Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.4, Aug. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.4.719

개인정보 비식별 조치를 위한 데이터 상황 기반의 위험도 측정에 관한 새로운 방법

김 동 현,[1†] 김 순 석[2‡]
[1]중앙대학교(대학원생), [2]한라대학교(교수)

A New Scheme for Risk Assessment Based on Data Context for De-Identification of Personal Information

Dong-hyun Kim,[1†] Soon-seok Kim[2‡]
[1]Chungang University(Graduate student), [2]Halla University(Professor)

요 약

본 논문은 최근 개정된 데이터 3법에 맞추어 조직 내 실무자들이 개인정보를 활용함에 있어 비식별 조치 수행 시 위험도에 따른 처리 수준 산정을 위한 새로운 측정방법을 제안한 것이다. 우리가 제안한 방법은 위험도 측정 시 데이터만이 아닌 데이터를 둘러싼 주위 상황을 고려하였고, 모든 분야에 적용이 가능하도록 범용 환경에서 데이터 상황을 크게 데이터 활용방법, 데이터 이용환경, 그리고 데이터(자체) 3가지 카테고리로 나누어 보다 체계적으로 분류하였으며, 제시된 분류에 따라 각 상황별 위험도에 기반하여 정량적으로 계산할 수 있도록 새로운 측정 방법을 제안하였다. 제안한 방법은 기존의 비식별 정보의 위험도 산정을 전문가들의 정성적 판단에만 맡기는 것이 아니라 일반 조직 내 개인정보처리자가 실무에 활용할 수 있도록 정량적인 방법으로 위험도를 산정할 수 있도록 설계하였다.

ABSTRACT

This paper proposes a new measurement scheme for estimating the processing level according to risk when performing de-identification in the use of personal information by practitioners in the organization in line with the recently revised Data 3 Act. Our proposed methods considered the surrounding circumstances surrounding the data, not just the data, for risk measurement, and divided the data situation into three categories more systematically so that it can be applied in all areas in a general-purpose environment, the data utilization environment, and the data (self) so that it can be calculated quantitatively based on each context risk according to the presented classification. The proposed method is designed to calculate the risk of existing de-identifiable information in a quantitative manner so that personal information controller in general organizations can use it in practice, not just in the qualitative judgment of experts.

Keywords: De-identification, Personal Information, Data Context, Risk Assessment

I. 서 론

2020년 2월 4일 개인정보보호법 등 이른바 데이터 3법이 개정되었다. 개인정보의 비식별(De-identification)은 데이터 3법 개정과 함께 개인정보의 보호 관점이 아닌 활용 관점에서 최근 들어 이슈가 되고 있다.

비식별은 주어진 원본 데이터셋(Dataset)에 대해 개인정보처리자 또는 제3자로부터 데이터 주체가 누구인지를 알아볼 수 없도록 처리하는 것을 말하며

Received(04. 16. 2020), Modified(06. 25. 2020), Accepted(06. 25. 2020)
† 주저자, kdonghyun@kisa.or.kr
‡ 교신저자, sskim@halla.ac.kr(Corresponding author)

이를 세분화하면 크게 가명처리와 익명처리로 나뉜다. 익명처리는 가명처리의 한 부분이긴 하지만 가명처리만으로는 충분한 익명처리로 보기 어렵기 때문에 추가적인 비식별 조치가 필요하다. Elliot 등[1]은 지난 2015년 익명처리에 대한 유형을 형식적 익명처리(Formal Anonymisation), 보장된 익명처리(Guaranteed Anonymisation), 통계적 익명처리(Statistical Anonymisation), 그리고 기능적 익명처리(Functional Anonymisation) 4가지로 분류한 바 있다. 형식적 익명처리는 어떤 형태로든 주어진 데이터셋에서 이름과 같은 직접식별자를 제거하거나 가리는 것을 말한다. 보장된 익명처리는 확실히 보장되고 복구가 불가능하게 처리하는 방법으로 해당 데이터셋 내에서 식별될 위험이 어떠한 가정을 하더라도 사실상 제로여야 한다. 그러나 우리가 유용한 데이터를 공유하고자 할 때 일반적으로 보장된 익명처리는 이론적으로 가능하지 않을뿐더러 실용적이지 않다. 통계적 익명처리의 개념은 통계적 노출 제어(Statistical Disclosure Control, SDC)라는 기술 분야와 관계가 있다. SDC의 기본적인 원리는 재식별의 가능성을 제로로 만드는 것이 불가능하기 때문에 그 대신 노출 상황의 위험을 통제하거나 제한할 필요가 있다는 것이다. 형식적 익명처리와 보장된 익명처리는 모두 단순히 통계적 익명처리의 특별한 경우라고 말할 수 있다. 형식적 익명처리는 재식별화의 가능성을 1 이하로 낮추는 메커니즘이고, 보장된 익명처리는 이 가능성을 0으로 만들기 위한 메커니즘이다. 그러나 개인과 관련된 데이터를 공유하고 배포하려는 사람의 목표는 두 가지이다. 첫째는 유용한 데이터를 공유하고 배포하는 것과 둘째는 이 데이터의 기밀성과 프라이버시를 유지하는 형태로 만드는 것이다. 형식적 익명처리는 후자의 목적을 달성할 수 없으며, 보장된 익명처리는 전자의 목적을 달성할 수 없다. 통계적 익명처리는 이 양 극단 사이에 여러 입장이 있음을 인정하는 것으로 프라이버시 보호 모델에서 말하는 k-익명성 모델[2]이 대표적인 통계적 익명처리 방법에 해당한다. 끝으로 네 번째 유형의 익명처리는 기능적 익명처리이다. 기능적 익명처리를 논하기에 앞서 영국의 비식별 전문가그룹인 UKAN(UK Anonymisation Network)에서 발간한 익명처리 의사결정 프레임워크에서는 위험을 결정하는 요인에 대해 아래와 같이 6가지로 분류하고 있다[3].

- 익명처리된 데이터에 포함된 개인을 재식별 하려는 공격 동기(이 경우 문제가 되는 것은 ‘ 무엇이, 어떻게 일어났는가?’일 것이다)
- 노출(disclosure)에 따른 결과(이로 인해 개인이 재식별화를 시도하도록 하는 동기에 영향을 미칠 수 있다)
- 악의적인 의도 없이(이를 자연스러운 식별(spontaneous identification)이슈라 부른다) 노출이 발생하는 경우
- 위험 결정 요인에 영향을 줄 수 있는 데이터 접근 관리를 위한 정부의 절차, 데이터 보안 및 다른 인프라들
- 해당 데이터와 연관이 있을 수 있는(데이터가 비식별화 되면 노출/식별화를 위해 반드시 필요한) 다른 데이터/지식
- 해당 데이터와 다른 데이터/지식의 차이, 주로 데이터 차이(data divergence)라고 한다.

통계적 익명처리에 이러한 요인들을 고려하여 생성된 것이 네 번째 익명처리 유형인 기능적 익명처리이다. 기능적 익명처리에서는 Mackey와 Elliot[4]이 데이터 환경이라고 통틀어 말하는 상황(context)에 따른 요인을 다루고 있으며, 이 유형이 UKAN의 익명처리 의사결정 프레임워크에서 주장하는 바로 그 개념이다. UKAN에서 정의한 데이터 상황에 대한 정의는 다음과 같다.

데이터 상황은 데이터와 데이터 환경과의 관계를 가리키는 말로 고정 데이터 상황(데이터 공개 이후)과 유동 데이터 상황(데이터 공유나 혹은 공개 시)으로 나뉘며, 이때 데이터 환경은 다시 데이터, 행위자, 관리절차, 인프라의 4가지 구성요소로 이루어진다. 이것은 다시 환경중심솔루션 부분에서도 언급되고 있는데 데이터 환경을 고려하지 않고는 데이터가 익명처리 되었는지의 여부를 판단할 수 없다는 것이다. 즉, 환경 제어를 통해 데이터 자체를 제어하는 것으로 효과적인 데이터 익명처리가 가능하다는 것이다.

비식별 처리에 있어 이러한 데이터 상황에 대한 접근법은 2011년 영국 Duncan 등[5], 2015년 미국의 산업계 표준인 HITRUST와 Privacy Analytics사가 공동 개발한 맥락(Context)기반 위험도 측정 방법론[6,7,14,15], 그리고 2016년 미국 국립표준연구소에서 발간한 NIST SP 800-188(2nd Draft) 정부 데이터셋의 비식별화

(De-Identifying Government Datasets)[8]에서도 소개되고 있다. 한편 지난 2016년 6개 정부부처합동으로 발간된 개인정보 비식별 조치 가이드라인[9]에서도 비식별 조치된 정보의 적정성 평가시 데이터 상황과 유사한 재식별시 시도가능성과 재식별시 영향분석을 일부 고려하고 있다. 그러나 이러한 데이터 상황기반의 접근법에 있어 가장 중요한 요구사항은 첫째, 그 나라의 법과 제도적인 요소에 맞추어 적용되어야 하며 둘째, 특정분야에 치우치지 않고 범용성 있게 두루 적용될 수 있어야 한다는 것이다. 전자의 경우에 있어, 우리나라 가이드라인의 경우는 비록 법과 제도적인 요소가 고려되었다고는 하나 최근 개정된 데이터 3법과는 맞지 않을뿐더러 데이터 상황에 따른 전체 구성요소들 가운데 극히 일부만 고려되었다. 후자의 경우, 미국 HITRUST와 Privacy analytics사에서는 데이터 상황에 대한 접근법을 구체적으로 제시하고 있으나 우리나라의 법·제도나 상황에 맞지 않고 특히, 의료분야에 한정되어 제시하고 있다는 점이 단점이다. 이는 바꿔 말해 통신, 유통, 금융, 교육 등 전 분야에 적용할 수 없다는 것이다. 그 외 데이터 상황기반의 접근법들[4,5,8]은 모두 고려사항 정도로 제시하고 있어 내용이 구체적이지 않다.

그동안 일반 개인정보 보호의 위험을 식별하고 평가하는데 있어 학술적인 측면에서 상황(context)을 고려한 몇몇 연구들이 있어왔다. Nissenbaum[10]은 프레임워크를 통해 상황을 고려한 위험평가에 관한 개념적 메타 모델을 제안한 바 있으며, Bieker 등[11]은 상황을 프라이버시 위험 평가 프로세스에 포함시키려고 함으로써 프라이버시 위험 평가에서 상황을 포함하는 것의 중요성을 보여준 바 있다. Mulligan 등[12]은 피해 위험 측면에서 프라이버시 보호 위험을 조사하였고, Solove[13]는 프라이버시가 침해되거나 침해될 수 있는 모든 영역의 모든 측면을 고려하여 위험 평가에 관한 새로운 모델을 제안한 바 있다. 그러나 이러한 접근들은 일반 개인정보 보호의 관점에서 개인정보의 영향을 평가하는데는 유용하지만 개인정보보호 중 비식별 조치관점에서 특화되어 평가하기에는 한계가 있다. 한편 이와 달리 학술적 연구로 개인정보보호에 대한 위험 평가를 비식별조치 관점에서 측정하고자 한 시도들도 있어왔다. 그러나 이러한 시도들은 대부분 의료분야에 치중되어 범용적으로 적용하기에는 한계가 있다. 대표적인 예로 Emam[6,14,15]은 보건의료 데이터의 비식별 조치를 위험도와 상황에 기반하여 측정하고 평가하기 위한 측정 방법론을 제안한 바 있다. Emam은 앞서 언급한 Privacy analytics사의 CEO이기도 하다. Emam이 제안한 방법론은 2장 관련연구의 2.4 데이터 맥락(context)기반 위험도 측정 부분에서 보다 자세히 다루고자 한다. Prasser 등[16]은 생의학(biomedical) 데이터에 대해 전반적인 데이터 상황이 아닌 공격자 모델에 기반하여 위험도를 측정하고 평가를 시도한 바 있다. 공격자 모델은 검사나 기자, 마케터 등의 공격자의 시나리오를 기반으로 위험도를 측정하는 것으로 이 또한 데이터 중심이며 데이터 활용 목적이나 데이터의 민감도, 그밖에 데이터 공개 환경 등 데이터 주변의 상황을 전반적으로 다루고 있지는 않으며 생의학 데이터의 특성을 다른 분야에 범용적으로 적용하기에도 한계가 있다. 한편 Tomashchuk 등[17]은 의료분야가 아닌 일반 범용분야에 있어 처음으로 프라이버시 보호 요건을 준수하면서도 데이터의 유용성을 높이고자 하는 시도를 한 바 있다. 그러나 프라이버시 보호 요건으로 k-익명성의 최소 k값을 한계값으로 사용자가 미리 정하도록 하고 있어 데이터 주변의 상황을 고려하는 것이 아닌 오로지 데이터에만 한정하여 평가하고 있다는 한계가 있다.

따라서 본 논문에서는 이러한 단점들을 개선하고자 최근 개정된 우리나라의 데이터 3법에 따르면서 데이터 상황 접근법에 따라 각 상황을 우리나라의 환경에 맞추어 체계적으로 분류하고, 분류된 각 구성요소들을 위험도에 기반하여 정량화하여 측정할 수 있는 새로운 방법을 제시하고자 한다.

본 논문의 2장에서는 데이터 상황기반 접근법과 관련하여 기존에 제시된 방법들에 대해 살펴보고 3장에서는 우리나라 상황에 맞는 새로운 위험도 기반의 정량적 측정 방법을 제시, 분석한 후, 4장을 끝으로 결론을 맺고자 한다.

II. 관련연구

2.1 Duncan 등[5]이 제안한 환경제어

이 방법은 앞서 말한 UKAN의 익명처리 의사결정 프레임워크에서도 설명되고 있는 방법으로 데이터 익명처리에 대한 여부를 환경 제어를 통한 데이터 제어로 처리가 가능하다는 주장이다. 즉, 환경 제어를 다음과 같이 '누가(who)', '무엇을(what)', '어디서

(where), 어떻게(how)'라는 질문의 답을 통해 특징지어지는 것으로 정의하고 있다. 첫째, 누가(who) 데이터에 접근할 수 있는가? 둘째, 무슨(what) 분석이 이뤄지거나 이뤄지지 않는가? 셋째, 어디서(where) 데이터 접근/분석이 이뤄지는가? 그리고 어떻게(how) 접근하는가?

그러나 이 방법은 단순하기는 하나 구체적이지가 않으며 실제 데이터를 둘러싼 모든 상황을 고려하기에는 여전히 부족한 면이 많다.

본 논문에서는 Duncan 등[5]이 제안한 4가지의 환경제어를 그동안 우리들이 산업 현장에서 쌓은 경험과 요구사항들을 토대로 보다 다양한 상황을 고려하여 1. 데이터 활용방법으로 카테고리화하고 7가지(왜, 무엇을, 어디서(처리관점, 장소관점), 어떻게(공개관점, 제공관점, 이용관점)으로 확장하여 제시하고자 한다.

2.2 다섯안전(Five Safes)[8]

한편 미국의 경우 지난 2016년 12월에 발표된 표준 문건, NIST SP 800-188(2nd Draft) 정부 데이터셋의 비식별화(De-Identifying Government Datasets)를 통해 영국의 데이터 상황과 유사한 다섯 안전(Five Safes)이라는 개념을 소개하고 있다. 다섯 안전은 말 그대로 다섯 개의 안전을 위한 위험(혹은 접근) 범위를 제시한 것으로, 국가 통계로부터의 데이터를 연구 커뮤니티와 공유, 접근, 평가하기 위해 설계된 일반 프레임워크이다. 다섯 안전의 내용은 다음과 같다. 첫째, Safe projects - 자료의 사용은 적절합니까? 둘째, Safe people - 연구자들이 적절한 방법으로 그것을 사용한다고 신뢰받을 수 있습니까? 셋째, Safe data - 자료 자체에 공개 위험이 있습니까? 넷째, Safe settings - 접근 시설이 승인되지 않은 사용을 제한합니까? 다섯째, Safe outputs - 통계적 결과는 비폭로적(non-disclosive)입니까?

이러한 다섯 안전의 장점은 EU 개인정보보호법인 GDPR에서 정의한 데이터 컨트롤러로 하여금 단지 데이터만이 아닌 데이터 접근이나 평가 시 데이터 공개에 대한 여러 가지 다른 측면들을 고려하도록 강제할 수 있다는 점이다. 이러한 이유로 미국에서는 표준 문건 제정을 통해 정부 기관들이 자료 공개시 위험 분석을 체계화하여 이용할 수 있도록 하고 있다. 그러나 이 또한 고려사항일 뿐 구체적이지 못하다.

2.3 우리나라 개인정보 비식별 조치 가이드라인[9]

지난 2016년 6월 행정안전부를 비롯한 6개 정부부처합동으로 발간한 개인정보 비식별조치 가이드라인에서도 데이터 상황과 관련된 일부 고려사항들을 찾아볼 수 있다. 비식별 조치 이후 적정성 평가 시 데이터를 이용 또는 제공받는 자의 재식별 의도와 능력, 개인정보 보호 수준 등 재식별 시도 가능성을 분석하고 데이터가 의도적 또는 비의도적으로 재식별될 경우 정보주체 등에게 미칠 수 있는 영향을 분석하는 것이 바로 그것이다. 이 부분에서 가장 큰 오류는 이러한 고려사항들이 데이터에 대한 비식별 처리가 끝난 이후 적정성 평가 시에 진행된다는 점이다. 다시 말해 이러한 사항들은 평가 시가 아니라 비식별 처리 이전에 고려가 되어 그 가능성과 영향도에 따라 비식별 처리의 강도를 결정하고 처리가 이루어져야 한다는 것이다. 두 번째 오류는 상기 2가지 사항만으로는 데이터를 둘러싼 환경적인 요소들을 모두 고려하기엔 충분치 않다는 점이다. 즉, 데이터를 사용하려는 목적, 용도, 사용주기, 제공 상황 등 여러 요소들이 함께 고려될 필요가 있다.

따라서 본 논문에서는 기존 가이드라인의 단점을 극복하고자 가이드라인에서 제안한 데이터 이용환경에 관한 오류들을 수정하여 2. 데이터 이용환경으로 카테고리화하고 이를 확장한 새로운 방법론을 제안하고자 한다. 한편 기존 가이드라인과 제안 방법론과의 구체적인 차이점은 Table 3에서 제시하였다.

2.4 데이터 맥락(context)기반 위험도 측정 [6,7,14,15]

현재까지 알려진 비식별 처리 방법 중 데이터의 상황을 고려한 가장 고도화된 방법이 미국 산업계 표준으로 자리 잡고 있는 HITRUST와 Privacy Analytic사가 공동 개발한 맥락(Context)기반 접근법이다. 이 접근법은 미국 Privacy Analytics사의 CEO이자 오타와 대학 교수인 Khaled El Emam이 처음 제안한 것으로 현재 미국 HIPAA 프라이버시 규칙[18]의 의료분야 데이터 비식별을 위한 전문가 결정 방법에 주로 이용되고 있다. 그가 제안한 방법론에 따르면 비식별 조치 전 데이터의 중요도, 환자에 대한 잠재적 상해나 피해 정도, 그리고 데이터 공개 승인에 대한 타당성을 정성적으로 평가하여 선례를 참고로 위험에 대한 임계치를 산정한다.

그러나 이 방법은 첫째, 측정 자체가 정성적이고 둘째, 의료분야의 비식별에 최적화되어 있으며 셋째, 우리나라 비식별 조치 가이드라인과 마찬가지로 상기 사항들만으로는 데이터를 둘러싼 환경적인 요소들을 모두 고려하기엔 충분치 않다. 다시 말해 우리나라의 법과 제도, 데이터를 사용하려는 목적, 용도, 사용주기 등 여러 요소들이 함께 고려될 필요가 있다. 한편 비식별 조치 이후에는 데이터 상황을 고려한 위험도를 크게 데이터 자체에 대한 위험도와 맥락에 따른 위험도 2개의 부류로 나누어 정량적으로 측정하고 있다. 여기서 최종 위험도는 상기 2개의 위험도를 곱한 값으로 계산하며 각각의 위험도 계산은 다음과 같다. 첫째, 데이터 자체에 대한 위험도는 k-익명성 [2] 프라이버시 보호 모델에서 말하는 최대 재식별률 혹은 평균 재식별률로 계산한다. 둘째, 맥락 즉, 이용환경에 따른 위험도는 다시 고의적인 시도에 대한 발생 가능성, 의도치 않은 시도에 대한 발생 가능성, 그리고 위반할 가능성 3가지로 나뉘며, 위험도는 이들 중 최대값으로 계산한다.

우선 전자로 데이터 자체에 대한 위험도를 측정하는 데 있어 문제점은 측정 도구를 k-익명성 지표에만 한정하고 있다는 점이다. 예컨대 비식별 조치 시 k-익명성 모델을 적용하지 않을 경우는 측정 자체가 불가하다. 따라서 데이터 자체에 대한 위험도는 이러한 상황을 감안하여 데이터 구성이나 분포, 그리고 데이터에 대한 민감도를 함께 측정할 필요가 있다. 본 논문에서는 이를 3. 데이터(자체)로 카테고리화하여 보다 확장된 새로운 측정 방법론을 제안하고자 한다. 후자로 맥락 즉, 이용환경에 따른 위험도 측정 방법에 있어서의 문제점은 의료분야에 지나치게 편중되어 범용적으로 사용하기에는 한계가 있다는 점이다. 예컨대 위반할 가능성 측정의 경우 미 건강 관리 정보 및 관리 시스템 협회(HIMSS)가 지난 2012년에 시행한 설문조사 결과를 기반으로 측정하고 있다. 따라서 이용환경에 따른 위험도 측정은 범용 분야에 일반적으로 적용하기에는 한계가 있다.

III. 제안하는 데이터 상황기반 위험도 측정방법

3.1 데이터 상황에 대한 분류

우리가 제안하는 위험도 측정 방법은 먼저 데이터 상황에 대한 체계적이고 명확한 분류에서 시작된다. 우리는 우리나라의 개정된 법과 제도 그리고 이를 둘러싼 다양한 상황을 고려하여 크게 데이터 활용방법(7가지 관점), 데이터 이용 환경(2가지 관점), 그리고 데이터 자체에 대한 위험도(3가지 관점) 3가지 카테고리, 총 12가지 관점으로 나뉘어 Fig.1과 같이 분류하였다.

이러한 분류 기준에 대한 기반은 앞서 관련연구에

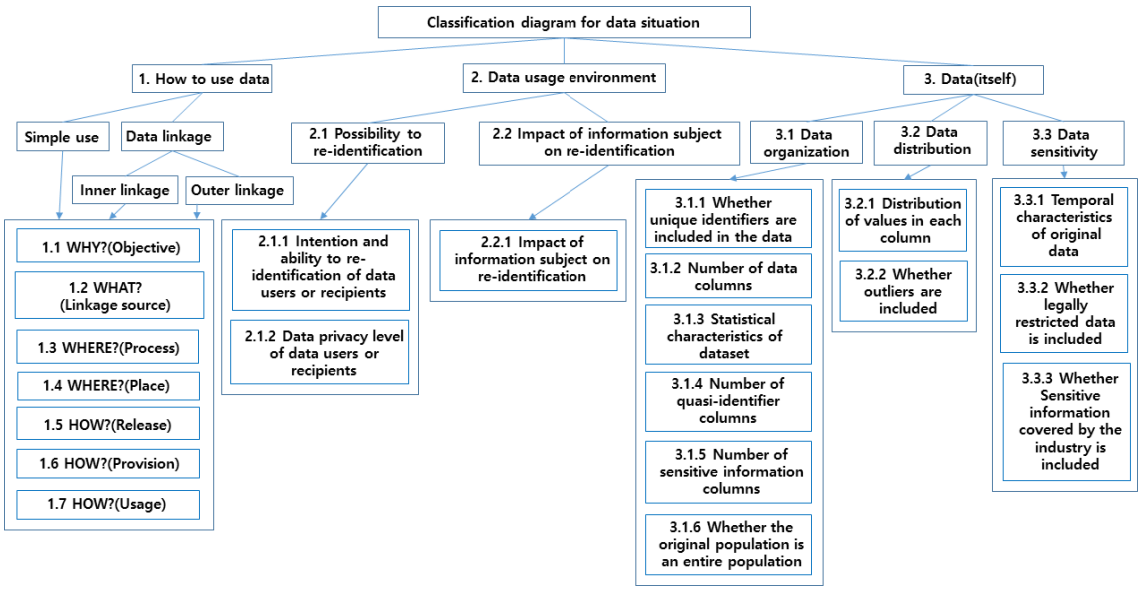

Fig. 1. The classification of data situation

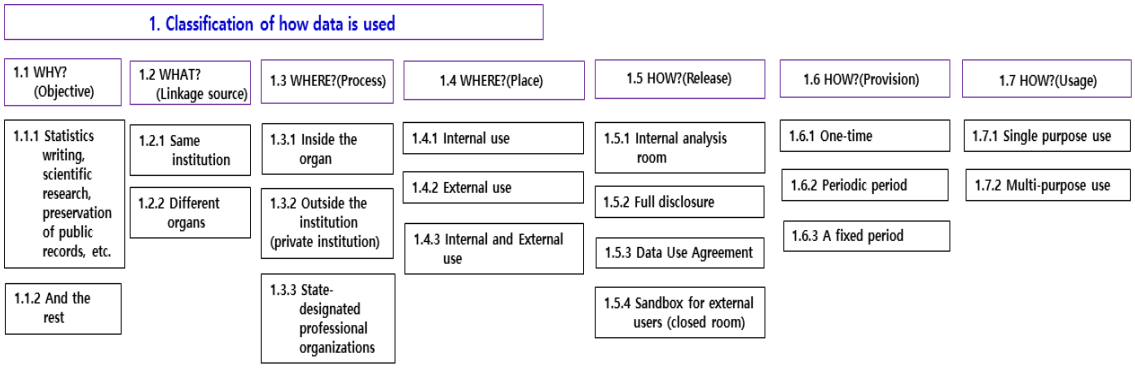

Fig. 2. The classification of how data is used

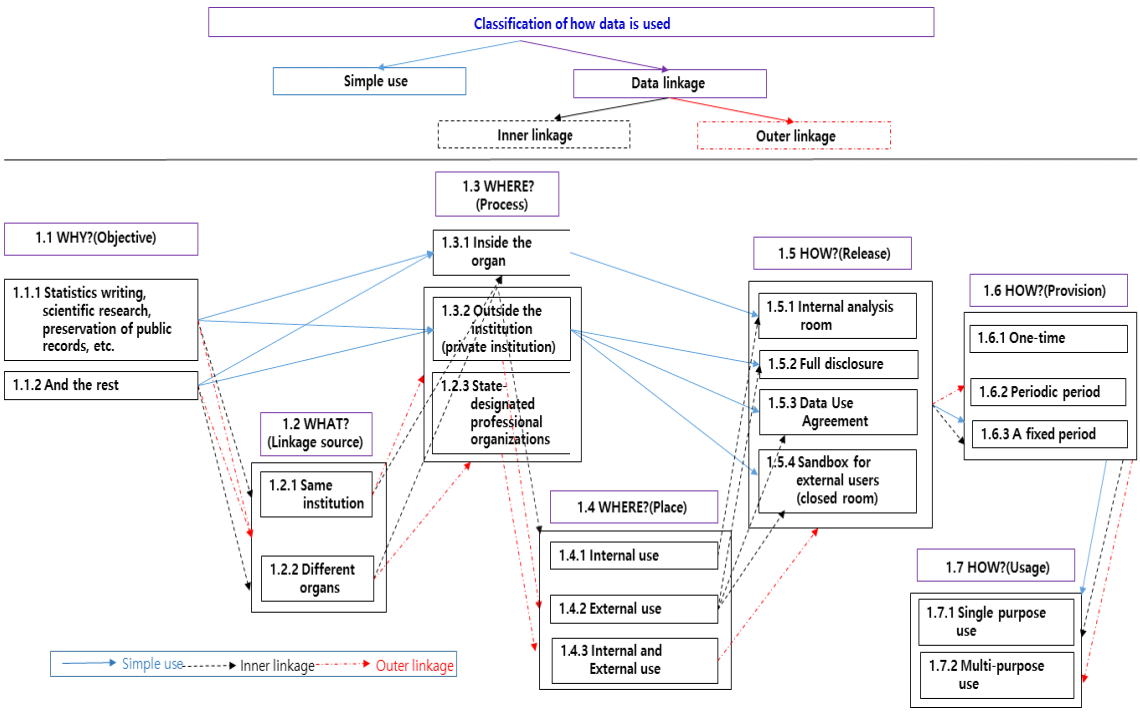

Fig. 3. Flow chart on how to use data

서 언급한 Duncan 등[5]의 환경제어와 우리나라 가이드라인[9], 그리고 미국의 데이터 맥락(context)기반 위험도 측정 방법[6,7,14,15]을 지난 2016년 가이드라인이 공표된 이후 지금까지 현업에서 비식별 조치에 대한 적성성 평가를 실제로 경험하면서 쌓은 노하우들과 업계의 요구사항들을 토대로 수정 확장한 것이다.

먼저 데이터 활용방법은 아래와 같이 분류된다. 단순사용, 그리고 데이터 결합(내부결합, 외부결합)에 따라, 1) 왜(Why) 비식별 조치를 하고자 하는가? 즉, 그 활용목적이 무엇인가?, 2) 무엇을(What?)(원천관점) 데이터 결합 시 동일기관 내 데이터를 가지고 결합을 하려는가? 아니면 서로 다른 기관의 데이터를 가지고 결합을 하려는가?, 3) 어디서(Where?)(장소관점) 데이터를 활용하려는 장소가 내부, 외부 혹은 모두 사용 등 어디에서 이루어지는가?, 4) 어디서(Where)(처리 관점) 데이터에 대한 비식별 조치가 어디(기관 내부, 외부, 전문기관)에서 이루어지는가?, 5) 어떻게(How)(3가지 관점으로 세분화됨) (① 데이터 공개관점) 데이터 접

Table 1. The Classification for data usage

1	How to use data			Basis
	1.1	Why?(Objective-Depending on the purpose of the data)		Expansion of environme-ntal control classificati-on proposed by Duncan et al. [5] (4 -> 7)
		1.1.1	The purpose of statistical preparation, scientific research, and preservation of public records in accordance with Paragraph 1 of Article 28-2(processing of pseudonym information, etc.) of the Personal Information Protection Act of Korea	
		1.1.2	1.1.1 Other purposes	
	1.2	What?(Linkage source)		
		1.2.1	When the source data to be combined is the same institution	
		1.2.2	When the source data to be combined is from different institutions	
	1.3	Where?(Process - From the point of view of de-identifying data)		
		1.3.1	De-identification within the institution(company)	
		1.3.2	De-identification of data outside the institution(company) (eg, as a private specialized institution, which is not possible under the current legislative decree, but can be introduced in the future)	
		1.3.3	De-identification of data by the Korea Personal Information Protection Committee or a specialized agency designated by the relevant ministries in accordance with the current enacted enforcement ordinance.	
	1.4	Where?(Place)		
		1.4.1	When the place to use the data is used only within the institution or the company	
		1.4.2	This applies when the place where the data is to be used is outside the institution or company (for example, a private specialized institution).	
		1.4.3	Both 1.4.1 and 1.4.2	
	1.5	How?(Release-In terms of data disclosure, depending on how the data will be released after the data has been de-identified)		
		1.5.1	Disclosures are only made within the company's or institution's own analysis room	
		1.5.2	Full disclosure to the general public	
		1.5.3	Disclosure by mutual agreement between the data provider and the data user.	
		1.5.4	Disclosure only within sandboxes with security facilities for external users	
	1.6	How?(Provision)		
		1.6.1	When data provision is one-time	
		1.6.2	When data is provided periodically (monthly, quarterly, semi-annually, etc.)	
		1.6.3	When data is provided for a predetermined period	
	1.7	How?(Usage)		
		1.7.1	When the purpose of using data is a single purpose	
		1.7.2	When the purpose of using data is multi-purpose	

Table 2. The Classification for data usage environment

2	Data usage environment		Basis
	2.1	Possibility of re-identification attempt: the intention and ability to re-identify data users or receiving organizations or companies and the level of personal information protection	Realistically expand the guidelines in Korea [9] and reflect provisions regarding the pseudonymization of the revised Personal Information Protection Act
		2.1.1 Refers to the intention of re-identification of data users or receiving organizations or companies, the ability to re-identify, and the possibility of linking with external information	
		2.1.2 Refers to the ability to protect the personal information of data users or receiving organizations or companies	
	2.2	Impact on information subject when re-identified: refers to the effect on data subject when data is re-identified intentionally or unintentionally	
		2.2.1 Same as above 2.2	

근/분석이 이루어지는가?, (② 데이터 제공관점) 어느 기간 동안 데이터 제공이 이루어지는가? (③ 데이터 활용관점) 활용목적 관점에서 단일, 다용도 혹은 계약상 등 어떤 목적으로 데이터를 활용하려는가? 이를 보다 세부적으로 분류하면 모두 19가지의 상황으로 분류할 수 있다(Fig.2, Table 1 참조).

한편 위 데이터 활용 방법을 단순사용일 경우와 데이터 결합(기관 또는 기업 내부에서의 내부결합과 외부에서의 외부결합)으로 분류하여 순서에 따른 흐름도를 기술해보면 Fig.3과 같다.

둘째로 Fig.1에 따라 2. 데이터 이용 환경

(Table 2 참조)과 3. 데이터(자체)(Table 5 참조)에 대한 세부적인 분류체계는 아래와 같다.

Table 2의 재식별 시도가능성과 재식별시 정보주체에게 미치는 영향 부분은 우리나라 개인정보 비식별 조치 가이드라인에서의 적정성 평가 시 사용하는 방법과 매우 유사하지만 다음과 같은 차이가 있다(Table 3 참조). 참고로 Table 4에서 밑줄로 표기된 부분이 기존 가이드라인[9] 지표를 보다 현실적으로 개선한 부분이다. 데이터(자체)에 대한 분류에 있어 제안하는 방법은 기존 개인정보 비식별 조치 가이드라인을 개선한 것으로 아래와 같은 차이가 있다

(Table 6 참조). Table 7은 데이터(자체) 분류에 있어 데이터 구성 부분에 대한 측정 일부를 예시로 든 것이다.

3.2 데이터 상황 분류에 따른 위험도 측정방법

상기 3.1의 분류도에 따른 위험도 측정 방법은 Table 8과 같다.

데이터 활용방법에 있어 단순사용, 내부결합 및 외부결합 사용 등 각 경우의 위험도에 따라 각 단계별로 매우 낮음(Very low)(1점), 낮음(Low)(2점), 보통(Normal)(3점), 높음(High)(4점), 매우 높음(Very High)(5점) 부여 후 점수를 합산하여 처리자가 반영하고, 데이터 이용환경의 경우 가명처리와 익명처리의 경우로 나뉘며 체크리스트를 이용, 5점 척도의 경우 1~5점을 부여하고 예/아니오의 경우 각각 5점과 1점을 부여 후 점수를 합산하여 처리자가 반영한다, 데이터(자체)의 경우 가명 및 익명처리 모두 동일하며 체크리스트를 이용, 5점 척도의 경우 1~5점을 부여하고 예/아니오의 경우 각각 5점과 1점을 부여 후 점수를 합산하여 처리자가 반영한다. Table 8에 따른 측정 항목과 반영비율에 따라 합산 점수가 가명처리의 경우 3단계(Level 1(52점 미만, 29.2%), Level 2(52점 이상~70점 미만, 44.2%), Level 3(70점 이상, 26.6%))로, 익명처리의 경우 5단계(Level 1(42점 미만, 9.8%), Level 2(42점 이상 53점 미만, 21.8%), Level3(53점 이상 68점 미만, 36.8%), Level 4(68점 이상 79점 미만, 21.8%), Level 5(79점 이상, 9.8%)로 최종 평가한다.

아울러 최종 평가 결과에 따라 결정되는 비식별 처리 수준은 Table 9와 같다. 한편 Table 9에 따라 결정된 비식별 처리 수준에 따른 세부적인 처리 방법, 절차, 그리고 처리 결과에 대한 적정성 평가 등에 대한 부분은 현재 후속 연구로 진행 중이며 조만간 결과를 발표할 예정이다.

Table 3. Difference between the guidelines for de-identification of personal information in Korea[9] and the proposed method

Classification #	Guideline in Korea [9]	The proposed scheme
Common	Applicable only for anonymous processing	- Distinguish between pseudonym processing and anonymous processing, that is, apply with different detailed indicators - In the case of anonymization, the indicators in 〈Table 2〉 2.1.2 are excluded. This is because anonymous processing is not subject to protection in accordance with Article 58-2 of the Revised Personal Information Protection Act.
2.1.1	No distinction between the internal and external use of an institution (company).	Classified as separate indicators
2.1.2	All yes / no evaluation by detailed indicator	Only the 2 indicators below are evaluated as yes / no, and the remaining indicators are used on a 5-point scale. 1) There is a possibility that the data user or requester may provide the data to a third user without prior permission 2) The data user or requester does not reflect the phrases such as prohibiting re-identification and restricting data provision to third parties in the contract related to data use (provision).
		The three indicators below are used separately on a 5-point scale. 1) Operates according to the management plan for storage and processing of data 2) Data is provided or provided in a secure manner with physical and technical protection measures in place. 3) Authorization and access history of personnel who can access data are managed
2.2.1		All changed to a 5-point scale and some modifications were made (refer to 〈Table 4〉).

Table 4. Detailed indicators on '2.2.1 Impact on information subject during re-identification' among classification systems for data use environment

Detailed indicators	Check
o When data is intentionally or unintentionally re-identified *, it is possible to create social confusion due to legal, moral and technical issues * When the identification information corresponds to sensitive information or includes information that directly invades the privacy of an individual, such as a phone number, email, or ID	5-point scale
o When data is intentionally or unintentionally re-identified, it may infringe the personal information or privacy of the relevant data subject.	5-point scale
o When data is intentionally or unintentionally re-identified, it may cause economic or uneconomical losses to relevant data subjects.	5-point scale
o When the data is intentionally or unintentionally re-identified, it may cause economic or uneconomic losses to the applicant institution.	5-point scale

Table 5. The Classification for data(itself)

3	Data(itself)		Basis
	3.1	Data organisation: The organisation of the original data itself	Revised and expanded the United States data context-based risk measurement method [6, 7]: reflects only k-anonymity -> expands to data composition, distribution and sensitivity
	3.1.1	Whether a unique identifier is included in the data	
	3.1.2	Total columns of data	
	3.1.3	Single or multiple statistical characteristics of dataset	
	3.1.4	Total number of quasi-identifier columns	
	3.1.5	Total number of sensitive information columns	
	3.1.6	Whether the original population is for all citizens	
	3.2	Data distribution: distribution of attribute values in each column of original data and whether outliers are included	
	3.2.1	Distribution of attribute values in each column (attribute)	
	3.2.2	Whether personally identifiable outliers are included	
	3.3	Data sensitivity: the sensitivity of the original data itself	
	3.3.1	Single, multiple, concatenated (behavioral or locational) temporal characteristics of the original data	
	3.3.2	Whether or not data restricted by the Personal Information Protection Act of Korea is included	
	3.3.3	Whether sensitive information covered by the industry group is included	

Table 6. Differences between the guidelines for de-identification of personal information in Korea[9] and the proposed method in the classification of data(itself)

<table>
<tr><th colspan="2"></th><th colspan="2">Guideline in Korea[9]</th><th colspan="3">The proposed scheme</th></tr>
<tr><td colspan="2">Measure-ment time</td><td colspan="2">When evaluating adequacy after de-identification</td><td colspan="3">Before de-identification</td></tr>
<tr><td rowspan="10">Metric</td><td colspan="2">-</td><td colspan="3">Whether a unique identifier is included in the data</td></tr>
<tr><td>Data size</td><td rowspan="9">Details are not provided</td><td rowspan="6">Data organisation</td><td>3.1.1</td><td rowspan="9">Refer to <Table 5></td></tr>
<tr><td rowspan="5">Specification by data column (range, number, etc.)</td><td>3.1.2</td></tr>
<tr><td>3.1.3</td></tr>
<tr><td>3.1.4</td></tr>
<tr><td>3.1.5</td></tr>
<tr><td>3.1.6</td></tr>
<tr><td rowspan="2">Distribution by data column</td><td rowspan="2">Data distribution</td><td>3.2.1</td></tr>
<tr><td>3.2.2</td></tr>
<tr><td colspan="2">-</td><td rowspan="2">Data sensitivity</td><td>3.3.1</td></tr>
<tr><td colspan="2"></td><td>3.3.2</td></tr>
<tr><td>How to measure</td><td colspan="2">Expert qualitative judgment</td><td colspan="3">Calculated by quantitative checklist type</td></tr>
</table>

Table 7. Measurement of data organisation parts in data(itself) classification(some examples)

	Detailed indicators	Check
Whether a unique identifier is included	o Does the column contain unique identifiers?	Appropriate / Inappropriate
Number of data columns	o The possibility of linking with other data is high due to the large number of columns (measurement of simple quantity).	Yes/No
Statistical characteristics of dataset	1. Dataset represents a single statistical attribute - Example: In the case of card company data, there are only attributes for card use.	□ 1
	2. Dataset represents several statistical properties - Example: In the case of card company data, it includes various properties such as card use, point use, and card loan.	□ 5
Number of quasi-identifier (QI) columns	o The number of quasi-identifiers among the columns is high, so there is a high possibility of personal identification through linkage with other data.	Yes/No
Number of sensitive attribute (SA) columns	o The number of sensitive information in the column is high, so it is highly possible to identify an individual through linkage or inference with other data, and there is a high possibility of invading privacy to the identified object through identification.	Yes/No

Table 8. Risk measurement scheme according to data situation classification

Measurement field	Measurement item			Reflection ratio (%)	
1. How to use data	Simple use, internal linkage, external linkage			40	40
2. Data usage environment	Possibility to attempt re-identification	Re-identification intention and ability	In the case of pseudonymization	(7.5)*	30
			In the case of anonymization	15	
		Personal information protection level (Only for pseudonymization)		(7.5)*	
	Impact of information subject on re-identification			15	
3. Data(itself)	Data organization			15	30
	Data distribution			6	
	Data sensitivity			9	
Total reflection ratio				100	100
* Only for pseudonymization					

Table 9. Processing level according to the final evaluation result

	Final risk level		Frequency rate (%)	Final score
Pseudonymization	Level 1	Normal	29.2	〈 52
	Level 2	High	44.2	〉= 52 & 〈 70
	Level 3	Very High	26.6	〉= 70
Anonymization	Level 1	Very Low	9.8	〈 42
	Level 2	Low	21.8	〉= 42 & 〈 53
	Level 3	Normal	36.8	〉= 53 & 〈 68
	Level 4	High	21.8	〉= 68 & 〈 79
	Level 5	Very High	9.8	〉= 79

3.3 측정 방법에 대한 예시

우리가 제안한 측정 방법에 대한 예시로 가명처리 단순 사용시 데이터 활용방법을 예시로 들고자 한다.

가명처리 단순사용의 경우 데이터 활용 방법에 대한 흐름은 Fig.4와 같다. Fig.4에서 각 단계별로 매우 낮음(Very low)(1점), 낮음(Low)(2점), 보통(Normal)(3점), 높음(High)(4점), 매우 낮음(Very high)(5점)을 부여한다. 이어 Table 10의 위험도 산출표에 따라 만일 합산 점수가 10(22%, 4Case)점 이하일 경우 보통(Normal), 11점~13점(45%, 8Case)일 경우 높음(High), 14점 이상(33%, 6Case)일 경우 매우 높음(Very high)으로 판정한다.

예로 회사 내부에서 사용하며 주기적인 분석을 필요로 하고 단일 목적으로만 사용하는 경우, Fig.4의 흐름도에 따라 기관내부(2점), 내부분석실(1점), 주기적(4점), 단일목적 사용(3점)이 부여되며 합계는 총 10점이다. 따라서 Table 9의 위험도 산출표에 따라 위험도는 보통(Normal)에 해당하고 위험도 최종 점수는 24점으로 산정된다. 따라서 이 경우 Table 8에 따라 데이터 활용방법에 따른 위험도 점수가 총 40점 중 24점이 부여되며 나머지 데이터 이용환경과 데이터 자체에 대한 위험도 60%를 반영하여 최종 위험도 점수가 산출된다.

3.4 비교분석

우리는 Table 11을 통해 지금까지 제안한 방법과 기존에 제안된 방법들을 비교 분석하고자 한다.

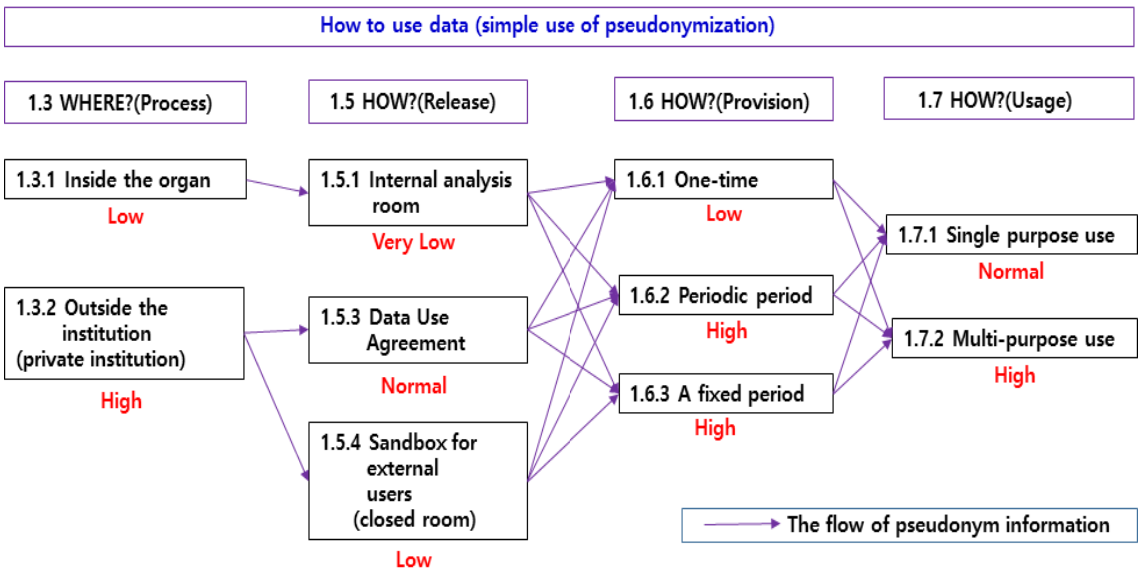

Fig. 4. Flow Chart(How to use data(simple use of pseudonymization))

Table 10. Risk calculation table for simple use of pseudonymization

Total score	Frequency		Frequency sum	Risk	Risk score
	Count	ratio			
8	1	5.56%	22.2%	Normal	24
9	1	5.56%			
10	2	11.11%			
11	3	16.67%	44.5%	high	32
12	2	11.11%			
13	3	16.67%			
14	4	22.22%	33.3%	Very high	40
15	2	11.11%			

Table 11. Comparison of existing risk measurement methods with the proposed method

	Duncan et al[5]	Emam' scheme[6,7,14,15]	Guideline in Korea[9]	The proposed scheme
Measurement before de-identification	O	O	X(조치 후 측정)	O
How to measure	Qualitative	Qualitative	Quantitative and qualitative	Quantitative
How to use data	Who, what, where, how, divided into 4	X	X	Why, what, where and how, divided into 7
Data usage environment	X	Importance of data, potential injuries / damages to patients, validity of approval (mainly dependent on precedent)	Analysis of the possibility of attempting re-identification and its effect on the data subject when re-identification occurs (quantitative)	Improving 'Guidelines in Korea'[9] (Refer to 〈Table 3〉)
Data(itself)	X	k-anonymity only	Data specification and de-identification status analysis (qualitative)	Improving 'Guidelines in Korea'[9] (Refer to 〈Table 6〉)

기존 가이드라인[9]에서는 데이터 이용환경에 대한 위험도를 비식별 조치 이후 적정성 평가시 수행하는데 반해 제안하는 방법은 이와 달리 조치 이전에 측정한다. 그 이유는 비식별 조치를 수행하는 실무자 입장에서 사전에 이러한 환경을 고려하여 측정함으로서 충분한 사전 조치를 수행할 수 있고 조치된 이후에 또 한번 측정함으로써 조치가 제대로 이루어졌는지를 자체적으로 평가할 수 있기 때문이다.

첫 번째 측정 요소로서 데이터 활용 방법에 있어 기존 Duncan 등[5]이 누가, 무엇을, 어디서, 어떻게로 단순 분류하였다면 우리는 이를 보다 구체화하여 왜, 무엇을, 어디서, 어떻게로 분류하였다. 4가지 구성은 동일하지만 범주가 다르고 세부 항목도 7가지로 제시하였다. 여기서 누가에 대한 부분이 빠진 이유는 이 부분을 제안 방법에서는 데이터 이용 환경 부분에서 보다 상세하고 다루고 있기 때문이다. 아울러 나머지 3가지 항목은 비식별 조치의 실제 환경에 맞게 보다 구체화하여 7가지로 확장하여 제안하였다.

두 번째 측정요소인 데이터 이용 환경에 있어 기존 Emam의 방법[6, 7]은 앞서 2.4절에서 제시한 바와 같이 의료 환경에 치중되어 있어 범용환경에 적합하지 않다. 한편 이 보다는 우리나라 가이드라인[9]이 보다 구체적으로 제시되어 있는데 이를 제안하는 방법에서는 국내 개정 개인정보보호법을 반영하여 가명처리와 익명처리로 구분하였으며 기존 지표 중 현실에 맞지 않은 부분은 보다 현실성 있게 개선하였다(보다 자세한 사항은 Table 4의 밑줄친 부분을 참조하기 바란다.) 아울러 기존 예/아니오로 평가되던 것을 5점 척도로 변환한 것도 차이가 있다. 예로 '데이터가 의도적 또는 비의도적으로 재식별 되었을 때 신청기관에게 경제적 또는 비경제적 손실을 발생시킬 수 있음'에 대한 항목의 경우 기존 예/아니오로 평가되었지만 현실에서는 평가자가 이를 단순히 예/아니오로만 평가하기 힘든 점이 그러하다.

세 번째 측정요소인 데이터(자체)에 있어 기존 Emam[6,7,14,15]의 방법은 단순히 k-익명성의 k 값만을 이용하여 위험도를 측정하고 있으며 우리나라 가이드라인에서는 현황분석에 있어 정량적이지 않으며 구체적이지 않는(Table 6 참조) 반면에 제안하는 방법에 있어서는 Table 5와 Table 6에서 보는 바와 같이 데이터 구성, 분포, 그리고 민감도로 보다 다양화하여 현실에 맞게 개선하였다.

결론적으로 볼 때 제안하는 방법은 기존 Duncan 등[5]이 제안하는 환경제어 부분과 우리나라 가이드

라인[9]과 Emam[6,7,14,15]의 방법에서 제안한 데이터 이용 환경 및 데이터(자체)에 대한 부분을 종합적으로 고려하여 보완하고 개선하였다.

3.5 개정 데이터 3법에 대한 분석

지난 1월 9일 국회 본회의를 통과한 개인정보보호법 등 이른바 데이터 3법이 8월 5일부로 시행될 예정이다. 개인정보 비식별 조치와 관련하여 가장 눈여겨 볼 점은 개인정보에 가명정보의 개념이 추가되었고 개인정보보호법 28조의2 1항의 통계작성, 과학적 연구, 공익적 기록 보존 목적에 따라 개인 동의없이도 가명처리된 가명정보를 통해 활용이 가능해졌다는 점이다. 아울러 이 법의 시행으로 인해 기존 비식별 조치 가이드라인은 자동으로 폐기된다. 기존의 가이드라인은 기본적으로 익명처리에 초점이 맞추어져 있어 이번에 시행되는 가명처리 규정에는 적합하지 않기 때문이다. 본 논문에서는 이러한 법 개정에 맞추어 데이터 상황에 따른 분류체계를 아래와 같이 반영하였다. 첫째, 1. 데이터 활용 방법에 있어 활용 목적(1.1 왜(Why?))을 가명처리와 익명처리로 각각 구분하여 Table 1의 1.1.1항과 1.1.2항에 분류하였다. 둘째, 2. 데이터 이용환경에 있어서도 가명처리의 경우와 익명처리의 경우를 구분하여 분류하였다. 다시 말해 2.1.2 데이터 이용자 또는 제공받는 기관 또는 기업의 개인정보보호 능력에 대한 항목의 경우 익명처리시 제외하도록 구성하였다(Table 3 참조). 왜냐하면 익명처리의 경우 개정 개인정보보호법 제58조의 2에 따라 더 이상 보호의 대상이 아니기 때문이다.

3.6 활용측면에 있어 제안하는 측정 방법에 대한 기대 효과

지난 2016년 6월 행정안전부를 비롯한 6개 정부부처가 합동으로 데이터를 안전하게 활용할 수 있도록 비식별 조치 가이드라인을 제정한 바 있다. 그러나 기존 가이드라인은 다른 이유들도 많지만 특히 비식별 조치한 정보들에 대한 적정성 평가시 k-익명성을 강제 조치하도록 함으로 인해 개인정보를 활용하려는 기업이나 기관들로부터 외면을 받아왔다. 주된 이유는 데이터의 왜곡이 심하여 분석을 위한 품질이 저하된다는 것이다. 예컨대 일부 금융이나 통신 분야를 제외한 의료분야의 담당 전문기관인 사회보장정보원이나 교육 분야의 전문기관인 한국교육학술정보원의 경우 지난 2016년 이후 활용 사례를 단 한 건도 가지고 있지 않다. 또한 가이드라인의 안전성 측면에서도 지난 2017년 시민단체들로부터 검찰에 고발을 당하는 등 크고 작은 문제점들이 노출된 바 있다. 또한 우리는 여러 현장 실무자들로부터 비식별 조치에 대한 기본 개념이 부족하여 방법론부터 애로를 겪는 사례들을 자주 접한 바 있다. 특히 스타트업이나 소기업들은 더욱 그러하다.

그럼에도 불구하고 지난 2월 5일 개인정보보호법 등 이른바 데이터 3법의 개정을 계기로 개인정보를 가명조치함으로서 과학적연구나 통계작성 등에 활용할 수 있는 길이 열렸다.

본 논문에서 제안하는 방법은 이러한 애로를 겪는 실무자들을 위해 현장에서 비식별 조치 이전에 데이터 상황에 대한 전반적인 위험도를 측정, 진단하고 조치 수준을 결정할 수 있도록 방법론을 제안하고 돕고자 하는데 그 목적이 있다. 아울러 개정된 이른바 데이터 3법에 따라 활용 목적에 맞게 가명처리와 익명처리를 구분하여 측정하도록 제시하였다.

제안한 방법은 향후 기업이나 기관 등 조직 내 개인정보 비식별 조치를 다루는 실무자나 혹은 전문가들이 개정된 데이터 3법의 테두리 내에서 실제 비식별 조치 수행 시 어떠한 시각과 방법을 통해 안전하고 적절하게 조치를 수행해야 하는지에 대한 기초 자료로서 활용 가능성이 매우 높을 것으로 판단된다.

IV. 결론 및 향후 연구방향

우리는 지금까지 개인정보 비식별 조치에 있어 데이터 상황을 고려한 위험도 기반의 측정에 대한 새로운 방법을 제안하였다. 제안한 방법은 데이터에 대한 상황을 크게 데이터 활용방법, 데이터 이용환경, 그리고 데이터(자체) 3가지 카테고리로 나누어 보다 체계적으로 분류하였으며 기존 Duncan 등[5], Emam의 방법[6,7,14,15], 그리고 우리나라 개인정보 비식별조치 가이드라인[9]에 비해 위험도의 산정을 외부 전문가에게만 맡기는 것이 아니라 일반 조직 내 개인정보처리자가 정량적으로 위험도를 산정할 수 있도록 일종의 가이드 형태로 구현하였다. 한편 Table 9에 따라 결정된 비식별 처리 수준에 따른 세부적인 처리 방법, 절차, 그리고 처리 결과에 대한 적정성 평가 등에 대한 부분은 현재 후속 연구로 진행 중이며 조만간 결과를 발표할 예정이다.

References

[1] Elliot, M. J., Dibben, C., Gowans, H., Mackey, E., Lightfoot, D., O'Hara, K., and Purdam, K, "Functional Anonymisation: The crucial role of the data environment in determining the classification of data as (non-) personal," CMIST work paper 2015.

[2] Sweeney L, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(3), pp. 557-570, 2002.

[3] UKAN(UK Anonymisation Network), "The anonymisation decision making framework," 2016.

[4] Mackey, E and Elliot, M. J, "Understanding the Data Environment," XRDS: Crossroads, 20(1), pp. 37-39, 2016.

[5] Duncan, G. T, Elliot, M. J and Salazar-Gonzalez, J. J, "Statistical Confidentiality," New York: Springer, 2011.

[6] Khaled El Eman, "Guide to the De-identification of Personal Health Information," CRC Press, pp. 203-221, 2013.

[7] HITRUST and Privacy Analytics, HITRUST Data De-identification Methodology, Training course material, 2019.

[8] NIST 800-188(2nd Draft) De-Identifying Government Datasets, Dec. 2016.

[9] Joint government departments in Korea, Guidelines for de-identification of personal information, June. 2016.

[10] Nissenbaum HF, "Privacy in Context: technology, policy, and the integrity of social life, Stanford, California," Stanford Law Books, 2010.

[11] Bieker F, Friedewald M, Hansen M, Obersteller H, and Rost M, "A process for data protection impact assessment under the european general data protection regulation," Lecture notes in computer science, Proceeedings of 4th annual privacy forum, pp. 21-37, 2016.

[12] Mulligan DK, Koopman C, and Doty N, "Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy," Philos Trans Ser A Math Phys Eng Sci, 374(2083), pp. 1-17, 2016.

[13] Solove DJ, "A taxonomy of privacy," Univ Pa Law Rev, 154(3), pp. 477-564, 2006.

[14] Khaled El Emam, "Risk-based de-identification of health data," IEEE Security & Privacy,8(3), pp. 64-67, 2010.

[15] Khaled El Emam and Luk Arbuckle, "Anonymizing health data," O'Reilly book, pp. 29-33, 2013.

[16] Fabian Prasser, Florian Kohlmayer, and Klaus A. Kuhn, "The Importance of Context: Risk-Based De-Identification of Biomedical Data," Methods of Information in Medicine, Schattauer, June. 2016.

[17] Oleksandr Tomashchuk, Dimitri Van Landuyt, Daniel PleteaKim Wuyts, and Wouter Joosen, "A data utility-driven benchmark for de-identification methods," International Conference on Trust and Privacy in Digital Business, TrustBus 2019, Lecture Notes in Computer Science book series, volume 11711, pp 63-77, 2019.

[18] HIPAA(Health Insurance Portability and Accountability Act) Privacy Rule, Dec. 2012.

〈저자소개〉

김 동 현 (Dong-hyun Kim) 정회원
2013년 2월: 동국대학교 정보보호학과 석사
2020년 8월: 중앙대학교 융합보안학과 박사과정
2010년 10월~현재: 한국인터넷진흥원 데이터활용지원팀 책임연구원
〈관심분야〉 개인정보보호, 비식별, 데이터 위험관리

김 순 석 (Soon-seok Kim) 종신회원
2003년 2월: 중앙대학교 컴퓨터공학과 박사
2003년 3월~현재: 원주 한라대학교 컴퓨터공학과 부교수
〈관심분야〉 개인정보보호, 비식별