Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.30, NO.2, Apr. 2020 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2020.30.2.253

가명 데이터 활용연구 - 기술적 처리방법 및 기업의 활용방향을 중심으로 -*

김 정 선†‡
SK텔레콤 AIX센터 (부장)

Research on the Use of Pseudonym Data - Focusing on Technical Processing Methods and Corporate Utilization Directions -*

Jung-Sun Kim†‡
SK Telecom AIX Center (General Manager)

요 약

본 연구는 본격적인 데이터 경제 활성화를 위한 데이터 3법 통과 이후 기업의 가명데이터의 활용과 관련한 기술과 적용 프로세스에 대하여 살펴보고, 기업이 가명 데이터를 활용하기 위해 준비해야 할 사항과 그 과정에서 발생하게 되는 불확실성을 제거하는데 기여하고자 하였다. 앞으로 기업에서는 기존 IT 시스템 관점의 정보보호 관리체계를 확장하여 데이터 프라이버시 보호 관점에서의 관리와 통제 및 제3자 제공 관점에서의 관리가 필요하다. 또한 내부 이용자가 활용하는 데이터 이용환경에서조차 적절한 가명데이터 이용통제가 이루어져야 한다. 가명화 데이터 활용에 따른 시장 변화 및 이종 데이터 결합 활용을 통한 경제적 창출 효과는 매우 클 것이며, 이를 위해 조속한 시간 내에 데이터 활용 및 거래 활성화를 위한 적절한 비식별 조치 기준 및 위험도 판단 기준 수립이 준비되어야 한다.

ABSTRACT

This study examines the technologies and application processes related to the use of pseudonym data of companies after the passage of the Data 3 Act to activate the data economy in earnest, and what companies should prepare to use pseudonym data and what will happen in the process It was intended to contribute to the elimination of uncertainty. In the future, companies will need to extend the information security management system from the perspective of the existing IT system to manage and control data privacy protection and management from a third party provisioning perspective. In addition, proper pseudonym data use control should be implemented even in the data use environment utilized by internal users. The economic effect of market change and heterogeneous data combination due to the use of pseudonymized data will be very large, and standards for appropriate non-identification measures and risk assessment criteria for data utilization and transaction activation should be prepared in a short time.

Keywords: Data Economy, De-identification, Pseudonym

Received(02. 20. 2020), Modified(1st: 03. 27. 2020, 2nd: 04. 06. 2020), Accepted(04. 06. 2020)

† 주저자, jungsun.kim@sk.com
‡ 교신저자, jungsun.kim@sk.com(Corresponding author)

I. 서 론

GDPR(General Data Protection Regulation) 시행 이후 국내에서도 개인정보보호법 내 가명정보의 개념을 도입하는 것으로 2018년 상반기 4차 산업혁명 위원회가 주최한 두 차례의 해커톤을 통해 합의과정을 거친 바 있다. GDPR은 2018년 5월 25일부터 시행된 EU의 개인정보보호 법령으로, EU 국가 내 사업장을 운영하는 기업뿐 아니라 전자상거래 등을 통해 해외에서 EU 주민의 개인정보를 처리하는 기업에도 적용되며, 법령 위반 시 전 세계 연간 매출액의 4% 또는 2000만 유로 중 더 높은 금액이 과징금으로 부과된다.

해커톤 시행 이후 법안 마련의 절차를 거쳐 현재 데이터 경제3법으로 불리우는 개인정보보호법, 신용정보보호법, 정보통신망법이 8월 5일 발효를 앞두고 있다. 개정 법안에 따르면, 통계 작성, 과학적 연구 및 공익적 기록 보존 등을 위해 정보 주체의 동의 없이 가명정보 사용이 가능해진다. 또한 제3자에게 가명 정보를 제공할 수도 있다. 개정된 데이터 3법에서는 개인정보 관련 개념 체계를 개인정보, 가명정보, 익명정보로 명확히 구분하였으며, 익명정보는 더 이상 개인정보보호법의 적용 대상이 아님을 분명히 하고 있다. 그러므로 현 시점에서는 가명정보의 처리 방법과 활용 방법과 관련하여 명확한 가이드라인에 대한 준비가 필요하다.

기존 법안의 적용 내에서는 개인정보와 개인 식별정보에 대한 범위와 구분에 대한 모호성과 가이드라인 등이 제시하는 k익명성 등에 대한 적용과 분석에 필요한 준식별자 및 민감정보 등에 대한 비식별화 처리기법 상의 혼란 등이 존재하고 있다. 따라서 데이터의 가치를 훼손시키는 방식으로 비식별화가 이루어지고 결과적으로 데이터 경제 활성화를 통한 지능정보사회의 가치 구현이 저해되는 모습이 나타나고 있다. 이에 새롭게 적용되는 시행령과 가이드라인은 데이터 처리, 공유 및 활용 시 기업에서 발생 가능한 불확실성을 제거해 주어야 한다.

본 연구에서는 기업에서 가명 데이터를 활용하기 위해 기술적으로 준비해야 할 사항과 그 과정에서 발생가능한 불확실성을 제거하기 위한 준비사항 등을 제안함으로써 향후 본격적으로 도래할 데이터 경제시장 창출 및 활성화에 기여하고자 한다.

II. 가명처리 기술 및 활용 현황

2.1 가명화의 개념 및 처리 기술의 현황

비식별화란 본질적으로 개인정보를 구성하는 세 가지 요인인 특정 데이터가 한 개인과 대응됨(Single-out), 특정 데이터와 특정 개인이 연결됨(Linkability), 주어진 데이터 세트 내에서 특정 개인을 추론할 수 있음(Inference)을 예방조치 하는 것이다. 연결성과 추론을 방지하기 위해 직접 식별자에 조치를 취하는 가명화, 준식별자를 포함한 속성정보도 비식별화하여 세 가지 조건을 모두 반영한 것이 익명화이다.

비식별화는 가명화, 익명화보다 상위 개념으로 전 세계적으로 보편화된 개념이지만 미국 내 의료정보에 대하여 비식별화 지원 및 안전성 진단을 수행하는 HITRUST Alliance에 따르면 익명화의 개념을 가명화와 비식별화 개념의 상위로 보기도 하고 있어 전 세계적인 용어 통일은 이루어지지 않고 있는 모습이다[1]. 또한 기업이 비식별화를 진행하는 방법으로 데이터 세트에 직접적인 조치를 취하는 방법(삭제, 범주화, 치환 등)과 자료를 제공, 이용하는 방식을 통제하는 관리적 방법이 활용될 수 있다.

가명화(Pseudonym)는 개인과 관련된 데이터에 대하여 추가 정보 없이는 특정 개인에 대한 데이터에 대하여 다시 연결할 수 없도록 처리하는 것을 의미한다. 예를 들어, 이름, 핸드폰 번호, 이메일 주소, 연령, 국적 및 직장 이름이 제출되어 있는 경우, 가명화를 통해 특정인을 식별할 수 있는 식별자(이름, 핸드폰 번호 등)를 제거하고 국적과 같이 개인을 특정할 수 있는 데이터를 비식별화한 후 기존 식별계 정보와 분리시켜 관리해야 한다. 합목적적으로 추가 정보를 활용하여 특정인에 대한 정보와 다시 연결이 가능한 가역성(reverse engineering)을 가진다는 것이 익명 데이터와는 다른 점이다. 이에 활용되는 가명화 처리 기술에는 기존 정보를 다른 정보로 대치하거나 암호화 수행 또는 일부 정보를 가리는 마스킹 개념 등이 활용된다.

GDPR에 따르면, 가명처리란 추가적인 정보 없이는 개인을 식별할 수 없는 정보로 정의하면서, 가명정보도 개인정보의 일부분으로 보고 있다[2]. 통과된 데이터 3법 중 개인정보보호법 제2의 1-2항에서는 '가명처리란 개인정보의 일부를 삭제하거나 일부 또는 전부를 대체하는 방법으로 추가 정보 없이는

특정 개인을 알아볼 수 없도록 처리하는 것'으로 정의하고 있다[3]. 사실 법이란 매우 개념적이고 규범적인 성격을 갖는데, 가명처리란 그러한 개념적이고 규범적인 법테두리 안에 다분히 기술적인 조치와 개념이 함께 반영되었다는데 그 차별성이 있다. 그렇기 때문에 수범기관에서는 데이터 가명화에 적용되는 기술에 대한 정확한 이해와 준비의 필요성이 생긴다. GDPR에서 개인 식별자를 정의하고 있는 핵심 기준은 '개인을 특정할 수 있다'는 것과 '개인을 식별할 수 있다'는 부분을 분리하는 것이다. 즉, 데이터 세트 내에서 단일 값을 가진 레코드가 존재하더라도 식별자에 대한 비식별화 조치가 적절하게 이루어졌다면 가명화된 데이터로 인정한다는 것이다.

그림 1은 개인의 체중정보와 특별지원수당 지급금과의 상관관계 분석을 위하여 생성한 데이터세트이다. 원본 데이터에는 개인의 이름, 주소, 생년월일 등의 정보가 들어 있었으나 해당 정보를 삭제하고 해시함수를 이용하여 연구 코호트 번호를 생성한 사례이다. 해당 데이터 세트 내에서는 단위 데이터의 특정성은 살아 있으나 식별성은 제거된 상황이다.

GDPR에서 규정하는 개인식별자(PII)는 매우 포괄적이다. GDPR article 4(5)항에 따르면, 개인정보는 "식별되거나 식별 가능한 자연인(데이터 대상)에 관련된 정보이며, 식별 가능한 자연인은 이름, 식별 번호, 위치 데이터, 온라인 식별자나 해당 자연인의 신체적, 생리학적, 유전적, 정신적, 경제적, 문화적 또는 사회적 신원에 대한 한 개 이상의 인자를 참조하여 직간접적으로 식별할 수 있는 사람을 의미한다"고 규정되어 있다[4]. 여기에는 IP 주소와 쿠키 데이터가 포함될 수 있으며, GDPR이 정보접근 요청(Subject Access Request), 잊힐 권리/삭제 권리, 데이터 확률 등의 새로운 개념을 도입하면서 이제 개인정보 주체는 자신에 대해 각 기업에서 어떤 데이터가 수집되는지 알 권리가 있고, 실제 이메일과 소셜 플랫폼부터 HR, CRM 시스템까지 모든 곳에 개인 식별자가 존재할 수 있으므로 해당 데이터에 대해 민관의 적극적인 관리 활동이 요구되어 진다.

name, adress, birthday	Special allowance	Body mass index	Research Cohort Reference Number
[redacted]	<2y	15	QA5FRD4
	>5y	14	2B48HFG
	<2y	16	RC3URPQ
	>5y	18	SD289K9
	<2y	20	5E1FL7Q

Fig. 1. An example of pseudonym data

GDPR 내에서 규정하는 데이터의 일반적 처리에 대한 광범위한 규칙 목록의 주요 내용은 아래의 5가지 영역으로 요약할 수 있다.

- 합법적 관심사 : 데이터를 수집하고 저장하는 유효한 이유가 있어야 함.
- 동의 : 상업적 또는 마케팅 목적으로 개인에 대한 특정 유형의 개인 데이터를 사용하려면 먼저 개인의 허가를 득해야 함.
- 잊혀질 권리 : 개인은 자신의 개인 데이터에 대한 액세스 권한을 취소할 수 있으며 기업은 이때 데이터를 삭제, 관리하는 루틴을 가지고 있어야 함.
- 익명화/가명화 : 개인 데이터는 익명으로 저장하거나 가명화를 통해 저장해야 함. 이로써 개인정보성 데이터에 대한 부적절한 접근이 불가능함.
- 개인정보의 이동 : 개인 데이터를 외부 당사자와 공유하는 데 동의하면 당사자가 해당 데이터에 대한 액세스 권한을 취소할 수 있는 프로세스도 동시에 관리되어져야 함. 여기에는 파트너의 데이터 웨어하우스에서 데이터가 지워지도록 하는 의무가 포함됨.

통과된 개인정보보호법 개정안에는 위의 주요 방향 중 동의 및 가명화/익명화 부분만이 반영되었다. 아직은 정보이동권이나 잊혀질 권리에 대한 부분은 반영되지 않고 있지만, 국제법과의 정합성을 추구하는 장기적 방향성 하에서 국내에서도 얼마 지나지 않아 정보 주체의 권리를 강화하기 위한 동의제도 개선과 정보이동권에 대한 이슈가 구체화될 것으로 보인다. 신용정보보호법 내 정보이동권에 대한 항목과 개인정보에 대한 동의 관련 개선 부분이 이를 보여주고 있다.

2.2 기업 내 가명 데이터 활용을 위한 준비사항

현재 기업들의 데이터 활용 프로세스에서 개인정보보호 관리를 위한 정보보호 관리체계(ISMS-P) 인증제도가 시행되고 있다[5]. 일반적으로 기업에 적용되는 정보호보 관리체계로는 가명데이터 활용에 대한 데이터의 안정성, 신뢰성 모두를 담보하기 어려운데, ISO27001과 NIST 표준 및 지침 또한 비식별화 및 가명화 처리와 데이터 프라이버시 영역과 관련해서는 정비를 시작한지 얼마되지 않았다[6]. 또

한 해당 보호관리체계는 공공기관 및 기업 내부에서 데이터 이용을 관리하기 위한 프레임웍을 적용하고 있는데 반해 데이터 3법 개정과 동시에 고려되어야 할 프레임웍은 외부 제공 및 결합을 위한 관리구조이다. 그러므로 비식별화 및 가명데이터를 관리하기 위한 기업의 데이터 시스템 구조와 프로세스는 향후 정비되어야 할 점이 많을 것으로 보인다. 이를 위해 기업 가명데이터 관리를 위한 필수적인 준비사항은 다음과 같다.

개정된 개인정보보호법 제28조 4의 ①~②에는 가명정보에 대한 안전조치 의무가 적시되어 있다. 이에 따르면 추가정보를 별도로 분리하여 보관·관리하여야 하며 안전성 확보에 필요한 기술적, 관리적 및 물리적 조치를 취하여야 한다. 이에 따라 기업들은 데이터 레이크(data lake)에 활용이 가능한 데이터가 수집되면 보안이 필요한 데이터와 그렇지 않은 나머지 데이터를 분리하여 관리하여야 하며, 내부 이용자가 활용하는 환경에서조차 적절한 접근 권한에 따른 가명화 처리를 취하여야 할 것이다. 즉, 가명정보 개념 도입에 따라 기업의 데이터 관리 체계의 이원화를 통해 식별계와 가명계로 분리된 데이터 이용환경의 설계가 필요하다.

다음으로는 가명화 처리과정에서 데이터를 유출당하거나 비정상적 공격에 의한 해킹의 경우에도 민감한 개인 데이터에 접근할 수 없고, 민감한 개인 데이터를 한 사람이나 원소스로 쉽게 추적할 수 없도록 조치해야 한다. 그러면서도 이후 특정 권한을 보유한 처리자는 나중에 중요한 데이터를 읽을 수 있는 상태로 렌더링하는 방법으로 데이터를 안전하게 저장 가능해야 할 것이다. 가명화의 이러한 가역적 데이터 관리 프로세스에 있어서는 무단 접근으로부터는 데이터를 안전하게 보호하면서도 특정 수준의 접근 가능성을 허용하는 부분이 관리의 핵심이 된다.

또한 가명데이터의 개념이 개인정보보호법 개정안에 반영되면서 가장 크게 달라지는 부분은 동의를 받지 않아도 개인정보를 가명화하여 제3자에게 제공하거나 결합, 활용할 수 있는 부분이다. 이를 기반으로 자체 보유 데이터 이외의 외부 데이터를 결합 활용하여 보다 나은 가입자 분석을 통한 개인 맞춤형 서비스를 제공하거나 인공지능 알고리듬 개발이 가능하게 될 것이다.

ENISA(2018)[7]에 따르면, 가명화는 단일 단순 속성/식별자에 적용하는 기술은 아니며, 따라서 정보처리자가 누구이며 어떤 조건인지 그리고 데이터를 이용하게 되는 목적과 제공자를 고려한 상황(Context)에 따라 가명화에 대한 처리 수준을 고려할 것을 권고하고 있다. 즉, 내부 이용 시에 정해진 정보보호 준칙에 따라 비식별처리를 진행하되, 외부에 제공하게 될 경우에는 데이터 자체의 특성과 상황별 개인정보 처리자를 고려하여 비식별화 수준을 결정하여야 한다. 이때 참고 가능한 정보로는 안전조치 의무의 준수 여부(법 28조의4), 개인정보보호 인증 취득 여부(법 32조의2), 개인정보 영향평가(공공기관, 법 33조) 시행 여부, 결합 전문기관의 보안 안전성 수준 등이 고려될 수 있다.

새로 도입되는 데이터 3법의 개정 요건에 대한 해석과 향후 대통령령의 구체적인 내용에 따라 개정법의 운영이 달라지게 될 가능성도 존재한다. 우선 개정된 개인정보보호법 제27조 7항에는 가명데이터의 개인정보 규제 적용 배제 조항이 명기되어 있다. 이에 근거하여 기존까지 파기관리 할 수 밖에 없었던 데이터들에 대한 보관 연한에 대해서도 새로운 해석이 가능해 질 수 있다. 특히 개정된 신용정보보호법에는 해당 정보주체와의 상거래관계가 종료된 날로부터 최장 5년 이내(그 이전에 목적이 달성된 경우, 목적이 달성된 날로부터 3개월 이내)에 개인신용정보를 삭제할 의무를 부과하고 있으나, 가명 처리된 개인 신용정보의 경우에는 이에 대한 예외가 인정된다.

그리고 개정된 개인정보보호법 제15조의 양립가능성 개념에 대해서도 향후 해석이 분분할 가능성이 있다. 해당 조항에 따르면, '개인정보처리자'는 당초 수집 목적과 합리적으로 관련된 범위 내에서 정보주체에게 불이익이 발생하는 지 여부, 암호화 등 안전성에 필요한 조치를 하였는지 여부 등을 고려하여 양립가능한 목적에 따라 데이터를 이용할 수 있는데 양립가능한 범위 및 이용 목적에 대한 해석이 분분할 가능성이 있기 때문이다. 그러므로 관련 시행령에서는 데이터 활용에 대한 여지와 자유도를 허용하고 개인에 대한 막대한 불이익과 안전성 저해가 예상되는 경우에만 관련 활용을 배제하는 네거티브(negative) 규제를 적용하는 것이 바람직할 것이다[8].

데이터 3법 통과 시 기대되는 가명데이터 활용 측면의 시장 변화는 다음과 같다.

우선 가명계를 구축하기 위한 시스템 투자가 활발해질 것이며, 이를 위한 클라우드 활용 시 별도의 위·수탁 동의 절차가 불필요해진다. 이에 따라 자연스러운 클라우드 시장 활성화가 가능해질 것이다.

다음으로 개별 기업의 신규 서비스 출시 시에도 추가적인 개인정보 활용 동의 없이 기업 내 가명 데이터 연계 분석 및 활용이 가능질 것이다. 기존에는 동일 기업에서 제공되는 서비스별 개별 동의가 필요한 상황이었으므로 그에 따른 시간과 비용의 절감이 가능해질 것으로 기대된다.

또한 가명정보를 제3자에게 제공하는 것도 가능하므로 비식별화를 대행하거나 가명정보의 결합 Key를 생성해주는 TTP(trusted 3'rd party)의 출현도 가능해질 것이다. 결합대행기관의 경우 기존 공공기관으로 한정된 부분을 적격한 조건을 갖춘 민간기업으로의 확대까지도 가능할 것이다.

마지막으로 기업 간 데이터 결합 활성화에 따른 데이터 상품 출현과 데이터 거래소의 활성화가 기대된다. 특히 이를 통해 AI생태계에 필수적인 올바른 학습용 데이터 셋의 확보가 용이해 질 것이다. 대부분의 기업들은 가입자 서비스를 제공하면서 수집된 자기 고객의 정보 외에는 가지고 있지 못하다. 이러한 데이터는 고객 유지 활동에는 도움이 될 수 있으나, 신규고객 발굴을 위한 활동에는 아무런 인사이트를 제공할 수 없다. 그러므로 직접적인 마케팅 활용에 대한 선택적 동의는 별도로 받아야 할지라도 가명데이터 수준의 이종 데이터 결합세트는 기업 활동에 추가적인 또 다른 인사이트 제공이 가능하다.

2016년도 발의된 비식별조치 가이드라인의 이종 데이터 결합 프로세스는 사전 비식별화 및 적정성 평가, 결합키 생성 후 전문기관을 통한 결합 및 키 삭제, 반출대상 데이터 식별화 및 사후 적정성 평가로 이루어진다. 그러나 해당 프로세스에 대한 비용과 시간이 만만치 않아 시행 기업들로부터의 개선 요구가 다수 발생했고 이러한 점들은 이후 개정될 가이드라인에 반영될 필요가 있다. 특히 결합 전, 후 비식별화를 이중 수행함으로써 발생되는 데이터 품질 결함 및 삭제율의 문제는 시급한 개선이 필요하다.

데이터 3법이 개정된 배경인 데이터 생태계 활성화를 위해서는 향후 수범기관의 불확실성을 제거하는 노력이 필요한 데 이를 위해서는 명확한 가명처리의 기준과 처리방법, 결합 프로세스, 익명화에 대한 수준과 기준 및 위험을 판단할 수 있는 매트릭스 정의 등이 포함되어져야 할 것이다.

① 기관 내 가명데이터 연계, 결합

우선 가명화 처리된 데이터의 경우 그림 2와 같이 기관 내에서 추가적인 정보를 활용한 재식별화가 가능해지며 개별 서비스의 수집 데이터를 가명화한 후 결합 활용하게 된다. 이때 해당 기관에서는 해당 프로세스의 적정성, 이용목적 내 활용, 적절한 가명정보 관리에 대한 거버넌스(governance) 체계가 필요하다.

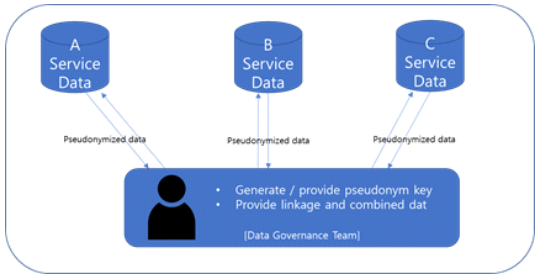

Fig. 2. An example of an enterprise's internal pseudonym data utilization

② 외부 데이터 연계, 결합

기존 비식별조치 가이드라인에서 규정된 이종 간 데이터 결합의 방법 외에도 안전하면서도 데이터의 활용성을 높일 수 있는 다양한 방법의 고려가 가능하다.

우선 전문기관의 성격을 정부가 지정한 공공기관으로 제한을 두지 않고 적격한 조건을 보유한 민간기업 등으로의 확대가 가능하다. 미국의 경우 데이터의 결합은 브로커리지 등의 민간기업에서 주로 진행하고 있으며, 일부의 경우를 제외하고는 데이터 결합의 법적 제약이 없다. 공공데이터의 경우 영국 행정데이터 작업반에서는 다양한 데이터 결합방법에 대한 장·단점을 다음의 표 2와 같이 분석하고 있으며 이중 제3의 신뢰기관을 통한 결합의 안정성이 가장 높다고 분석하고 있다[9]. ADRN(Administrative Data Research Network)의 경우 결합키를 발행하는 기관과 결합을 대행하는 기관을 이원화하여 운영하고 있다.

의료 데이터에 한정적이긴 하지만 결합과 관련한 장점을 13가지로 정리한 웰컴 트러스트의 보고서에 따르면, 결합데이터 이용 시 다음과 같은 혜택이 발생 가능하다고 한다[10].

- 실험군과 대조군의 정확한 분류 진행 가능
- 연구 가능 주제영역의 확대
  - 의료 분야 종단적 연구의 확대 가능
  - 회귀분석(retrospective analysis) 및 기대분

석 prospective analysis)에 유용
- 통계 데이터의 확대 가능
- 데이터의 일관성 검증 및 추가 속성 확보
- 드문 사건의 분석에 유용
- 다수준 모델링을 통한 환경적 요소 발견 가능
- 시뮬레이션 모델 개발 가능
- 조사방식에 비해 시의적절한 분석 수행 가능
- 데이터 수집 비용 감축
- 국가 간 데이터 연계 시 환경효과 비교 가능
- 학제 간 연구의 촉진 가능

그러나 동 보고서에서는 데이터 연계 시 발생 가능한 문제점에 대해서도 아래와 같이 지적하고 있다.

- 통계적 문제 발생 가능 : 데이터의 부족, 연계 필드가 없거나 데이터의 편향성, 부적절한 가정, 실험군의 대표성 상실, 대조군의 부족, 데이터 수집 시점의 불일치 등의 문제를 포함
- 기술적 문제 발생 가능 : 연계 데이터 이용 절차의 법적, 구조적 복잡성 및 데이터 연계 수행 방식, 연구자의 데이터 접근 방식 정의의 문제
- 제도적 문제 발생 가능 : 결합 데이터 이용에 관한 법적문제와 개인권리 침해와 사회적 이익의 균형여부를 확보하는 윤리적 문제, 시민사회와 기업, 학계간의 데이터 연계 공유에 대한 관점 차이로 비롯되는 문화적 문제를 포함

그러므로 개정될 가이드라인에서는 데이터 연계 결합절차의 표준화 및 통계적 위험도 접근 방식이 명확하게 제시되어야 한다. 또한 개인정보보호위원회의 독립과 더불어 이해관계자의 단순화, 신뢰받는 제3기관의 도입 및 연계 방식과 절차에 대한 구조화도 필요하다. 결합 데이터 활용에 대해서는 기업 및 행정기관의 정보공개의 원칙과 더불어 결합 활용 사례의 장점을 공유함으로써 시민사회와의 관점상의 간극을 좁혀나가는 일도 병행되어야 할 것이다.

Table 1. Analysis of advantages and disadvantages according to the combine model

	Advantages	Disadvantages
Model1 Single Center	· Minimize data transmission (minimize risk of transmission process)	· Cheating can be prevented · No structure to guarantee the anonymization of personal information · Rely entirely on the honesty of data users
Model2 Firewall single center	· As a single institution, supervision is easy and efficient · Effective for evaluating the quality and bias of data linkage · Possible to collect meta-data and programs to increase understanding of administrative data sets.	· Provide structural barriers to prevent fraud · As a single institution, supervision is easy and efficient · Effective for evaluating the quality and bias of data linkage · Possible to collect metadata and programs to increase understanding of administrative data sets.	· Although it provides a barrier against fraud, it is not externally visible · It depends on whether the research center strictly complies with the functional separation (firewall) · Increased risk during data transmission
Model3 TTP	· Providing a visible structural security barrier to prevent fraud · Structure to guarantee anonymization of personal information · It is possible to collect metadata and programs to increase understanding of administrative data sets.	· Management supervision is difficult and inefficient due to the involvement of many institutions · Increased risk during data transmission · Measurement and quality of linkage cannot be measured because two organizations are linking
Model4 Multilateral security contract	· Providing a visible structural security barrier to prevent fraud (no access to personal level data) · No personal information transmission (Analysis method can be controlled so that only the matrix that preserves privacy is transmitted and the result is not revealed)	· Computer / statistics still under development · There are still parts that cannot be analyzed

III. 국내 빅데이터 플랫폼 및 센터구축 사업개요 및 가명데이터 활용의 한계

2019년 정부는 공공과 민간이 협업하여 보유하거나 외부에서 수집한 데이터를 플랫폼에서 분석, 유통하고, 혁신서비스를 발굴하는 등의 데이터 생태계 조성을 위한 '빅데이터 플랫폼 및 센터구축' 사업을 추진하였다. 해당 사업은 3개년 일정으로 10개 플랫폼 및 100개의 센터를 지정하여 총 1,516억이 투자되고 첫해 641억의 예산으로 진행된 대규모 사업이다. 해당 사업의 성과를 통해 그동안 시장에 개방된 적이 없는 1,400여 종의 새로운 데이터와 17종의 데이터 기반의 혁신 서비스가 실시될 예정이다.

그러나 해당 사업을 통해 제공되는 데이터는 금융 빅데이터 플랫폼의 DUA(data usage aggrement) 형 분석 서비스와 일부 뉴스미디어 컨텐츠 등을 제외하면 모두 집계형 데이터만이 대상이다. 집계형 데이터의 경우 명확한 이용 목적 하에 현황 파악을 하는 기술통계량 분석에는 유용할 수 있으나, 모델을 만들거나 AI학습용 알고리듬을 개발하는 데이터로서는 한계가 있다. 머신러닝, 딥러닝 기술에서 추론과정은 입력된 데이터 값과 모델 파라미터 값의 수치 연산으로 이루어지는데, 현재 빅데이터 거래 플랫폼에 게시된 데이터들은 단편적이거나 업데이트가 되지 않는 통계형으로 학습을 지속하기에는 부적절한 구조를 가진다. 이에 데이터세트 내에 속성인자에 대한 정보를 포함한 구조가 필요하며, 지속적인 업데이트가 필요하다. 이러한 원본형 데이터가 외부에 제공될 수 있다면, 이를 가공하는 사업자, 서비스를 개발하여 제공하는 사업자 및 분석을 통해 인사이트를 제공하는 컨설팅 등의 분야에서 활성화가 가능한 생태계가 만들어질 것이다. 이는 기존 플랫폼의 양면성을 넘어서는 다면성 생태계 구조(multi-side platform)로 변모될 수 있음을 의미한다[11].

그러므로 정부가 추진한 10대 플랫폼이 활성화 되고 이에 대한 지속 가능성을 확보하기 위해서는 해당 플랫폼에서 다루어지는 데이터의 구조 자체가 달라져야 하며, 이를 위해 적절한 수준의 플랫폼 보안에 대한 검증과 인증이 병행되어야 한다.

데이터 관점의 정비를 위해서는 현재 10개의 개별 플랫폼간 메타데이터의 표준화 관리가 필요하다. 또한 오픈소스형 데이터 관리, 배포 플랫폼인 CKAN(Comprehensive Knowledge Archive Network) 기반의 데이터 카달로그에 대한 플랫폼 간 공유를 통해 보다 빠르고 손쉬운 사용자 검색 환경이 제공되어야 할 것이다.

그리고 데이터 3법이 발효되는 2020년 8월 5일 이전에 데이터 활용에 대한 적정성 및 위험 측정기준 등 관련 데이터가 플랫폼을 통해 거래되는 것이 적정한가의 평가 프로세스에 대한 정비도 이루어져야 할 것이다.

플랫폼 인프라 관점에서는 법을 통해 정해진 결합전문기관의 수행 역할 확대에 대한 준비가 이루어져야 한다. 현재 공공기관이 수행하는 가명데이터 결합기관의 역할을 빅데이터 플랫폼이 수행하기 위해서는 여러 가지 사전 준비사항이 필요하다. 그리고 결합데이터의 활용 및 시장 저변을 확대하기 위해서 기존 빅데이터 플랫폼 중 적정한 기준을 충족하는 곳에 대해서는 해당 권한을 부여할 필요가 있다.

마지막으로 현재 개인정보보호위원회로 일원화된 보호기구에서 볼 수 있듯이 제도 및 정책 관점에서 데이터 산업 진흥을 위한 각 부처의 역할 정비도 필요하다. 현재 금융기업은 금융위원회가, 정보통신망법을 준용하는 기업은 과학기술정보통신부가, 의료영역에 대해서는 보건복지부 위주로 각각 정책 및 제도를 준비 중이나, 이종 데이터간 결합이나 성공적인 신산업 창출을 위해서는 서로 엇박자가 나는 제도 및 정책을 조율하기 위한 일원화된 진흥기구의 필요성도 제기된다.

IV. 결 론

본 연구에서는 기업에서 가명 데이터를 활용하기 위해 기술적으로 준비해야 할 사항과 그 과정에서 발생하게 되는 불확실성과 관련된 관리 프로세스 등을 도출하기 위해 GDPR의 적용 기준 및 국내 빅데이터 플랫폼 사업의 현황을 살펴보았다.

국내 데이터 산업 생태계의 조성 및 활성화를 진행하고, 정부의 10대 플랫폼 및 센터 구축사업 등 진정한 다면성 플랫폼으로의 진화를 위해서는 데이터 이용 환경과 관련하여 다음과 같은 준비가 필요하다.

우선 가명데이터의 수집, 분석, 활용 환경에 대한 정비가 고려되어야 한다. 기존 내부 목적으로 활용되던 식별계 데이터 레이크 영역과 데이터의 대외 가치화를 위한 가명데이터 활용 환경을 분리하여 관리해야 할 필요가 있다. 물론 이러한 인프라는 개정된 데이터 3법에 따라 외부 클라우드를 활용함으로써 구현이 가능할 것이다.

다음으로 거래 및 유통이 가능한 적정 수준의 가명화에 대한 명확한 기준이 마련되어야 한다.

데이터3법에 규정되어 있는 가명정보 및 그 기술적 기반이라고 할 수 있는 가명처리는 개념적으로 볼 때 종래 통계처리나 익명화 처리와는 다른 수준의 개념임에도 불구하고, 정책 현장에서는 사실상 발생 가능한 리스크로 인해 익명수준의 데이터 활용으로 그 방법론이 논해지고 있으며, 논자에 따라 이를 이해하는 방식이 각기 다르다. 그러므로 데이터에 대한 가명화 처리방법에 대한 기준 마련 및 수준에 대한 사회적 합의가 필수적이다.

이를 위해 개인정보 처리자에 대한 수준과 이용목적 및 환경에 따른 맥락 요소를 반영하여 비식별화 수준이 결정되어야 할 것이다. 비식별화 단계에서 가명과 익명에 대한 기준과 관련하여 HIPAA의 세이퍼 하버(safe harbor)와 같이 직접 식별자에 대한 명확한 기준을 제시하거나, 위험도 측정에 대한 전문가 결정방법으로 진행할 것인지[12]에 대한 명확한 가이드라인이 제공되어야 할 것이다. 다만, 현재 국내환경 및 시민사회의 정서 상 당분간은 전문가 결정방법으로의 진행이 보다 현실성 있는 대안으로 생각된다. 이를 위해 비식별조치 전문가 양성 등이 병행되어야 한다.

그림 3에서 제시된 내용은 미국의 HITRUST 산하 Privacy Analytics라는 기관의 리스크 측정 모델로 대외 제공형 데이터 활용 시 적용된다. 수치 0.5는 k익명성의 k=2 수준의 비식별화 처리이고, 0.05는 k=20 수준의 비식별화 처리를 의미한다[13]. 미국 의료기관에서는 전문가의 비식별 인증심사 시 0.05 수준 정도를 최대 위험도로 설정하여 비식별 조치를 수행하며, 실제 현장에서는 0.5~0.13 사이의 수준으로 준식별자에 대한 비식별 조치 수준을 적용 중이다.

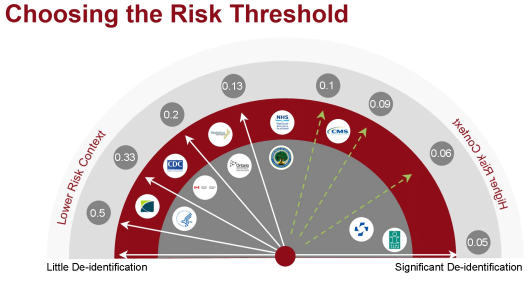

Fig. 3. Data Risk Measurement of Privacy Analytics

마지막으로, 데이터의 가명처리 및 결합대행을 위한 제도 기반 및 기술적, 물리적, 관리적 기준이 필요하다. 현행 데이터의 비식별화는 '비식별 조치 가이드라인'에 기재된 방식에 따라 기업 자체적으로 수행하여야 하나, 전문 인력이 부족하고 대용량 데이터의 경우 별도의 솔루션이 요구되는 한계가 존재한다. 또한 결합을 위하여 민관이 지불해야 할 시간적, 비용적 부담도 만만치 않은 것이 현실이다. 그러므로 시행령에 반영될 데이터 결합기관에 대한 지정과 적용 프로세스는 데이터 경제 활성화를 충분히 고려하는 방법이어야 한다.

데이터 활용에 리스크 제로는 없다. 창과 방패의 관계처럼 보호와 공격의 기술은 나날이 발전하고 있다. 데이터를 안전하게 활용하기 위한 제도적 장려와 더불어 문제가 생겼을 경우 즉시 대응이 가능한 명확한 규칙이 필요할 뿐이다.

현재 우리나라는 명확치 않은 데이터 활용에 대한 법적 근거 등으로 시장이 위축되어 있고, 시민단체의 고발 사건 등에 따라 데이터 활용에 대한 리스크가 과대하게 부풀려질 수밖에 없는 상황이다. 그러므로 가명정보를 활용함에 있어 과학적, 기술적, 사회적으로 합의 가능한 수준의 명확한 기준을 수립, 적용하는 부분이야말로 데이터의 활용 관점에서 매우 중요한 요소임에 분명하다.

이미 개정된 개인정보보호법 상 개인정보 재식별화 및 유출에 대한 부분은 정보 처리자에 대한 형사적 책임과 매출 3%에 달하는 과징금의 징벌적 배상을 요구하고 있다. 개인정보 오남용이라는 리스크에 대한 몫은 오롯하게 정보처리 주체에게 있다. 이에 따라 보다 신뢰사회의 자율적인 책임을 전제로 어렵게 도입된 가명 데이터의 공익적, 산업적 활용이 활발하게 일어날 수 있는 기반 마련이 우선시 되어야 할 것이다.

References

[1] El Emam, Khaled, and Luk Arbuckle. "Anonymizing health data: case studies and methods to get you started," O'Reilly Media Inc.,p.5-6 2013

[2] Mourby, Miranda, et. al., "Are 'pseudonymised' data always personal data? Implications of the GDPR for

administrative data research in the UK," Computer Law & Security Review 34.2, pp. 222-233, 2018
〔3〕 Amendment to the Privacy Act 2, 2020
〔4〕 Mourby, Miranda, et. al., "Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK," Computer Law & Security Review 34.2, pp. 222-233, 2018
〔5〕 Kyung-tae Park and Se-hyun Kim, "A Study on The Preference Analysis of Personal Information Security Certification Systems : Focused on SMEs and SBs," Journal of Information Security and Cryptology, 24(5), pp. 911-918, 2014
〔6〕 S. Garfinkel, De-Identification of Personal Information, NISTIR 8053, 2015. 10
〔7〕 ENISA(European Union Agency for Network and Information Security), Recommendations on shaping technology according to GDPR provisions, An overview on data pseudonymisation, Nov. 2018
〔8〕 Geun-hye Kim, "An Exploratory Study on the Transition of the Regulation System for the Introduction of the Fourth Industrial Revolution Technology," Korean Journal of Information and Chemistry 20(3), pp. 59-88, 2017
〔9〕 Institute for Information Human Rights, "A Study on the Introduction of Data Linkage Assistance System," 2017
〔10〕 Green, E., Ritchie, F., Mytton, J., Webber, D. J., Deave, T., Montgomery, A., Chowdhury, S., "Enabling data linkage to maximise the value of public health research data," pp. 54-60, 2015
〔11〕 Evans, David S., et al. Platform economics: Essays on multi-sided businesses, Competition Policy International, 2011.
〔12〕 Cheol-Joong Kim, Kwang-Soo Lee, Pil-Woo Lee, Eui-Jin Moon, Byung-Joo Song, Kyung-Taek Song, and Soon-Seok Kim, "A Study on the Information of Domestic Non-Identification for Secondary Use of Medical Information," Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, 6(8), pp. 15-23, 2016
〔13〕 El Emam, Khaled, "Principle of De-identification', Privacy Analytics, an IQVIA Company, pp. 10-11. 2019

〈저자소개〉

김 정 선(Jung-Sun Kim) 정회원
2015년 2월: 이화여자대학교대학원 디지털미디어 박사
1999년~현재: SK텔레콤 AIX센터 재직 중
〈관심분야〉 빅데이터, AI, 데이터경제, 비식별화, 가명화