되어 있으며, 기존 비식별 조치 가이드라인[5]과의 가장 큰 특징으로는 개인정보를 가명처리 시 개인정보의 유형, 특성 등을 고려하여 법령을 준수하는 범위 내에서 처리 절차와 방법을 자율적으로 판단할 수 있도록 안내하고 있으며, 2단계(가명처리)를 다시 4단계로 세분화하여 가명정보의 처리(제공) 환경을 검토하고, 항목별 위험도를 분석하는 '위험도 측정' 절차가 추가되었다.

동 가이드라인의 위험도 측정 절차를 살펴보면 첫 번째 대상선정 단계에서의 재식별 가능 항목을 제거하고, 두 번째 위험도 측정 단계에서의 데이터의 맥락을 분석하며, 마지막으로 데이터 자체에 대한 재식별 가능성을 검토함으로써 가명처리의 적정성을 유지하고 있다. 여기서 맥락 분석은 가명정보처리자의 개인정보 보호수준 및 다른 정보 보유여부 등을 검토하고, 처리(제공)환경에 따라 처리자 내부 활용 또는 내부의 다른 부서에 제공 또는 외부의 제3자 제공으로 구분함으로써 가명처리에 대한 위험도를 다르게 산정하도록 제시한 것이다.

동 가이드라인은 가명정보의 처리 목적, 처리(제공) 환경, 정보의 특성 등을 종합적으로 고려하여 위험도를 측정하고, 위험도 측정을 통해 정의한 가명처리 수준을 가명처리 전에 수립함으로써 기존 '비식별 조치 가이드라인[5]'의 문제점을 해결하였다. 다만, 위험도 측정 시 이용기관의 개인정보 보호수준 및 다른 정보 보유여부에 대한 검토는 현실적으로 어려우며, 데이터 자체에 대한 위험성을 구체적으로 측정하는 절차는 제시되고 있지 않다. 그 외 동 가이드라인을 기반으로 한 금융[10], 보건의료[11], 공공[12] 분야에서 동 가이드라인에 기반을 둔 분야별 가명처리 안내서가 마련되었으며 가이드라인별 차이점은 Table. 1과 같다.

Table 1. Comparing the current status of domestic de-identification guidelines

(A : Anonymisation, P : Pseudonymization)

Type	Target	step	Assessment	Characte ristic
De-Identifica tion guidelines[5]	A	Step.4 (None)	manda tory	Required application of privacy model
Pseudonymi zation guidelines[6]	P	Step.4 (Step.2)	auto nomy	-
Financial sector guidelines[10]	A, P	same as above	P (auto nomy), A (manda tory)	Obligation to keep records
Health sector guidelines[11]	P	same as above	manda tory	DRB[2] installation and operation
Public sector guidelines[12]	P	same as above	auto nomy	Pre-conformity review for availability

3. Canada IPCO De-Identification Guideline

캐나다의 경우 `16년 6월 IPCO(Information and Privacy Commissioner of Ontario)에서 'De-identification Guidelines for Structured Data[13]'라는 가이드라인을 제정하고, 비식별 조치를 한다고 해서 데이터의 재식별 위험이 완전히 제거되는 것이 아니며 재식별 위험을 허용 가능한 수준 또는 매우 작은 수준으로 만들 수 있는 관리적 도구라 정의하고 있다. 또한, 이러한 비식별 조치 시 위험성을 매우 작게 만들기 위해 위험성에 기반을

2) 데이터 공개검토위원회(DRB : Disclosure Review Boards)

둔 접근법(Risk-Based Approach)을 Table. 2와 같이 제시하고 있다.

Table 2. The 9-step De-Identification measure model presented in the Canadian Guidelines

Step.1	Open model selection
Step.2	Variable classification
Step.3	Determining the Re-identification Risk Threshold
Step.4	Measuring the risk of the data itself
Step.5	Measuring risk in data context
Step.6	Total risk calculation = data risk × context risk
Step.7	Data de-identification
Step.8	Data utility assessment
Step.9	Record each step of the procedure

동 가이드라인은 비식별 조치 전 재식별 위험에 대한 한계치를 결정하고 설정한 한계치에 따라 비식별 조치 수준을 결정 및 적용하도록 제시하고 있으며, 위험도 측정의 기본 기술은 k-익명성[8] 사용을 제시하고 있다. 이러한 위험도 측정 절차는 `16년에 마련된 '개인정보 비식별 조치 가이드라인[5]'의 적정성 평가 단계에서도 인용을 하고 있음을 시사하고 있다.

4. HITRUST De-Identification Framework

`15년 미국 HITRUST(Health Information Trust Alliance)사에서는 건강 정보와 관련한 관계자들 간의 위험 정보와 컴플라이언스를 공유하고 개인정보 비식별 조치를 위해 일관되고 관리 가능한 방법론을 제공하기 위한 'De-Identification Framework[14]'를 발간하였다. 본 프레임워크의 주요 내용으로는 여러 단계의 비식별 조치 수준에 대한 정의와 각각에 따른 구체적인 유스케이스를 추천하고 비식별 조치 방법론에 대한 평가와 방법론의 실무 적용 시 보증을 위한 재식별 위험과 기준을 평가하고 있다. 또한, HITRUST CSF(Common Security Framework)를 이용하여 비식별정보의 이용, 저장, 유지관리와 관련한 위험들을 경감시키기 위한 프레임워크 및 부록을 통해 기타 위험에 대한 대처방안들을 제공하고 있다.

그 외에 Privacy Analytics사의 설립자이자 오타와 대학교수인 Khaled El Emam은 이러한 프레임워크를 구체적으로 설명하기 위한 책(Anonymizing health data)[15]을 발간하였는데 그 중 비식별화 위험도 측정의 핵심으로 '비식별화 방법론에 대한 평가'는 관리적인 요소인 Program Methodology(관리 유무, 문서화, 관리자 식별, 독립적인 조사 등 4개)와 데이터 자체에 대한 De-Identification Methodology(한계 값 설정, 위험 측정, 식별자 구분, 위험 식별, 데이터 안전조치, 프로세스 정립, 잔여위험 관리, 유용성 검토 등 8개) 총 12가지 평가지표를 제시하고 있으며, '재식별 위험 관리 방법'으로는 완화 제어(Mitigating Controls), 프라이버시 침해(Invasion of Privacy), 동기 및 능력(Motives & Capacity)에 관한 사항을 판단하여 측정을 하도록 제시하고 있다.

5. UKAN The Anonymisation Decision-Making Framework

영국 정보보호위원회(ICO, The Information Commissioner's Office)는 `12년 '개인정보 익명화 실무 지침(Anonymisation : managing data protection risk code of practice)'을 발표하고 EU Directive 95/46/EC Recital 26에 근거하여 '익명화된 정보(Anonymous data)'는 데이터보호법(Data protection Act)에 적용되지 않는다고 명시하고 있다. 영국의 익명화 네트워크(The UK Anonymisation Network, 이하 'UKAN')는 이러한 실무 지침을 구체적으로 설명 및 보충하고, 나아가 실무자가 현장에서 참조할 수 있는 문헌을 제공하기 위해 `16년 '익명화에 관한 의사 결정 프레임워크(The Anonymisation Decision -Making Framework)[16]'를 마련하였다. 본 프레임워크의 주요 특징으로는 ICO의 '개인정보 익명화 실무 지침에 비해 보다 실무적인 내용과 기술적인 내용을 많이 포함하고 있다. 효과적인 익명화를 위해서는 데이터가 활용되는 환경, 즉 데이터 맥락(context)을 중요한 요소로 고려하여 절차를 수집해야 하며 변수들의 특징 및 활용되는 환경적 요인에 따라 식별이 가능한 정보와 식별이 불가한 정보로 구분되기 때문에 데이터의 활용 목적과 제공 범위 등의 이용 환경을 충분히 고려해야 한다고 제시하고 있다[17]. 이처럼 데이터만이 아니라 데이터를 둘러싼 주위 환경을 고려하는 접근을 '데이터 상황적 접근법(data context approach)'이라고 표현한다. 익명화 결정 프레임워크를 수립하는데 기반이 되는 다섯 가지 원칙은 Table. 3과 같다.

Table 3. UKAN Anonymisation Decision-making Framework 5 Principles

Step.1	You can't just look at the data and make a decision whether it's safe to share or distribute it
Step.2	But still need to know the data
Step.3	Anonymization is a safe way to handle data, but it must be possible to use data safely and usefully even after anonymization
Step.4	Realistically, zero risk (no risk at all) is impossible when you want to produce useful data
Step.5	Means to manage risk should be determined in proportion to the risk factors and their impact

Table. 3 원칙의 기반 하에 프레임워크를 구성하는 핵심요소로 4개의 익명화 유형(형식적, 보장된, 통계적, 기능적)을 제시하고 있으며, 기능적 익명화는 형식적, 보장된, 통계적 익명화의 단점을 모두 보완한 것으로 데이터와 데이터 환경(Environment) 모두를 포함한 데이터 상황을 고려할 수 있는 절차를 제시하고 있다.

6. NIST De-Identifying Government Datasets

미국 국립표준연구소인 NIST는 정부 데이터셋을 비식별화 하여 사용할 수 있도록 NIST SP800-188 (De-Identifying Government Datasets)[18] 표준을 마련하였다. 이 표준은 미국 정부의 데이터셋에 대한 비식별화 방법의 선택과 사용, 평가 방법을 가이드하고 있으며 데이터 비식별화 절차의 관리를 위한 프레임워크를 제공한다. 또한, 이 문서의 최종 목적은 고의적인 데이터 공개를 통해 발생할 수 있는 특정 개인 정보의 노출 위협을 최소화하는데 그 목적이 있다고 제시하고 있다. 동 지침은 비식별을 수행하는데 있어 크게 8단계의 절차를 소개하고 있으며, 이에 따라 각 절차 내에서 수행해야 할 기술들은 Table. 4와 같다.

Table 4. NIST's De-Identification procedures

Step.1	Identify the purpose of de-identification
Step.2	Assessing threats arising from data disclosure
Step.3	Data life cycle
Step.4	Data Sharing (Public) Model
Step.5	The Five Safes
Step.6	DRB installation and operation
Step.7	De-identification standard
Step.8	Education, training and research

동 지침은 Step.2에서 비식별화 된 데이터 공개로부터 발생하는 위험 평가를 과학적이고 객관적인 요소를 기반으로 이뤄져야하며, 데이터셋에서 개인에 대한 최선의 이익과 데이터 보유 기관의 책임 및 사회에 대한 기대 편익을 고려해야 한다고 제시하고 있다. 즉 데이터 공개에 대한 재식별 가능성과 재식별로 인한 부정적 효과를 평가하도록 하고 있다. Step.5에서는 영국 Bristol University에서 제안된 Five Safes(Safe Projects, Safe People, Safe Data, Safe Settings, Safe Outputs)[19]를 데이터 접근 시 중요하게 고려해야 할 사항으로 제시하고 있다.

7. Risk Assessment Based on Data Context for De-identification

김동현 등[20]은 `20년 개정된 데이터 3법에 맞춰 실무자들이 개인정보를 활용함에 있어 비식별 조치 수행 시 위험도에 따른 처리 수준을 산정하기 위한 측정 방법론을 제시하였다. 해당 논문은 위험도 측정 시 데이터만이 아닌 데이터를 둘러싼 주위 상황을 고려하여 처리를 해야 하며, 이러한 데이터 상황을 3가지(데이터 활용방법, 데이터 이용환경, 데이터)로 분류하여 각 상황별 위험도 측정을 계량적으로 할 수 있도록 제안하고 있다. 해당 논문은 본 연구 주제와 매우 유사한 측정 방식을 제시하고 있지만 제안하는 비식별정보의 위험도 측정 방법론이 익명정보를 기준으로 정의하고 있어 현재 개정된 개인정보 보호법의 가명정보 활용 시 이용하기에 어려우며, 세부 지표 또한 가명정보 처리에 맞춰 데이터 활용 환경 및 이용환경 등을 재정의하고 검증할 필요가 있다.

8. Implications for Prior Research

앞서 살펴본 선행연구를 통해 이러한 데이터 상황기반의 위험도를 측정하는 접근법에 있어 가장 중요한 원칙은 첫째, 그 나라의 법과 제도적 요소에 맞추어 적용되어야 하며 둘째, 특정 분야에 치우치지 않고 범용성 있게 적용이 될 수 있어야 하고 셋째, 위험에 대해 계량적으로 측정이 가능해야 객관적인 측정 지표가 제시될 수 있다는 것이다. 국내외의 제도 현황[5][6][10-13]의 경우 법과 제도적인 요소를 고려하여 절차를 제시하였지만 구체적으로 각 절차별 어떤 사항을 검토해야 하는지에 대한 기술은 없었으며, 데이터 상황에 기반을 둔 위험도 측정방식[14-18]은 비식별정보에 대한 위험도 측정 방안을 제시하고는 있으나, 익명정보를 중심으로 하고 있으며 특히, 의료분야에 한정이 되어 제시하고 있다는 단점이 있었다. 마지막으로 김동현 등[20]이 제시한 위험도 측정 방법론이 본 연구의 맥락과 가장 유사하나 익명정보를 기반으로 제시하고 있다는 점에서 가명처리 시 활용하기 어렵다는 문제가 있었다. 본 논문은 이러한 단점들을 개선하고자 최근 개정된 우리나라의 데이터 3법에 따르면서 데이터 상황 접근법에 따라 각 상황을 국내의 환경에 맞추어 체계적으로 분류하고, 위험도 측정의 경우 계량적으로 측정 할 수 있는 방법을 제3장을 통해 제시하려 한다.

III. Research Proposal

1. Improvements to the De-Identification Process

앞서 살펴본 선행연구들을 통해 문제점을 보완할 수 있는 가명처리 중심의 비식별 조치 절차를 Fig. 2와 같이 제안하였다.

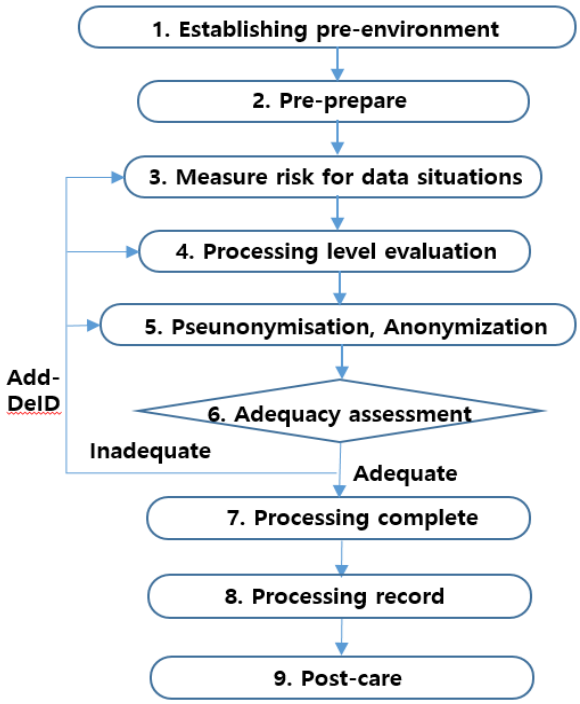

Fig. 2. New Proposed De-Identification Procedure

Fig. 2의 절차를 현재까지 발간된 국내 비식별 조치 관련 가이드 등[5][6][10-12]과 비교해보면 1. 데이터를 처리 전 데이터 활용을 위한 사전 환경을 구축하는 부분과 3. 데이터 상황에 따른 위험도를 계량적으로 측정하는 부분, 그리고 마지막으로 8. 비식별 조치 과정 전체를 기록하는 단계가 추가되었다. 상기 3가지 신규 추가 절차에 대해 가명정보는 개인정보로서 익명정보와 달리 활용할 수 있는 범위를 개인정보 보호법에서 제한[21]하고 있기 때문에 사전에 처리 환경을 구축하지 않을 경우 가명처리를 수행하여도 활용을 못하거나 위법을 하는 상황이 발생할 수 있다. 다음으로 데이터 상황에 따른 위험도를 계량적으로 측정하는 부분은 본 장의 제2절에서 구체적으로 설명을 하고, 과정을 기록하는 단계는 캐나다 IPCO[13], 영국 UKAN[16], 미국 NIST[18] 등의 국외 문헌에서 데이터 활용에 대한 투명성과 표준화 된 처리방법을 관리하기 위한 절차로 제시하고 있는 절차를 반영하여 추가하였다.

2. Proposal of risk measurement method based on data context

Fig. 2의 '3. 데이터 상황에 대한 위험도 측정 단계'는 데이터 상황에 대해 비교적 구체적으로 위험성 측정 절차를 제시하고 있는 UKAN[16], Five Safes[19], Duncan[22] 및 김동현 등[20]이 제안한 위험도 측정 방법을 가명정보 처리를 중심으로 재구성하여 제안하였으며, 제안한 관점에 대해서는 비식별 조치 관련 경험이 풍부한 해당분야 전문가로 학계 교수(경력 22년) 및 비식별 조치 컨설턴트 2인(경력 18년, 15년)과 3차에 걸친 검토 회의를 통해 Table. 5와 같이 새롭게 정의하였다.

Table 5. Proposed data context-based risk measurement method

Main Category	Middle Category	Sub-Category
A. How to use data	A.1 Who(User)?	2
	A.2 How(Use Type)?	5
	A.3 How(Combination status)?	3
B. Data usage environment	B.0 Level of personal information protection	14(41)
	B.1 Trust level of user organization	2(7)
	B.2 Impact of information subject on re-identification	1(2)
C. Data attribute	C.1 Data organization	8
	C.2 Data distribution	2
	C.3 Data sensitivity	5

첫째, A.데이터 활용방법의 A.1 누가(User) 항목은 비식별정보를 활용하는 주체를 의미하며, A.2 어떻게(Use Type)는 A.1에서 파생되는 활용 유형을 의미한다. 예를 들면 A.1에서 기관 내부에서 활용할 것인지 또는 외부로 제공할 것인지 결정을 하고, A.2에서 보다 구체적인 활용 유형을 결정할 수 있도록 하였다. A.3의 경우 가명정보 결합에 대한 부분으로써 이러한 3가지 구성을 통해 가명정보를 활용할 수 있는 모든 경우를 고려하여 정의하였다. A. 데이터 활용 방법에서의 위험 측정은 데이터가 재식별될 가능성이 높은 환경에 비례해서 계산을 한다. 예를 들어 자사가 보유한 데이터를 기업 내부에서 외부에 노출을 하지 않고 가명처리 후 데이터를 분석하는 경우 외부의 공격 또는 온라인상의 공개되어 있는 다른 정보 등과의 재식별 위험성이 현저히 떨어진다. 반대로 자사의 데이터를 동종 업계인 외부로 제공하는 경우 유사한 데이터를 통해 재식별될 가능성이 높아지게 된다. 즉, 내부에서 자체 활용할 경우 재식별에 대한 위험도가 감소하기 때문에 낮은 점수를 배점하고, 외부에서 타부서 등이 활용하는 경우 재식별 위험이 높아지기 때문에 높은 점수를 배점 받을 수 있도록 설계하였다.

둘째, B.데이터 이용환경은 각 구성에 따라 계량으로 측정할 수 있는 세부 지표(Sub-Category)를 제공하고 있다. 그 중 B.0 개인정보보호 수준은 개인정보 보호법에서 규정하고 있는 안전성 확보조치(개인정보 보호법 제28조의4 및 제29조)를 기반으로 41개의 체크리스트를 만들어 제공

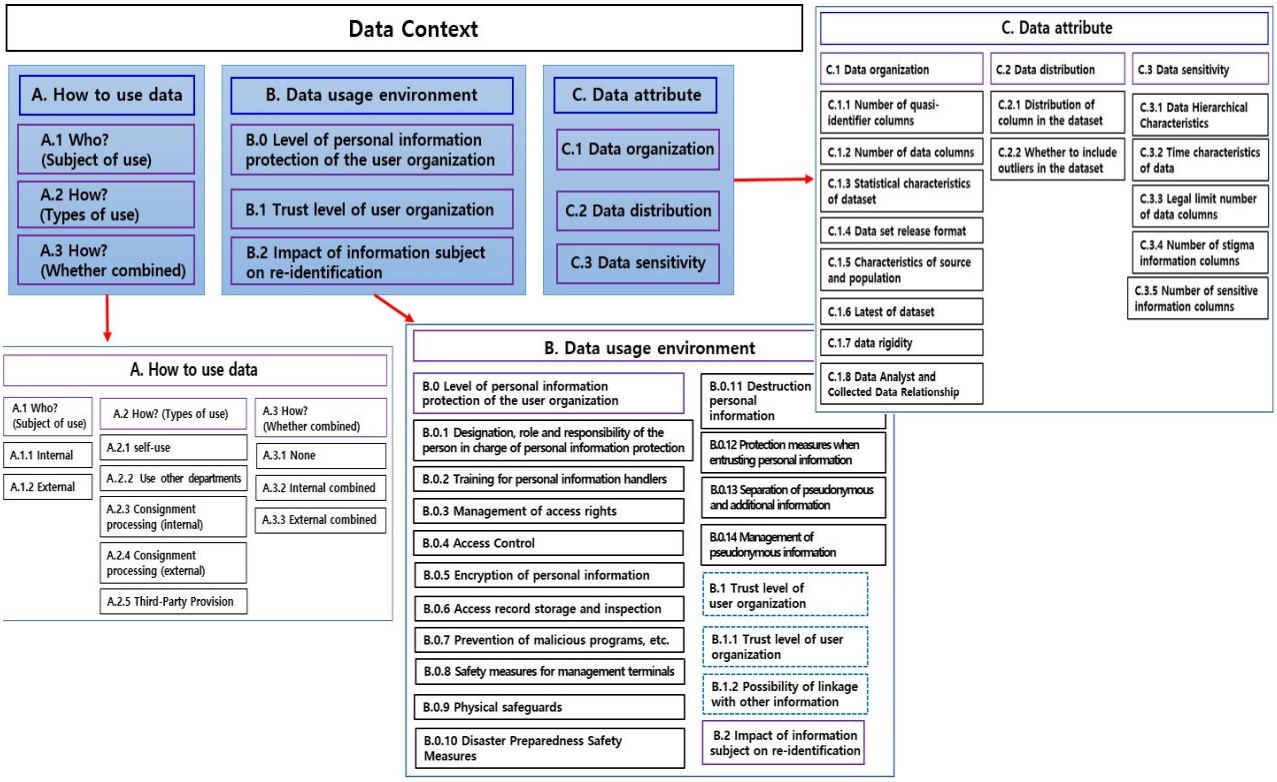

Fig. 3. Proposed data context-based risk measurement method as a whole

하고 있으며, 해당 체크리스트는 법률에서 필수적으로 규정하고 있는 사항으로 모두 적정이여야만 만족을 할 수 있도록 설계하였다. B.1 이용기관의 신뢰수준은 비식별정보를 사용하는 기관의 신뢰도를 측정하는 것으로 계약서 체결 등을 통한 관리적인 요소와 기술적인 요소로 제공하는 정보와 유사한 정보를 활용하는 기관이 보유하고 있는지 등을 측정한다. B.2 재식별 시 정보주체에 미치는 영향은 비식별정보가 재식별 되었을 때 사회적 혼란을 가져오거나 개인의 프라이버시를 심각하게 침해하는 경우 등을 측정한다.

셋째, C.데이터는 C.1 데이터의 구성에서 식별자와 준식별자의 컬럼 개수, 데이터의 제공형태, 최신성, 원본과 모집단의 특성 등을 측정하고, C.2 데이터 분포는 데이터 셋 내의 컬럼 값들의 분포와 극단치, 특이치 여부 등을 측정한다. C.3 데이터 민감도는 데이터의 계층적, 시간적 특성과 낙인성, 민감정보의 컬럼수를 측정한다.

마지막으로 A.데이터 활용 방법, B.데이터 이용환경, C.데이터 자체에서 측정된 값을 모두 합산하여 최종 위험도를 판단하게 되며, 최종 위험도에 따른 처리 기준은 기관 또는 기업마다 판단하는 기준이 다르기 때문에 활용하는 곳에서 적정하게 배점을 정의하여 활용할 수 있다. 예를 들어 제안하는 데이터 상황 기반의 위험도 측정 방식을 활용하여 나온 결과의 위험도가 높을 경우 일반적인 수준보다 강한 비식별 조치 수준을 적용하고, 위험도가 낮을 경우 비식별 조치 수준을 낮추어 보다 유용한 가명정보를 생성할 수 있도록 제안하였다. 제안한 위험도 측정 방법론에 대한 전체 구조도는 Fig. 3과 같다.

IV. Validation of data context-based risk measurement indicators

1. Research Method

제3장에서 제안한 데이터 맥락 기반의 가명처리 위험도 측정 방법론 지표에 대한 타당성 및 적정성을 검증하기 위해 비식별 조치 관련 경력 15년 이상 및 비식별 조치 경험이 있는 실무자를 대상으로 델파이조사와 전문가 대상 표적집단면접(Focus Group Interview)을 실시하였다. 설문 구성은 리커트(Likert) 5점 척도를 활용하여 측정하였으며, 타당성이 2점 이하인 경우 각 항목에 대한 세부 의견을 수렴하였다. 제시하고 있는 방법론에 대한 절차의 타당성과 세부 변수 구성에 대한 적절성을 판단하는 설문 정의는 Table. 6과 같다.

Table 6. Define the Appropriateness review survey

var	question	
V-1	Is it appropriate to divide the risk by data context into A. data usage method, B. data usage environment, and C. data (self)?	
V-2	How to use data	Is it appropriate to classify data utilization methods into who (subject to use), how (type of use), and how (with or without combination)?
V-3	Data usage environment	Is it appropriate to divide the data usage environment into the level of personal information protection, the level of trust of the user organization, and the impact on the data subject in case of re-identification?
V-4	data (itself)	Is it appropriate to classify the risks to the data itself by the composition, distribution, and sensitivity of the data?

다음으로 계량적인 체크리스트가 제공되는 B.데이터 이용환경과 C.데이터 변수에 대한 영향도 측정 지표는 Table. 7과 같다.

Table 7. Define the impact review survey

var	question
E-1	There are economic (monetary) benefits for data users to re-identify data * [Standard] In the case of a declaration of conflict of interest, the standard is generally equivalent to 10 million won
E-2	Data users have non-economic (political, opinion manipulation, self-interest, etc.) benefits in re-identifying data. * [Reference] A minimum of 1 point (very unaffected) to a maximum of 5 points (very impactful) are given depending on the impact of non-economic profits
E-3	A data user has violated the Personal Infor mation Protection Act and subordinate statutes within the past 3 years * [Note] '1' points for 'no', '5' points for 'yes'
E-4	The phrases such as prohibition of re- identification and restrictions on provision to third parties are not reflected in the 'use (provision) related contract' between the data user and the provider * [Note] '1' points for 'no', '5' points for 'yes'
E-5	Data users have experience in processing (business entrustment, joint execution, etc.) of personal information related to pseudonymous information provided by the data provider (including original infor mation before processing of pseudony mous information or other pseudonymous information that has processed the same personal information) * [Standard] The higher the number of executions, the higher the score
E-6	There is business-accessible information that data users can link to to re-identify the information provided. * [Note] '1' points for 'no', '5' points for 'yes'
E-7	There may be data that can be combined with information provided on the Internet, such as public data disclosure.
F-1	When data is re-identified due to inappropriate disclosure/use/violation/processing, it may lead to social confusion due to legal, moral and technical issues * [Reference] Disposition of laws and fines due to illegality, embezzlement, self-interest, re-identification due to insufficient pseudonym handling, ethical issues due to AI use, etc.
F-2	When data is re-identified due to improper disclosure/use/violation/processing, it may infringe the privacy or privacy of the relevant data subject.
F-3	When data is re-identified due to improper disclosure/use/violation/processing, it may cause economic/non-economic loss to the relevant data subject * [Reference] The image of the subject of information is damaged, the use of crimes due to theft, etc.
F-4	When data is re-identified due to improper disclosure/use/violation/processing, it may cause economic/non-economic loss to the provider * [Reference] Decline in economic feasibility of business items, decline in stocks, deterioration of corporate image, etc.
G-1	Number of columns of direct/indirect identification information of data
G-2	Total number of columns in the dataset
G-3	Statistical characteristics of datasets
G-4	The scope of the dataset
G-5	Characteristics of source and population
G-6	The latest data set
G-7	Distribution of column values in the dataset
G-8	Whether to include outliers in the dataset
G-9	Time characteristics of data
G-10	Legal limit number of data columns
G-11	Number of stigma information columns
G-12	Number of sensitive information columns

2. Research Result

설문조사의 신뢰도 및 품질을 유지하기 위해 1차 조사는 단순한 설문조사가 아닌 직접 만나서 제안하는 방법론을 설명하고 현장에서 토론과 의견을 청취하는 조사로 진행하였다. 공공, IT/통신, 금융, 의료분야의 관련 분야 경력 10년 이상 전문가 10인을 대상으로 `21.8.19~8.25일까지 진행하였으며, 응답 값에 대한 신뢰도 측정을 위해 IBM SPSS 26.0을 이용하여 Cronbach α값을 측정을 하였다. 조사결과는 Table. 8과 같다.

Table 8. Delphi 1st Investigation Results

variable	Mean	SD	Cronbach α
V-1	4.5	0.642	0.837
V-2	4.2		
V-3	3.8		
V-4	4.3		
E-1	4	1.749	0.849
E-2	3.7		
E-3	3.7		
E-4	3.7		
E-5	4.5		
E-6	3		
E-7	3.5		
F-1	4	1.022	0.904
F-2	4.3		
F-3	3.8		
F-4	3.7		
G-1	3.2	1.250	0.922
G-2	3.8		
G-3	4		
G-4	4.5		
G-5	3.5		
G-6	3.7		
G-7	4.5		
G-8	4.3		
G-9	3.7		
G-10	3.7		
G-11	4		
G-12	4		

조사 결과 본 연구에서 제시한 위험도 측정 모델의 타당성 검토(V 지표)에 대한 전체 평균은 4.2, 표준 편차는 0.64, Cronbach α값은 0.837로 높은 수준의 타당성과 신뢰도를 나타내었다. 다만, '데이터 이용환경(V-3)'에서 제3자 제공 시 외부환경에 따른 재식별 위험도 분석도 포함되었으면 좋겠다는 의견이 있었다.

이용기관의 신뢰수준(E 지표)의 경우 전체 평균 3.7로 평균을 상회하였으나 표준 편차가 1.749로 설문 응답자의 분야별로 큰 차이가 있었으며 Cronbach α값은 0.849로 높은 수준의 신뢰도를 나타내었다. 다만, '최근 3년 이내에 법령에 대한 위법사실(E-3)'의 경우는 큰 회사일수록 크고 작은 이슈에 노출될 가능성이 높기 때문에 일관된 수준의 측정이 어렵다는 의견과 '접근 가능한 정보가 있거나 인터넷 상에 존재할 수 있다는 사실만으로 위험(E-7)'을 가정하기엔 기존 법령상에서 규정하고 있는 조항들이 있으므로 너무 높은 가중치로 관리되는 것은 지양해야 한다는 의견이 있었다.

재식별 시 정보주체에 미치는 영향(F 지표)의 경우 전체 평균 3.95, 표준 편차 1.022, Cronbach α값은 0.904로 평균 이상의 타당성과 높은 수준의 신뢰도를 나타내었다. 다만, 전반적으로 재식별되었을 때 경제적 수준에 대한 가치적 판단이 다르기 때문에 평가하기가 어렵다는 의견이 있었다.

마지막으로 데이터 자체에 대한 영향(G 지표)의 경우 전체 평균 3.9, 표준 편차 1.25, Cronbach α값 0.922로 전체적으로 높은 타당성과 신뢰도를 나타내었지만 데이터의 통계적 특성, 최신성, 시간적 특성을 어떻게 판단할 것인지 잘 모르겠다는 의견이 있었으며, 응답자의 인터뷰 결과 원본정보에 대한 고려 없이 가명정보만을 대상으로 판단을 하여 응답에 오류가 있었다는 것을 파악하였다.

델파이 1차 조사 이후 공공기관 실무자 4명, 학계 1명, 기업 3명, 변호사 3명, 비식별 컨설턴트 1명 등을 대상으로 3시간에 걸친 표적집단면접(FGI)을 실시[3]하였으며, 면접에 참여한 전문가 모두 절차에 대한 타당성은 긍정적인 의견이 도출되었다. 다만, 계량지표의 경우 5점 척도로 측정하는 부분에 대한 점수체계의 기준이 제공되어야 한다는 의견과 이용기관의 신뢰수준을 어떻게 측정할 것인가에 대해 다양한 이견이 제시되었다. 이를 위한 해결 방법으로 '사전 자료요구서' 등을 활용하여 관리적인 절차로 보완해야한다는 결론이 도출되었다.

지금까지의 1차 델파이조사와 표적집단면접(FGI) 결과를 반영하여 지표를 보완(중복 및 영향도가 낮은 항목 삭제)하고, 모집단 수를 50명으로 확대[4]하여 2차 델파이조사를 실시한 결과는 Table. 9와 같다.

Table 9. Delphi 2nd Investigation Results

variable	Mean	SD	Cronbach α
V-1	4.46	0.705	0.765
V-2	4.08		
V-3	4.08		
V-4	4.13		
E-1	4.42	0.609	0.667
E-2	4.38		
E-3	4.33		
E-4	4.79		
E-5	4.13		
E-6	-		
E-7	4.08		
F-1	4.42	0.816	0.791
F-2	4.54		
F-3	-		
F-4	4.25		
G-1	-	0.846	0.753
G-2	3.63		
G-3	3.75		
G-4	3.38		
G-5	4.08		
G-6	3.33		
G-7	4.00		
G-8	4.54		
G-9	3.42		
G-10	3.96		
G-11	4.17		
G-12	3.92		

3) `21. 9. 14., Zoom을 이용한 온라인 회의

4) 모집단 수가 적은 이유로 개인정보 보호법 개정이 시행된 지 1년밖에 되지 않았으며, 아직 데이터 활용 활성화가 미진하여 관련 전문가가 타분야에 비해 상대적으로 적음

2차 조사 결과 1차 조사에 비해 타당성에 대한 전체적인 평균점수가 높게 도출되었으며, 표준편차 역시 큰 차이를 보이지 않았다. 조사 표본의 증가로 Cronbach α값 평균은 0.744로 준수한 수준의 신뢰도를 나타내었다.

Ⅴ. Conclusions

지금까지 델파이 1차 조사와 표적집단면접(FGI), 델파이 2차 조사를 통해 본 연구에서 제안한 데이터 상황기반의 위험도 측정방법에 대한 타당성과 영향도를 측정하였다. 위험도 측정을 위한 세부 변수 구분의 타당성에 대해서는 1차 조사평균 4.2점과 2차 조사평균 4.1로 비교적 구분에 대한 타당성이 높은 것으로 조사되었다. 다만, B.데이터 이용환경 중 B.2 이용기관의 신뢰수준의 경우 이용기관에 대한 현황 파악이 어렵기 때문에 사전에 현황파악을 위한 질문지를 보내 응답을 받고 현황을 파악할 수 있는 관리적 절차가 추가되어야 할 필요가 있다.

다음으로 B. 데이터 이용환경의 세부지표별 영향도 측정의 경우 델파이 1차와 2차 조사 모두 인터넷 상에 제공받은 정보와 결합 가능한 데이터가 존재할 수 있다는 항목이 가장 낮은 영향도를 나타내었으며, 데이터 이용자와 제공자간의 계약서가 미흡하다는 항목에 대해서도 법률 상 가명정보를 처리하기 위해서는 계약을 체결하도록 제시하고 있어 위험도 측정에 큰 영향은 없을 것이라는 결론을 도출할 수 있었다.

재식별 시 정보주체에 미치는 영향의 경우 표적집단면접(FGI)을 통해 국내에서는 개인정보 보호법을 통해 재식별을 금지하고 있기 때문에 위험도를 측정하지 않아도 된다는 의견이 있었지만 빅데이터가 처리되는 특성상 국외로 이전 또는 제공될 가능성이 있으며 의도치 않은 재식별이 발생할 경우 이러한 위험도 측정을 통해 데이터를 처리하는 자가 안전하게 검토했다는 법적 증거자료로도 활용될 목적으로 필요성을 유지하였다. 다만, 재식별이 되었을 때 정보주체에게 경제적/비경제적으로 손실을 발생할 수 있다는 기준과 개인에 대한 프라이버시 침해에 대한 수준을 어떤 기준으로 판단할 것인지에 대한 부분은 추가로 연구를 할 필요가 있었다.

마지막 C.데이터 자체에 대한 영향도 측정 결과로는 1차 델파이 조사를 통해 불필요한 항목을 도출할 수 있었으며, 2차 델파이 조사 결과 대부분의 항목에서 보통 이상의 영향도가 있다는 것을 파악할 수 있었다. 위와 같은 조사결과를 반영한 최종 지표의 변경 이력은 Table. 10과 같다.

Table 10. Changes in indicators and final considerations based on survey results

var	final	review
V-3	supplement	Added managerial procedures for measuring user trust level
E-1	supplement	Consolidate similar indicator * Need to establish standards for economic/non-economic benefits
E-2		
E-6	delete	Inability to understand what data users have
E-7	keep	The deletion was considered, but further research on possible combination and personal information infringement was needed
F-1	keep	Deletion was reviewed, but maintained in consideration of overseas environment and legal action in case of re-identification
F-2		
F-4		
F-3	delete	2.3.2 Similar to indicators
C-1	delete	Unidentifiable through encryption when pseudonymization

국외에서는 지금까지 가명정보 활용에 대해 별도의 처리 절차나 기준은 기업 또는 기관의 책임 하에 안전하고 자유로운 활용을 권고하고 있다. 그에 반면 우리나라는 강력한 개인정보 보호의 기반 하에 개인정보를 빅데이터로 활용할 수 있는 기반이 뒤늦게 시작되어 이제야 가명정보 활용에 대한 연구가 다양하게 추진되고 있으나, 아직까지 가명처리 시 고려해야 할 사항에 대한 연구 결과가 거의 없다는 것을 확인할 수 있었다. 가명정보는 추가정보를 사용해서 언제든 개인을 식별할 수 있는 원본정보로 복원을 할 수 있기 때문에 데이터를 처리하는 자 또는 데이터를 제공받는 자 입장에서 재식별 가능성을 검토하여야 한다.

본 연구는 선행연구 등을 통해 이러한 재식별 가능성 등을 검토하기 위한 체크리스트 방식으로 제시되는 유일한 방법론이란 점에서 의의가 있다. 또한 현재 가명정보를 활용하는 기관에서는 데이터를 활용하기 전 외부전문가를 활용한 검증방법을 이용하고 있는데 이러한 위험성을 기존 정성적이 아닌 정량적으로 측정할 수 있다는 점에도 의의가 있다. 본 연구 결과를 통해 가명처리를 하는 실무자들이 가명정보의 위험성을 측정하는데 실무적으로 도움을 줄 수 있을 것이라 판단하며, 향후 보다 안전한 개인정보 활용을 위한 가명처리에 관한 연구가 활발히 진행되기를 바란다.

REFERENCES

[1] Economist, "The worlds most valuable resource is no longer oil, but data," 2017.

[2] N. David, "How to Plan, Participate and Prosper in the Data Economy," Gartner Research, Mar. 2011.
[3] Joint Government Departments in Korea, "Plans to revitalize the data and AI economy," 2019.
[4] IMD, "World Digital Competitiveness Ranking 2021 Report," 2021.
[5] Joint Government Departments in Korea, "Guidelines for De-Identification of Personal Information," 2016.
[6] Personal Information Protection Commission, "Guidelines for processing pseudonym information," 2021.
[7] EU General Data Protection Regulation, "Recital(26)," 2018.
[8] Sweeney L, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Vol. 10, No. 3, pp. 557-570, July. 2002.
[9] ISO/IEC 20889, "Privacy enhancing data deident ification terminology and classification of techniques, Annex A," 2018.
[10] Financial Services Commission, "Guidelines for processing anonymisation and Pseudonymization Information in the financial sector," 2020.
[11] Ministry of Health and welfare, "Guidelines for Health and Medical Data Utilization," 2020.
[12] Ministry of the Interior and Safety, "Guidelines for processing Pseudonymization Information in the public sector," 2021.
[13] Information and Privacy Commissioner of Ontario, "De-Identification Guidelines for Structured Data," 2016.
[14] HITRUST, "De-Identification Framework," 2015.
[15] K. Emam, L. Arbuckle, "Anonymizing health data," O`Reilly book, pp. 29-33, 2013.
[16] M. Elliot, E. Mackey et al, "The Anonymisation Decision making Framework," UK Anonymisation Network, 2016.
[17] F. Prasser, F. Kohlmayer et al, "The Importance of Context: Risk-Based De-Identification of Biomedical Data," Methods of Information in Medicine, Vol. 55, No. 4, pp. 347-355, Aug. 2016.
[18] NIST 800-188(2nd Draft), "De-Identifying Government Datasets," 2016.
[19] D. Tanvi, R. Felix et al, "Five Safes: Designing data access for research,", University of Bristol, 2016.
[20] Dhkim, Sskim, "A New Scheme for Risk Assessment Based on Data Context for De- Identification of Personal Information," Journal of The Korea Institute of Information Security and Cryptology, Vol. 30, No. 4, pp. 719-734, Jun. 2020.
[21] Personal Information Protection Act, "Article 28-2," 2021.
[22] G. Duncan, T. Elliot et al, "Statistical Confidentiality," New York Springer, 2011.

Authors

Dong-Hyun Kim received Ph.D degree in convergence security from Chung-Ang University, Korea, in 2022. He has been conducting personal information surveys and policy improvement for 6 years from 2010, and since 2016, the Data Utilization Support Team has been working to utilize safe personal information as big data. He is interested in Personal Information Security, De-Identification & De-Identified Information Risk Measure.