통계연구(2025), 제30권 제1호, 28-60/ DOI https://doi.org/10.22886/jkos.2025.30.1.28

안전한 데이터 연계와 활용을 위한 다자간 피처 스토어(Multi-Party Feature Store) 구축 및 운영에 관한 논의

박민정[1] . 정성규[2]

요약

데이터 연계·활용·거래를 위한 플랫폼으로서 다자간 피처 스토어(multi-party feature store)의 개념과 운영 방안을 제안한다. 다자간 피처 스토어는 현재 국내의 여러 제약조건을 반영하여 최신 개인정보보호 강화 기술(Privacy-Enhancing Technologies; PET)을 사용한 플랫폼으로, 여러 기관의 데이터를 연계한 마이크로데이터 공유를 지원하여 고도화된 통계 분석을 가능하게 하고 인공지능 모델 개발을 지원한다. 이 플랫폼은 사용자 중심의 데이터 접근성과 효율성을 대폭 향상하는 것을 목표로 하며, 4가지 주요 특성을 가진다. 첫째, 사용자는 데이터센터를 방문하여 민감한 원본 데이터에 직접 접근하기 이전에, 재현자료(synthetic data)를 활용한 데이터 트윈(data twin)을 통해 제약 없이 편리하게 사전 분석을 수행할 수 있어 시간과 비용을 절감할 수 있다. 둘째, 통계청의 통계등록부 등 모집단에 대한 기준 데이터(hinge data)를 활용하여 다기관 데이터를 효율적으로 연계함으로써, 데이터 통합과 고도화된 분석 및 머신러닝 모델 개발을 지원한다. 셋째, 일련의 과정 이후, 사용자는 보안 클라우드 환경에서 안전하게 연계된 데이터를 반복적으로 분석하고 그 결과를 API 형식으로 구독할 수 있다. 넷째, 데이터 제공자와 플랫폼 운영 기관을 위해서는 제공된 데이터와 피처(속성)의 가치 및 수요에 따라 보상이 이루어지는 체계를 구축하여 지속 가능한 데이터 플랫폼 생태계를 조성한다. 본 논문에서는 다자간 피처 스토어를 제안하게 된 배경인 데이터 연계·융합 및 활용을 저해하는 각종 요인과 이를 해결하는 최신의 개인정보보호 강화 기술을 먼저 논의한 뒤, 통계청이 현재 보유한 자원과 최신 기술을 활용한 다자간 피처 스토어 플랫폼의 청사진을 제시하고, 운영 방안 및 기대효과에 대해 논의한다.

주요용어 : 프라이버시 강화 기술(privacy enhancing technologies), 데이터센터, 데이터 트윈(data twin), 재현자료(synthetic data), 기준 데이터(hinge data), 통계등록부(statistical register), 다자간 피처 스토어(multi-party feature store)

1. 서론

우리는 오랜 기간 ‘4차 산업혁명’이나 ‘21세기의 원유인 데이터’ 등의 표현을 들어왔다. 4차 산업혁명은 인공지능(AI), 사물인터넷(IoT), 로봇 기술, 드론, 자율주행차, 가상현실(VR) 등이 주도하는 차세대 산업혁명을 일컫는다. 많은 사람이 4차 산업혁명을 통해 구현되는 다양한 신기술을 바탕으로 우리 생활이 혁신적으로 변화할 것이라고 예상한다. 일례로, 최근에는 챗GPT 등을 통해 혁신적인 AI 서비스가 가능할 수 있음을

1) 교신저자. 대전시 서구 한밭대로 713, 통계청 통계개발원 연구기획실, 서기관. E-mail: mjstat@korea.kr
2) 서울시 관악구 관악로 1, 서울대학교 자연과학대학 통계학과, 교수. E-mail: sungkyu@snu.ac.kr

경험하고 있다. 이러한 4차 산업혁명의 핵심 요소인 신기술들은 대부분 방대한 양의 데이터를 기반으로 개발된다.

다양한 분야에서 혁신적인 발전을 이루고자, 사람들은 데이터를 수집·정제·분석·활용하기 위해 노력하고 있다. 마치 과거에 원유를 정제하고, 이를 활용해 동력을 생산하고, 그 동력으로 자동차 등 각 생산 분야에서 신기술을 구현했던 것과 유사하다. 이뿐만 아니라 원유를 정제하여 나일론, 아크릴 등을 만드는 합성섬유 산업, 혹은 플라스틱 등을 생산하는 석유화학 산업처럼, 데이터를 기반으로 하는 새로운 산업의 출현을 기대하기도 한다. 즉, 과거에 원유를 이용해 다양한 유형의 산업 발전을 이루었던 것처럼, 산업 각 분야에서 데이터를 자원으로 활용하여 새로운 비즈니스 창출을 꾀하고 있다. 소셜 미디어나 검색엔진 등에서 생성되는 텍스트 데이터를 수집·정제·활용하여 구축한 '초거대 언어 모델(LLM)'은, 데이터의 활용 가능성과 그 가치를 충분히 입증하고 있다.

그러나 현실적으로는 산재한 데이터를 연계·융합하여 활용하는 데 많은 어려움을 겪고 있다. 쉽게 크롤링이 가능한 온라인 텍스트 데이터를 수집할 수 있었지만, 보다 정교하고 많은 정보를 가지고 있는 건강·질병 관련 자료, 세금 기록, 신용카드 데이터 등에는 일반 이용자의 접근을 허락하는 것조차 쉽지 않다. 대국민 서비스 과정에서 생성되어 정부와 공공기관이 보유하는 공공데이터마저도 국민에게 온전히 개방하기에는 많은 걸림돌이 있다. 즉, 데이터 연계는 물론이고, 연계 이전 단계인 민감한 고부가가치 데이터에 대한 원활한 접근이나, 개별 공공데이터의 전면 개방조차도 쉽지 않은 상황이다.

이 논문에서는 현재 우리가 따라야 하는 여러 제약조건을 고려하여, 데이터를 연계·융합하여 활용할 수 있는 새로운 플랫폼 모델을 제안하고자 한다. 제4절에서 소개하는 신규 플랫폼 모델은 '다자간 피처 스토어(muti-party feature store)'이며, 이 플랫폼은 ①재현자료(synthetic data)를 활용하는 '데이터 트윈(data twin)', ②데이터의 연계를 원활하게 하는 '기준 데이터(hinge data)'의 활용, ③원자료의 연계가 이루어지는 데이터센터(망분리 보안구역), ④선택된 피처에 대한 안전한 분석이 이루어지는 보안 클라우드(cloud) 공간으로 구성된다. 이러한 본 논문의 핵심 주제는 제4절에서 다룬다.

제안하는 새로운 모델을 설명하기 위하여, 먼저 제2절에서는 데이터 연계·융합 및 활용을 저해하는 각종 요인을 분석하고, 제3절에서는 이러한 걸림돌을 회피할 수 있게 돕는 신기술인 '개인정보보호 강화 기술(privacy enhancing technologies; PET)'을 UN이 발간한 가이드라인을 기반으로 설명한다. 나아가 이러한 기술을 반영하기 위한 우리나라 통계청의 최근 노력과 성과를 소개한다. 즉, 2절과 3절에서는 4절의 다자간 피처 스토어를 제안하게 된 배경 및 관련 기술을 설명하여 다자간 피처 스토어의 필요성, 당위성 및 가능성을 제시한다. 마지막 5절에서는 결론을 제시한다.

2. 데이터 연계·활용 저해 요인 분석

데이터를 활용하기 위한 많은 요청에도 불구하고 데이터를 연계하여 활용하는 것은 현실적으로 매우 어렵다. 이 절에서는 주요한 원인을 여러 측면에서 살펴본다. 우선

데이터센터와 관련된 데이터 거버넌스 및 법제도 문제를 살펴보고, 데이터 공급에 대한 보상 측면의 이슈를 알아본다. 다음으로 기술적 측면에서 개인정보보호 이슈와 데이터 표준화의 문제를 정리한다. 마지막으로 이용자의 편의성 문제를 논의한다.

참고로 본 논문에서는 마이크로데이터(microdata)라는 용어를 사용하도록 한다. 개인, 기업 등 개체 수준으로 구성된 정제된 원본 데이터를 마이크로데이터라고 하며, 이는 일반적으로 기관이 보유한 관계형 데이터베이스 시스템에서 추출하는 정형 데이터 형식을 가진다. 공학 분야에서는 이를 테이블 데이터(table data)라고 지칭하기도 한다.

2.1 데이터 거버넌스 및 법제도 문제

현재 우리나라에서 외부에 공개하기 어려운 건강·질병, 세금, 신용점수 등에 관한 마이크로데이터를 분석하고자 할 때는 해당 기관의 데이터센터를 방문해야 한다. 여기서 데이터센터란 망분리 보안 구역으로, 이용자가 정해진 업무 시간에 방문하여 데이터를 분석할 수 있는 공간이다. 대부분의 데이터센터는 반출 심의 절차를 운영하여 이용자의 데이터 분석 결과 반출 여부를 결정한다. 따라서 마이크로데이터를 물리적으로 보호하며, 반출물에 대해서도 철저히 심의하므로 혹시라도 발생할 수 있는 개인정보 누출을 차단한다. 즉, 데이터센터는 개인정보 보호법에 명시된 ‘안전성 확보에 필요한 기술적·관리적·물리적 조치’를 구현하였으므로, 현재의 법제도하에서 민감한 정보를 포함한 마이크로데이터를 제공할 수 있는 가장 일반적인 수단이다.

데이터 보유기관들은 데이터 분석 수요에 대응하기 위하여 각자의 다양한 명칭으로 데이터센터를 운영해 왔으나, 망분리 보안구역이라는 운영상 특징은 동일하다. 우리나라의 경우, 통계청, 국세청, 건강보험공단, 국민연금공단 등 많은 기관이 데이터센터를 운영한다. 현실적으로 데이터센터는 민감한 마이크로데이터를 일반 이용자가 분석할 수 있도록 제공하는 거의 유일한 수단이었으며, 해외의 다양한 데이터 보유기관들도 우리와 유사한 방식으로 오랫동안 데이터센터를 운영해 오고 있다.

우리나라의 데이터센터는 데이터 도메인 별로 서로 다른 법령의 적용을 받는다. 신용정보 데이터의 경우 신용정보법의 적용을 받으며, 신용정보 데이터를 제공하기 원하는 데이터센터는 ‘데이터 전문기관’으로 지정받아 운영해야 한다. 그 외의 일반적인 데이터센터는 데이터산업법의 ‘데이터 안심구역’으로 지정받아 운영한다. 최근 개인정보보호위원회는 ‘개인정보 안심구역’ 지정을 시작하기도 했다. 전문기관 혹은 안심구역으로 지정받은 데이터센터는 자신이 보유한 데이터뿐만 아니라 다른 기관의 데이터도 함께 제공할 수 있다. 모든 분야의 데이터를 제공하기 위해서는 법령별로 모두 지정받아야 한다. 예를 들면 통계청의 통계데이터센터는 데이터 전문기관 및 개인정보 안심구역으로 지정받았고, 한국데이터산업진흥원의 데이터 안심구역 서울 센터에 입주해 있다.

다만, 안심구역으로 지정받아 서로 다른 데이터 보유기관의 데이터를 제공할 수 있는 데이터센터라도 각각의 데이터를 모두 연계할 수 있는 것은 아니다. 예를 들면 데이터 안심구역으로 지정받은 데이터센터에서 건강보험심사평가원의 마이크로데이터와 통계청의 마이크로데이터를 각 기관에게 할당된 서로 다른 컴퓨터를 통해 제공할 수

있지만, 이를 한 컴퓨터에 제공하여 이용자가 이들을 연계하여 분석할 수는 없다. 서로 다른 데이터 보유기관의 데이터를 연계하여 분석하기 위해서는, 신뢰받는 제3의 결합전문기관을 통한 가명정보결합절차를 거쳐야 한다. 일반적으로 데이터센터의 자체결합은 허용되지 않고, 가명결합한 데이터에 대해 목적 외 이용금지 조항도 적용된다.

종합하면, 우리나라에서는 데이터 도메인별로 데이터 거버넌스가 다양하고, 서로 다른 개인정보처리자가 보유하고 있는 마이크로데이터를 연계·분석하기 위해서는 제3의 결합전문기관을 통한 가명결합 절차를 준수해야 하며, 그 과정은 매우 복잡하다. 또한 가명결합된 데이터를 사용하는 데도 여러 제약조건이 있다.

한편, 기관이 보유한 데이터를 특정 데이터 안심구역에 제공하는 것도, 경우에 따라 개별법령에 따른 제약이 있다. 예를 들어, 대중교통법을 준수하기 위해서는 대중교통 승하차 정보를 데이터 안심구역에 제공할 수 없다. 대중교통법 제10조의9의 제2항이 “교통카드데이터를 제공하는 경우 집계자료 형태(제10조의8에 따라 제출받은 자료를 분류・합계・변형하는 등 통계처리하여 가공한 형태를 말한다)로 제공하여야 한다. 다만, 국가, 지방자치단체 및 교통 관련 연구기관이나 공공기관으로서 대통령령으로 정하는 기관이 교통 관련 정책수립, 업무수행, 통계작성 및 학술연구 등의 목적으로 요청하는 경우에는 그러하지 아니하다.”라고 명시하고 있기 때문이다. 따라서 법조문에 규정되어 있지 않은 일반적인 이용을 위하여 해당 마이크로데이터를 데이터센터에 제공하는 것은 현재의 법 규정으로는 허용되지 않는다. 대중교통 승하차 정보 원자료가 인공지능을 활용한 빠른 길 안내 등, 민간 산업 발전에 요긴할 것이 예측되더라도 현재 법 규정으로는 제공할 수 없다.

데이터센터는 상세한 수준의 마이크로데이터를 일반적 이용을 목적으로 제공할 수 있는 거의 유일한 수단이나, 데이터 거버넌스가 다양하고, 관련 규제가 복잡한 측면이 있다. 데이터센터를 직접 이용해 본 적이 없는 사람들은 데이터센터를 통해 민감한 데이터 제공·연계·활용 문제를 손쉽게 해결할 수 있다고 판단할 수도 있지만, 실상은 그렇지 않다. 데이터 거버넌스와 법제도적 문제 해결에 관한 광범위한 논의는 본 논문의 범위를 벗어나며, 이 소절에서는 기술적인 해법을 논의할 때 필요한 데이터센터 관련 거버넌스 문제를 일부 정리했다.

2.2 데이터 공급에 대한 보상과 윤리

시장경제 체제하에서 자원이나 상품 생산에 대하여 적절한 보상이 이루어지는 것은 자연스러운 일이다. 그러나 개별적으로 존재할 때는 쓸모가 없고, 대량으로 모여야만 가치가 있는 종류의 데이터[3]에 대하여 보상 방안을 도출하는 것은 쉽지 않다. 데이터를 구성하는 개인이나 개별 기업의 권리에 대한 목소리도 높지만, 개별 자료의 가치를 산정하고 일일이 보상하는 것은 현실적으로 어렵기 때문이다. 또한 실제로 데이터를

3) 마이데이터 서비스에 활용되는 개인의 데이터, 렌터카 정보 조회에 활용되는 특정 자동차에 대한 자동차등록정보 데이터 등과 같이 개별적 데이터가 중요한 경우도 있다. 개별적 데이터가 중요한 경우를 ‘정보성 데이터’라고 지칭한다면, 본 논문이 다루는 대량으로 모여야 가치가 있는 종류의 데이터는 ‘통계성 데이터’라고 할 수 있다.

제공하는 주체는 데이터를 수집한 기관이므로, 대가를 지불받는 대상자를 누구로 지정할 것인가에 대한 문제도 있다. 다른 한편에서는, 데이터를 거래의 대상으로 보는 것이 적절한지에 대한 근본적인 논의가 이루어지고 있기도 하다.[4)]

데이터 거래나 보상의 개념은 최근 다양한 측면에서 본격적으로 논의되기 시작했으며, 실제로 데이터 거래가 시작된 초창기에는 데이터 품질에 대한 고려보다는 데이터 용량에 따라 가격을 산출하는 등의 단순한 방식이 고려되었다. 데이터 거래에 관한 일부 국가별 사례를 살펴보면 다음과 같다. 우선 최초의 데이터 거래소는 2015년 설립된 중국의 귀양 빅데이터거래소이다. 귀양 거래소에서는 공공데이터를 위주로 데이터 유통이 이루어지고 있었으나, 설립 4년차를 기준으로 3천여 개의 기업회원이 의료, 금융, 전자상거래, 에너지, 교통 등, 30여 개 분야의 민간 데이터도 판매하고 있다. 최근 중국에서는 지방 정부뿐만 아니라 기업 주도의 빅데이터 거래 플랫폼도 발전 중이다.

다음으로 미국의 경우에는 민간을 중심으로 금융데이터의 공유가 활발하게 이루어지고 있다. 민간단체인 금융데이터거래소(financial data exchange, FDE)가 설립되었으며, FDE는 개인 금융 데이터의 안전한 공유를 목적으로 대형 은행과 핀테크 기업 등이 제휴하여 운영되고 있다. 또한 미국에서는 데이터 브로커(data broker)가 합법적으로 활동하고 있으며, 연방정부와 주 정부의 공개된 정보, 블로그 등의 공개된 정보, 금융회사 보유 데이터 등을 연계·가공하여 판매한다. 데이터 브로커 사이의 거래도 이루어진다.

우리나라에도 미국과 유사하게 2020년 금융보안원이 출범한 금융데이터거래소가 운영되고 있어, 민간 데이터 거래가 본격적으로 이루어지기 시작했다. 통계청 통계데이터센터와 같이 민간 데이터를 일부 구매하여 이용자에게 비교적 낮은 이용료를 받고 제공하는 데이터센터도 있다. <표 2.1>은 유료 여부와 상관없이 국내에서 운영되고 있는 몇몇 데이터 제공·거래 플랫폼을 소개한다. 이 외에도 교통, 통신, 삼림, 농식품, 소방 안전 등 다양한 분야에서 각각의 플랫폼이 존재한다. 이처럼 데이터 거래를 위한 플랫폼이 활성화되고 있는 만큼, 각 플랫폼에서도 필요할 뿐만 아니라, 데이터 연계·활용에 있어 필수적인 요소인 '데이터 거래 시 가치평가와 보상기준'에 대한 논의와 실제적인 해결 방안 모색이 더 적극적이고 구체적으로 이루어질 필요가 있다. 관련 논의는 4.5절에서 이어가도록 한다.

4) 마이클 샌델(2012년)의 "돈으로 살 수 없는 것들(무엇이 가치를 결정하는가)" 및 빅토어 마이어 쇤베르거와 토마스 람게(2018)의 "데이터 자본주의" 등을 참고할 수 있다.

<표 2.1> 국내 빅데이터 플랫폼 사례

명칭	운영 주체	역할
문화 빅데이터 플랫폼 (www.bigdata-culture.kr)	한국문화정보원	도서, 체육, 예술, 숙박, 레저, 음식 등 다양한 문화 분야의 고품질 데이터를 개방해 데이터 유통·거래 생태계 조성
한국데이터거래소 (kdx.kr)	MBN, 삼성카드, CJ올리브네트웍스, SK텔레콤 등	국내 최초('19.12 출범) 민간 데이터 거래소로 유통·소비 분야의 빅데이터 플랫폼을 제공
국가암데이터센터 (www.cancerdata.re.kr)	국립암센터, 보건복지부	암 정복을 위한 빅데이터를 공개하여 연구자와 의료진의 활용을 지원
암 빅데이터 플랫폼 (www.bigdata-cancer.kr)	국립암센터	10대 암종별 임상데이터를 수집·제공하여 암 진단 및 치료 연구를 지원
환경 빅데이터 플랫폼 (www.bigdata-environment.kr)	한국수자원공사	생활환경과 자연환경 데이터를 수집·제공하여 맞춤형 수질정보 서비스 및 대기질 야외활동 추천 서비스 등 제공
금융권 공동 데이터 플랫폼 (www.datop.or.kr)	금융권 공동 구축	금융 데이터를 수집·분석해 금융 서비스 혁신과 데이터 기반 의사 결정을 지원
금융 빅데이터 플랫폼 (www.bigdata-finance.kr)	비씨카드	보험, 증권, 통신 등의 데이터를 수집·제공해 소상공인 창업지도 서비스 및 국민 금융생활 플래너 서비스 등 제공

2.3 개인정보보호 이슈

이 소절에서는 서로 다른 기관이 보유한 마이크로데이터를 연계하여 활용하고자 할 때, 개인정보보호 문제가 어떻게 걸림돌이 되는지 기술적인 측면에서 살펴보고자 한다. 많은 사람이 마이크로데이터를 외부에 직접 공개하면 개인정보 노출이 일어날 수 있음을 인지하고 있다. 따라서 노출위험(disclosure risk)이 큰 마이크로데이터를 외부에 공개할 때는, 유용성(utility)이 낮아지더라도 자료를 일부 변형하여 노출위험을 낮추는 비식별화(de-identification) 조치를 한다. 원자료 그대로 이용할 수 있도록 제공하기 위해서는 데이터센터를 이용한다. 망분리 보안구역인 데이터센터를 이용하면 개인정보 보호법에 명시된 '안전성 확보에 필요한 기술적·관리적·물리적 조치'를 이행하는 것이기 때문이다.

따라서, 재사용 금지 조항 등 법적인 제약 사항이 해소되었다고 가정한다면, 단일 마이크로데이터뿐만 아니라 가명정보결합절차를 통해 연계한 서로 다른 기관의 마이크로데이터까지도 데이터센터를 통해 이용자에게 안전하게 제공할 수 있다고 판단할 수 있다. 그러나 다음의 간단한 사례는 특정 마이크로데이터 자체가 아니라, 데이터센터를 통하여 그것을 분석한 결과만을 반출한다고 하더라도 개인정보보호 이슈가 여전히 발생할 수 있음을 보여준다.

이 예제의 마이크로데이터는 성별, 연령, 교육정도, 혼인상태, 집계구 번호로 구성되어 있으며, 집계구는 통계작성을 위한 아파트 1~2개 동 혹은 인구 500명 정도의 작은 구역을 말한다. 특정 집계구에 대해서 여러 조건을 만족하는 빈도수를 산출하고, 데이터센터는 익명성 확보를 위해서 5미만의 빈도수를 NA 표기하여 반출을 허가했다고 가정하자. 즉, NA로 표기된 값은 1~4명일 수 있어 불확실성을 가지므로, 노출위험을 감소시킨다고 판단한다. <표 2.2>에서 왼쪽 표는 반출을 신청한 분석 결과물을 보여준다. 개인정보를 노출시키지 않기 위하여 마이크로데이터 자체를 공개하지 않고, 절차가 다소 복잡하지만 데이터센터를 통해 마이크로데이터를 분석할 수 있게 제공하고, 왼쪽 표의 집계 결과만을 외부에 공개했다.

<표 2.2> 분석 결과물에서 개인정보 노출 예제

반출 신청한 분석 결과

이용자의 조건 선택	제공 빈도수
남 + 50~65세 + 박사 + 기혼	5
남 + 50~65세 + 박사 + 기혼, 이혼	6
남 + 51~65세 + 박사 + 기혼, 이혼	5
남 + 50세 + 박사 + 이혼	NA

간단한 추론 결과

남, 박사	51~65세	50세	합 (50~65세)
기혼	5	0	5
이혼	0	1**	1
합	5	1	6

그러나 <표 2.2>에서 왼쪽의 분석 결과를 간단히 재구성하면 오른쪽 표의 추론 결과와 같은 정보를 얻을 수 있다. 즉, 이 집계구에 거주하는 남자, 50세, 박사가 1명 존재하고 그의 혼인상태가 이혼임을 알 수 있다. 이 정보는 개인이 식별되고 개인의 민감한 정보가 드러나지 않게 하려고 반출물에서 NA로 처리하여 불확실성을 부여하여 감춘 정보이다. 그러나 위의 간단한 예제는 마이크로데이터 자체를 공개한 것과 다를 바 없이, 분석 결과에서 개인정보가 노출될 수 있음을 보여준다.

이러한 현상을 이해하고 적절한 해결책을 찾기 위해서는 우선 보안(security)과 정보보호(privacy) 문제를 분리하여 이해할 필요가 있다.

보안(security)은 데이터에 대한 사람들의 접근을 기술적으로 관리하는 것으로, 데이터를 외부로부터 보호하고 허가받지 않은 사람들에게는 공개되지 않도록 하는 것이다. 반면 정보보호(privacy)는 누구나 사용할 수 있도록 마이크로데이터 또는 집계 결과를 공개했을 때, 개인의 정보를 보호하는 것이 목적이다. 즉, 데이터 공개 시 개인정보의 노출(disclosure) 방지를 목적으로 한다. 국내에는 정보보호라는 용어를 사용하며 두 개념이 분리되지 않은 경향이 있으나, 보안(security) 기술은 데이터 미공개를, 정보보호(privacy) 기술은 데이터 공개를 전제로 한다는 큰 차이가 있다. 물론 프라이버시의 정의가 자신의 정보에 대한 자기 결정권이므로, 보안(security)과 정보보호(privacy) 기술 모두 프라이버시 보호라는 공통의 목적을 가진다.

개인정보보호 이슈를 해결하는 방안을 찾기 위해서는 보안(security) 조치를 넘어서 정보보호(privacy) 관점에서도 접근할 필요가 있다. 데이터센터가 망분리 보안구역이

라는 강력한 보안(security) 조치를 취하여 개인정보 누출에 대해 안심했지만, 데이터 제공을 어렵게 하는 원인은 정보보호(privacy) 측면에서도 발생하기 때문이다. 이와 관련된 용어 및 최신 기술 활용에 관해서는 3절에서 다시 논의하도록 한다. 참고로 이 소절에서 설명한 보안(security)은 3절의 Input Privacy 분야에서 사용되는 기술에, 정보보호(privacy)는 Output Privacy 분야에 사용되는 기술 및 개념에 가깝다.

2.4 데이터 표준 제정 및 표준화

다른 분야와 마찬가지로, 데이터 분야에서도 표준을 제정하기 위한 꾸준한 노력이 있어 왔다. 표준(standard)이란 표준화를 위한 합리적인 기준으로, 합의에 의해 작성되고 인정된 기관에 의해 승인된 문서(specification)를 말한다. 표준화(standardization)란 최적의 체계를 만들기 위한 공통적이며 반복적인 사용을 위한 규정을 수립하면서, 해당 표준을 관련 생태계에 적용하고 발전시키는 일련의 활동이다. 표준의 제정·개정·폐지는 환경의 변화를 능동적으로 수용할 수 있도록 적극적으로 추진되어야 하며, 표준화가 효율적으로 이루어져서 표준을 생태계에 능동적이며 지속 가능한 형태로 적용하는 것이 중요하다.

적절한 표준 제정과 표준화 이행은 효율적인 데이터 카탈로그 제작을 가능하게 한다. 기업이나 기관이 보유한 데이터 자산은 이질적인 데이터로 구성되지만, 이에 관한 기술적인 세부 내용을 이용자들이 간단하게 이해할 수 있도록 지원하는 시스템을 데이터 카탈로그라고 할 수 있다. 주요 기능은 많은 데이터 중에서 사용할 수 있는 데이터를 쉽게 식별하고 이해할 수 있게 하고, 메타데이터를 투명하게 제공하며, 데이터의 계보나 품질에 대하여 통합된 정보를 제공하는 것이다. 나아가 다양한 소스에서 메타데이터를 수집하여 큐레이션하는 등, 데이터 검색과 탐색을 편리하게 하고 데이터 분석과 시각화 기능까지도 제공할 수 있다. 주지할 것은, 이렇듯 데이터의 효율적인 활용을 돕는 데이터 카탈로그 구현 및 효율적인 활용에 있어 데이터 표준화가 필수적이라는 것이다.

표준을 제정하는 것은 소수의 관계자와 전문가가 모여 일정 기간 작업을 수행하여 일반에 제안할 수 있다. 더불어 환경 변화를 시의적절하게 반영할 수 있도록, 표준을 주기적으로 업데이트하는 것이 중요하다. 표준화를 이행하는 것은 데이터를 다루는 불특정 다수가 제정된 표준을 따르도록 하는 것을 의미한다. 그러나 데이터를 다루는 업무 담당자들이 모두 데이터 전문가일 수는 없으며, 관련 업무에 단기간 종사하는 예도 많다. 또한, 데이터 표준은 낯설고 어려운 측면이 있다. 따라서 다수의 데이터 업무 담당자들을 대상으로 표준화를 수행하게 하는 것은 쉬운 일이 아니다. 표준, 데이터 카탈로그, 데이터맵 등 관련 논의에 관해서는 김학래(2021) 등을 참고할 수 있다.

2.5 데이터 이용자의 비용과 편의성

지금까지 데이터 공개와 관련하여 사회적 관심을 받으며 논의되었던 사항은 주로 개인정보가 노출될 위험이 어느 정도인지, 노출위험을 제어하기 위하여 비식별화 조치를 하면 데이터의 유용성이 얼마나 감소하는지, 노출위험이나 유용성을 어떻게 측

정할지 등이었다. 이에 따라, 법적인 이슈나 거버넌스 문제가 없다는 전제하에, 데이터를 외부에 공개하기로 결정하는 기준은 보통 K-익명성 등을 기준으로 일정 수준 이하의 노출위험을 보장하면서도 일정 수준 이상의 유용성을 가질 수 있는지 정도였다. 공표 기준에 맞도록 노출위험을 낮추기 위해서 비식별화 과정을 수행하며, 이때 유용성이 현저히 낮아져서 데이터가 쓸모없어진다면 데이터센터를 이용하여 원본 데이터를 제공하는 전략을 취한다. 즉, 개인정보 노출의 위험성이나 데이터 유용성을 중심으로 데이터 제공 방식을 결정한다고 할 수 있다.

그러나 사실 이용자의 편의성 확보는 데이터 활용 활성화에 있어 대단히 중요한 요소이다. 데이터센터를 활용하여 원자료를 제공하는 방식은 낮은 위험성과 높은 유용성을 담보하여 바람직하지만, 데이터센터 운영시간에 맞추어 센터가 위치한 특정 지역으로 이동해야 하는 이용자로서는 데이터센터를 통한 자료제공 방식을 편리하다고 느끼기 어렵다. 대부분 데이터센터의 이용자가 급증하고 있지만, 홈페이지 등, 온라인에서 일반적인 통계를 배포하는 방식에 비하면 데이터센터 이용자 수는 절대적으로 작다. 원격분석시스템, 온라인 안심구역 운영 등, 원자료를 온라인 방식으로 분석할 수 있게 제공하려는 노력도 있었지만 그 또한 이용자 수가 많지 않다. 이용자 편의성 측면에서 세련된 서비스가 부족하다는 것이 주요 원인 중의 하나일 것이다.

지금까지는 데이터 제공에 있어 이용자 편의성 문제에 대한 인식이 미비했으나, 사실은 민간의 다양한 IT 서비스와 같이 편의성이 데이터 이용 여부 결정에 있어 매우 중요하다. 본 논문이 제안하는 다자간 피처 스토어라는 플랫폼은 이용자의 편의성을 매우 중요한 요소로 고려하여 설계되었다. 더불어, 2.2절의 데이터 공급에 대한 보상에 관한 문제도 편의성 증진 관점에서 풀어가는 것이 효율적일 수 있다. 관련 논의는 4절에서 이어가도록 한다.

3. 프라이버시 강화 기술

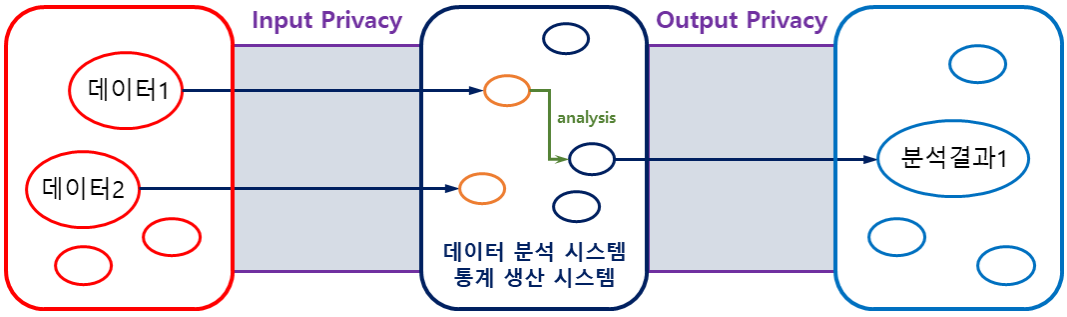
사람들은 개인이나 기업이 생성하는 데이터에 대한 접근을 허용할 때 발생할 수 있는 민감한 정보 유출에 대해 우려한다. 그러나 우리는 이미 은행 업무와 인터넷 데이터 전송 등에 있어 민감한 정보를 다룰 때 암호화를 널리 사용하고 있으며, 이러한 기술에 대한 신뢰성은 높다. 이렇게 일반적인 보안 기술이 널리 활용되고 있음에도 불구하고, 데이터 공유를 위한 각종 신기술인 '개인정보보호 강화 기술(Privacy Enhancing Technologies, 이하 PET)'에 대한 인식은 아직 저조하다. 이 절에서는 PET의 분류, 기존 용어와 관계 정리, 통계청의 신기술 활용과 이용자 편의성 향상을 위한 노력을 설명한다.

3.1 UN PET 가이드라인에 따른 기술 분류

이 소절에서는 UN의 PET 가이드라인(UN, 2023)을 중심으로 OECD의 PET 기술 백서(OECD, 2023)를 참고하여 데이터의 안전한 공유 보장을 위한 PET 분류를 논의

한다. 각각의 기술에 대한 세부적인 설명 역시 UN의 가이드라인 및 OECD 백서의 내용을 인용했다. 다만 이 소절에서는 UN의 가이드라인이 언급하는 세부 기술 중에서 본 논문의 주요 내용과 관련된 사항만 다루도록 한다.

UN의 가이드라인은 PET를 단순하고 간결하게 설명하기 위해, 관련 기술을 크게 '입력 단계의 개인정보보호(Input Privacy)' 및 '출력 단계의 개인정보보호(Output Privacy)'라는 두 범주로 분류했다. 본 논문에서는 '데이터 분석 시스템' 혹은 '통계 생산 시스템'을 가정하고, UN이 제시하는 두 범주를 시스템에 입력하는 데이터와 시스템에서 출력하는 데이터 각각에 대한 데이터 처리 혹은 보안 기술로 해석했다. 아래 <그림 3.1>은 이러한 개념을 나타낸다.



<그림 3.1> Input Privacy 및 Output Privacy 개념도

먼저, '입력 단계의 개인정보보호(Input Privacy)' 범주에 속하는 PET는 서로 다른 데이터 주체가 상대의 데이터를 명확하게 볼 수 없는 상태에서 시스템에 데이터를 입력할 수 있도록 하는 것을 목적으로 하며, 다음과 같은 접근방식이 널리 사용된다.

1. 신뢰할 수 있는 제3자를 찾아 사전 합의된 계약 이용 약관을 집행
2. 암호화 기반 접근방식을 사용

첫 번째 신뢰할 수 있는 제3자 접근방식이란, 데이터를 공유하려는 두 당사자가 함께 신뢰할 수 있는 국가 시스템이나 법인과 같은 제3자를 찾는 것이다. 이때 제3자는 각 당사자에게서 민감한 데이터를 전송받을 뿐만 아니라, 수요자들이 원하는 분석을 수행하는 역할도 담당한다. 그러나 국가에 따라, 민감한 데이터를 제3자와 공유하는 것을 법적으로 금지하는 경우도 있고, 이용자가 원하는 분석을 모두 수행해 줄 전문성을 갖춘 기관도 찾기 어려우므로, UN의 가이드라인에서는 이러한 제3자를 통한 개인정보보호 방식을 일반적인 경우로 논의하기는 어렵다고 정리하고 있다.

그러나 우리나라의 경우, 법적으로 '가명정보 결합전문기관' 제도를 운영하고 있어 여타의 국가와는 조금 다른 상황으로 볼 수 있다. 국내 데이터센터들이 결합전문기관으로 지정받으면, 가명정보 결합의 형태로 데이터를 공유하여 분석할 수 있기 때문이다. 또한 요청 당사자들이 데이터센터를 방문하여 민감한 데이터를 직접 분석하므로, 데이터센터가 분석 전문가를 보유하지 않아도 된다.

두 번째 암호화 기술 활용은 점점 더 주목받고 있는데, 대표적인 기술로는 '안전한 다자간 연산(Secure Multi-Party Computation, sMPC)', '동형 암호(Homomorphic Encryption, HE)' 및 및 '신뢰 실행 환경(Trusted Execution Environments, TEE)'이 있다. 참고로 이 세 기술은 OECD의 PET 기술백서[5]에서도 '암호화된 데이터 처리 도구(encrypted data processing tools)'로 분류하고 있다(OECD, 2023). 이들은 안전성 측면에서 훌륭한 기술이지만, 계산 비용이 많이 발생하는 것은 단점이다.

동형암호에서 발생하는 계산 비용으로는 ①데이터의 용량 확대, ②계산 비용 및 ③분석 솔루션 구축 비용이 있다. 우선, 원본 데이터에 비하여 '동형암호로 변환된 데이터의 용량'은 훨씬 크다. 따라서 대량의 데이터를 동형암호로 처리하기보다는 정확히 어떤 데이터를 암호화할 것인지 신중히 고려하는 것이 중요하다. 다음으로 안전한 다자간 연산과 마찬가지로 '동형암호 처리된 데이터의 계산시간'은 원본 데이터의 계산시간보다 훨씬 길다. 다만 최근에는 더욱 효율적인 알고리즘 개발로 이러한 계산시간의 비율이 줄어들고 있다. 마지막으로 '동형암호 연산 솔루션 구축 비용'을 고려해야 한다. 기존 데이터 분석 프로그램이 수행하는 각종 분석 도구를 대체하기 위한, 동형암호 처리된 데이터를 분석하는 프로그램이 필요하기 때문이다.

'신뢰 실행 환경(Trusted Execution Environments, TEE)'은 Input Privacy 및 코드 보장 문제를 완화하는 최신 CPU의 기능이다. TEE는 CPU의 하드웨어 및 관련 소프트웨어 라이브러리에 부분적으로 구현된다. TEE를 이용하면 데이터 및 데이터를 분석하는 코드에 대한 보안 관리가 보장되므로, 데이터 제공자로서는 데이터를 제공할 때 개인정보 노출위험에 대한 부담이 줄어든다. 한편 주요 단점은 CPU 활용 등에 있어 성능저하가 있을 수 있다는 것이나, 최근에는 기술적 발전으로 이러한 한계가 많이 완화되었다.

참고로 UN의 가이드라인은 이외에도 분산학습(Distributed Learning) 및 영지식증명(zero knowledge proofs)을 Input Privacy를 위한 기술로 소개한다. 또한, UN의 가이드라인에서는 언급되지 않았으나 OECD의 PET 기술백서에서 설명하는 기술 중, '데이터 난독화 도구(data obfuscation tools)' 범주로 소개하는 '국소 차등정보보호(local differential privacy)' 방식도 Input Privacy를 위한 기술로 이해할 수 있다. 영지식증명은 아직 초기 단계의 응용에 머물고 있으나, 국소 차등정보보호는 데이터 수집 단계에서 데이터 제공자의 민감정보를 보호하는 용도로 애플, 마이크로소프트 등의 빅테크기업에서 사용되고 있다.

다음으로, UN의 가이드라인은 '출력 단계의 개인정보보호(Output Privacy)'를 위한 기법으로서 재현자료(synthetic data)와 차등정보보호(differential privacy)를 제시하고 있다. 이때 '출력 단계의 개인정보'란, '입력 데이터를 분석하는 특정 시스템의 산출물 혹은 입력 데이터 분석 결과에서 드러날 수 있는 개인정보'이다. 만약 특정 시스템이

5) UN의 PET 가이드라인이 주요 세부 기술을 Input Privacy 및 Output Privacy라는 2개 범주로 분류하는 것과 달리, OECD의 PET 기술백서는 주요 세부 기술을 ①데이터 난독화 도구(data obfuscation tools), ②암호화된 데이터 처리 도구(encrypted data processing tools), ③연합-분산학습 기법(federated and distributed analytics) 및 ④ 데이터 책임 도구(data accountability tools)라는 4개 범주로 분류하고 있다.

통계생산 시스템이라면, 외부에 공표되는 마이크로데이터도 '출력 단계의 개인정보'에 해당할 수 있다. 따라서 Output Privacy 범주에 속하는 PET는 각종 통계, 개별 이용자의 데이터 분석 결과뿐만 아니라 외부에 공개하는 마이크로데이터에 적용되는 기법도 포괄할 수 있다. Output Privacy를 위한 재현자료 생성 및 차등정보보호 적용을 간단히 살펴보면 다음과 같다.

먼저, 재현자료는 개인정보보호라는 제약조건을 근본적으로 극복하기 위하여 제안되었다고 할 수 있다. 재현자료 생성 목적이 원자료에 있는 어떠한 개별 자료도 포함하지 않지만, 원자료가 가지고 있는 변수들 사이의 관계를 온전히 보존하는 새로운 데이터를 생성하는 것이기 때문이다. 이 방식은 다중대체 기법(Rubin, 1987)을 기반으로 완전 재현자료(fully synthetic data)의 생성 방안을 제시한 Rubin(1993)에서 비롯되었다. 완전 재현자료 생성은 다음과 같은 세 목적을 가진다(박민정과 김정연, 2017).

(1) 개체별 정보 노출 방지 등, 안전성 확보
(2) 타당한 추론을 가능하게 하는 적은 정보손실
(3) 일반적인 통계기법을 쉽게 적용할 수 있다는 의미에서 이용자의 편의성

한편, 노출위험이 높은 일부 값들만 재현하여 만든 자료를 부분 재현자료(Partially Synthetic Data)라고 하며, Little(1993)이 제안했다. 자료의 모든 정보가 민감하다고 보기는 어려운 경우도 많으므로, 모든 변수가 아니라 노출제어 처리가 필요한 일부 변수만 다중대체하자는 것이 핵심 아이디어이다. 대체 대상으로 고려하는 변수는 다음과 같은 두 가지 유형으로 나누어 볼 수 있다(박민정과 김정연, 2017).

(1) 키변수 : 조합하면 유일한 응답자를 식별해 낼 수 있는 응답자의 특성 변수
(2) 민감변수 : 소득, 특정 질병의 유무 등과 같이 응답자가 노출을 꺼리는 변수

실무적으로는 키변수를 그대로 사용하고 민감변수 일부를 재현하여 자료의 유용성 보존을 추구할 수 있다. 키변수를 이용하여 데이터를 동질성이 높은 그룹들로 나누고 각 그룹마다 재현자료를 생성하여 원자료와 유사성을 높이는 노력을 하기도 한다.

한편, 다중대체 기법은 베이지안 이론을 근거로 하므로, 이러한 재현자료 생성 방식은 베이지안 방식의 재현자료라고 표현할 수 있다. 새로운 방식으로, 최근에 딥러닝 이론을 기반으로 하는 재현자료 생성이 제안·발전되었으며, 구체적으로는 적대적 생성망(GAN) 등을 이용할 수 있다. 이 외에도 재현자료 생성을 위한 다양한 연구 성과가 발표되고 있다. 베이지안 방식에 의한 재현자료 생성 사례, 유용성 분석 등과 더불어 딥러닝 방식의 재현자료 생성에 관한 이론적인 소개 등은 박민정 등(2020) 및 김현태와 장가영(2023) 등을 참고할 수 있다.

다음으로 차등정보보호(Dwork, 2006)는 한 명의 개인이 실제로 자료의 분석에 포함되더라도 마치 포함되지 않는 경우와 같은 안전한 보안 환경을 목표로 한다. 즉, 특정 개인이 데이터베이스에 있는 경우에 생산된 정보와 없는 경우에 생산된 정보의 차이가 없어야 정보보호가 실현되었다고 보는 것이 차등정보보호의 의미이다(박민정 등,

2018a). 차등정보보호에 대한 확률적 정의에 필요한 가정은 다음과 같다.

- 데이터베이스 D_1과 D_2 각각에 포함된 개체 수의 차이가 한 개라고 하자.
- 주어진 쿼리에 임의의 노출제어 방법을 적용한 것을 확률함수 K라고 표현하자.

이러한 가정 하에서 ϵ-차등정보보호(ϵ-differential privacy)란 다음과 같은 조건을 만족하는 것이다. 집합 S는 모든 가능한 결과 중 임의의 집합을 나타낸다.

$$P[K(D_1)\in S] \leq e^{\epsilon}P[K(D_2)\in S] \quad for\ all\ S\subseteq range(K)$$

만약 노출제어 방법이 잡음을 첨가하는 방법이면 쿼리의 결과는 '잡음이 첨가된 분석 결과', $K(D)$이고, 확률 P는 '잡음이 첨가된 결과의 분포', $P[K(D)\subset S]$를 의미한다. 위에서 정의된 차등정보보호는 하나의 개인 또는 개체가 제외되어도, 공개된 자료, 즉 '잡음이 첨가된 분석 결과'로부터 얻어지는 정보가 유의미하게 변하지 않는다는 것을 의미한다. 간단한 예를 들면, 급여액이 매우 높은 '홍길동'이라는 개인의 정보가 포함된 D_1에서 평균 급여 $K(D_1)$를 집계하여 공개하여도 홍길동의 급여를 유추하기는 어렵게 만드는 것이 차등정보보호이다. 차등정보보호에 관한 상세한 설명은 박민정 등(2018a)을 참고할 수 있다.

이상적으로 재현자료는 노출위험이 없다고 간주하므로, 연구자들은 원자료와 유사하여 유용성이 높은 재현자료를 생성하는 데 주력했다. 그러나 부분 재현자료 등에서 노출위험이 없다는 것을 수학적으로 보장할 수는 없으므로, 차등정보보호 기준을 만족하는 재현자료 생성에 관한 연구도 진행되었다. 재현자료의 유용성 측정에 관하여는 안성빈 등(2023)에 정리되어 있으며, 차등정보보호를 만족하는 재현자료 생성에 관하여는 Wasserman(2012) 및 박민정 등(2019)을 참고할 수 있다.

3.2 용어 비교

UN의 가이드라인은 Output Privacy에 속하는 주요 기법으로 재현자료와 차등정보보호를 소개하면서, 이들이 비식별화 혹은 익명화(anonymization) 등으로 일컬어지는 기존 방식의 한계를 극복하기 위하여 제안되었다고 설명한다. 따라서 기존 방식이 무엇인지 명확히 이해할 필요가 있으나, 분야별로 용어의 사용이 달라 혼란이 있다. 이를 해소하기 위하여, 이 소절에서는 PET와 관련된 용어의 역사를 간단히 소개하고 분야별로 용어를 비교하여 정리한다.

국가통계(official statistics) 분야에서는 민감한 데이터를 보호하기 위한 각종 자료처리 기법을 매스킹(masking) 기법이라고 불러왔다. 각국의 통계기관은 자신이 배포·제공하는 마이크로데이터나 각종 통계에서 민감한 개인·기업 단위 데이터의 식별·재식별 방지하기 위하여 이러한 기법을 활용하면서, 여러 매스킹 기법을 직접 연구하기도 했다. 한편, 매스킹 처리의 효과를 판단하기 위하여 자료처리 전후 노출 위험성과 자료 유용성을 측정하기 위한 방법론에 관한 연구도 진행되었다.

이러한 매스킹 처리, 노출위험 측정 방법론 및 자료 유용성 측도 등, 각종 통계 및 마이크로데이터를 공표할 때 개인정보보호를 위한 기법 연구를 두루 포괄하여 통계적 노출제어(Statistical Disclosure Control)라고 한다(Hundepool 등, 2012). 반면, 국가통계 이외의 분야에서는 개인정보보호를 위한 데이터 처리 방안을 주로 비식별화 기법(de-identification techniques)이라고 불렀다. 비식별화에 관한 국제표준인 ISO/IEC 20889에서는 대다수의 통계적 노출제어 기법들이 비식별화 기법에도 포함된다고 설명한다(ISO/IEC, 2018).

통계적 노출제어 기법과 비식별화 기법의 차이를 설명하면 다음과 같다. 국제표준 20889에서 설명하는 비식별화 기법은 ①암호화 도구, ②가명화(pseudonymization) 도구, ③표본추출, 총계처리 등의 구조적 방법, ④삭제 방법, ⑤반올림이나 범주화와 같은 일반화 방법, ⑥잡음첨가나 자료교환과 같은 임의화 방법, ⑦재현자료 생성 등을 포함한다. 통계적 노출제어 연구 분야는 이 중 ③~⑥까지의 세부 기법들과, 이러한 기법을 적용하는 자료처리 전후 노출 위험성과 자료 유용성을 측정하기 위한 방법론을 다루는 것으로 정리할 수 있다.

위의 비식별화 기법 중, ①암호화 도구와 ②가명화 도구는 다양한 암호화 기법을 지칭한다는 측면에서 유사하지만, 암호화라는 표현은 보통 분석 대상이 되는 민감한 변수에 사용하는 경우가 많고, 가명화라는 표현은 주로 키변수(주민등록번호 등의 직접식별변수 또는 성명, 성별, 지역 등의 간접 식별변수 조합 등)를 암호화하여 대체식별번호를 생성할 때 사용하는 경우가 많다.

이 외에, 공학 분야에서 제안된 KLT 모형이 널리 사용되었다(개인정보보호위원회, 2022). 이는 비식별화 기법으로 자료를 처리했을 때 노출위험이 제어된 정도를 평가하기 위한 모형으로, K-익명성, L-다양성, T-근접성을 말한다. 이들은 통계적 모형을 사용하는 통계적 노출제어 연구 분야의 노출위험 측정과는 접근방식이 완전히 다르다. 대표적인 예를 들자면, K=3인 K-익명성 기준을 만족하는 것은 비식별화된 데이터의 키변수 조합별 데이터 개수가 최소 3 이상일 때 이루어지며, 확률적으로 노출위험을 측정하는 통계적 노출제어와는 그 결을 달리한다.

지금까지 살펴본 기술 분야 명칭과 세부 기법을 정리하면 <표 3.2>와 같다. 서로 다른 분야의 용어 및 분류를 이해하기 위하여 이 표를 참고할 수 있다.

<표 3.2> PET 관련 기술 분류 비교

	국가통계 연구 분야		ISO/IEC 20889		UN 가이드라인	
	매스킹 기법	통계적 노출제어	비식별화 기법	프라이버시 모델	Input Privacy 분야	Output Privacy 분야
sMPC 등의 보안 방식			*		○	
암호화 (HE 등)			○		○	
가명화			○			
구조적 방법 (표본추출, 총계처리 등)	○	○	○			
삭제 방법	○	○	○			
일반화 방법 (반올림, 범주화 등)	○	○	○			
임의화 방법 (잡음첨가, 자료교환 등)	○	○	○			
재현자료 생성			○			○
유용성 측도		○				
노출위험 측도		○				
KLT 모형				○		
차등정보보호				○		○

* sMPC는 동형암호와 밀접한 관련이 있으므로, ISO/IEC 20889에서는 이를 비식별화 기법 중 암호화의 일부로 설명함

※ 국제표준 ISO/IEC 20889이 다루는 주요 기술 중에서 본 보고서의 주제와 관련이 적은 내용(anatomization, linear sensitivity model 등)은 생략함

참고로 국내에는 비식별 조치 혹은 익명화 기법이라는 이름으로 이러한 기술들이 소개되어 왔다(관계부처 합동, 2016), '비식별 조치' 또는 '익명화'라는 표현이 노출위험에 대해서는 안전한 상태라는 느낌을 주지만, 이를 달성하기 위해서는 실무 담당자가 자료 유용성이 크게 손실되는 수준으로 자료를 변형해야 한다. 이러한 자료 변형은 자료 이용자의 반발을 일으킨다. 즉, 노출위험과 자료 유용성 사이의 상충관계(trade-off)에 대한 이해가 부족한 가이드라인 도입 초기에, 비식별이나 익명이라는 표현이 다양한 분야의 관계자에게 혼란을 가중하는 영향이 있었다고 볼 수 있다.

최근 개인정보보호위원회에서 기존 가이드라인의 폐기를 선언하면서 새로 발간한 가이드라인의 제목에서는 '가명정보 처리'라는 표현을 사용하고, 익명화 및 비식별이라는 용어를 가급적 사용하지 않았다(개인정보보호 위원회, 2022). 그러나 '가명정보 처리'라는 표현은 가명화(가명 처리)를 지칭하는 것이 아니며, 내용상으로는 ISO/IEC 20889의 비식별화 기법이나 프라이버시 모델 등을 다루고 있다.

한편, 통계청의 가이드라인은 ISO/IEC 20889의 비식별화(de-identification)라는 개념 및 분류 아래, 법규정상 익명처리와 가명처리의 개념을 정리했다(통계청, 2023). 통계청 가이드라인의 가명정보는 주요 키변수를 가명화한 데이터를 의미한다. 또한 익명처리는 매스킹 기법을 적용하는 것을 의미하며, K-익명성 측도를 기준으로 '익명성 K를 확보한 익명정보' 등의 표현을 사용한다(통계청, 2023).

용어에 관하여 조금 더 언급하자면, 박민정과 김항준(2016)에서는 베이지안 모형을 이용해 데이터의 통계적 특성을 재현한다는 의미로 synthetic data를 재현자료라고 번역할 것을 제안했다. 당시의 기존 '개인정보 비식별 조치 가이드라인'(관계부처 합동, 2016) 등에서는 synthetic data를 합성데이터, 인위자료 등으로 번역했었는데, 데이터 유용성에 민감했던 당시 기류에서 합성데이터라는 번역은 synthetic data에 대한 거부감을 일으켰기 때문이다. 이에 재현자료라는 표현으로 생성한 데이터의 유용성을 강조함으로써 synthetic data에 대한 수용성을 높이고자 했다.

다만 최근에는 노출위험에 대한 민감함이 예전보다 강해져 재현자료라고 표현하면 이전과는 반대로 프라이버시가 침해될 것 같은 인상을 줄 수 있다. 따라서 개인정보의 재현이라는 오해를 불식시키면서도 모집단의 분포를 재현한다는 의미를 강조하여, 통계적 재현자료(statistical synthetic data)라는 표현을 병행하는 방안을 제안한다. 이는 모집단의 통계적 특성을 재현하고 노출위험과 유용성을 통계적으로 검증하여 생성한 데이터라는 의미를 가지며, 언어모형 및 생성형 모형 등을 통해 생성된 인위 데이터를 포함하는 용어인 합성데이터와 구별된다.

해외에서는 과거에는 매스킹 기법을 적용한 데이터를 masked data라고 지칭했으며, 베이지안 모형을 이용해 생성한 데이터를 synthetic data로 불렀다. 최근에는 베이지안, 순차회귀모형, 딥러닝, 차등정보보호 적용 등 모형에 상관없이 원자료를 대신하도록 생성한 데이터를 모두 synthetic data로 부르는 추세이다. 다만, IT업계를 중심으로, synthetic data라는 용어를 원자료와 상관없이 생성형 인공지능으로 생성된 모든 인위 데이터를 모두 포함하는 뜻으로 더 넓게 사용하는 경향이 있으므로 주의해야 한다. 본 논문에서는 이와 같은 광의의 synthetic data를 합성데이터로, 모집단의 특성을 통계적으로 검토하여 높은 수준의 유용성을 가지도록 생성하였으면서도 노출위험을 제어한 데이터를 (통계적) 재현자료라고도 표현하기로 한다.

3.3 이용자 편의성 향상을 위한 통계청의 노력과 성과

국내외 대부분의 데이터센터는 유사한 방식으로 운영된다. 외부에 공개할 수 없는 마이크로데이터를 제공하여 이용자가 센터에서 이를 분석할 수 있도록 하며, 분석한 결과는 직원의 검수를 거쳐 반출을 허용한다. 만약 분석 수요가 많다면, 소수의 직원이

많은 분석 결과를 검토하는 데 오랜 시간이 요구된다. 또한 분석 결과의 유형이 다양하여 담당 직원의 경험과 전문성이 요구되며, 비식별화의 기준도 명확히 정립되지 않아 분석 결과 반출 업무에 어려움이 있다.

다음 <표 3.3> 및 <표 3.4>는 반출 결정 및 비식별화 조치에 관한 간단한 예제인데, 이를 통해 노출위험, 유용성 및 이용자 편의성 문제를 살펴보자. 분석 대상 마이크로데이터는 산업 분류, 대표자 성별, 분석 대상인 임의의 연속형 변수로 구성되어 있다. 다만 실제 반출을 요청하는 분석 결과는 이 예제와 같이 단순하지 않고, 분량이 방대하고 구조가 복잡한 경우가 많다.

<표 3.3> 분석 결과물에서 개인정보 노출 예제

Industry	Male CEO			Female CEO			Total		
	Counts	Avg.	Var.	Counts	Avg.	Var.	Counts	Avg.	Var.
A	238	122	28,993	76	98	7,440	314	220	36,433
B	2	18	35	347	84	29,186	349	102	29,222
C	64	119	7,604	32	97	3,108	96	216	10,713
Total	304	121	36,633	455	87	39,734	759	101	76,367

만약 데이터센터의 반출 기준이 K-익명성을 만족하는 것이고, 기준이 되는 K값이 3이라면, 위 <표 3.3>에서 붉게 표시된 빈도가 2인 분석 결과를 반출할 수 없다. 특히 평균(Avg.) 및 분산(Var.)이라는 두 개의 정보로 간단한 계산을 통해 마이크로데이터에 있는 두 개체의 데이터 참값을 쉽게 알아낼 수도 있다. 이는 마이크로데이터 원자료를 외부에 직접 제공하는 것과 다를 바 없는 결과이다. 따라서 이 셀들은 외부에 제공하지 못하고 삭제 처리하며, 이를 1차 삭제라고 한다. 그런데 이 분석 결과에서는 그룹별 통계와 전체 통계를 함께 제공하고 있으므로 전체값(Total)에서 다른 그룹값(Female)을 빼면, 삭제한 그룹(Male)의 값을 알 수 있다. 따라서 다른 셀들도 함께 삭제해야 하며 이를 2차 삭제라고 부른다. 이제 데이터센터는 다음 <표 3.4>와 같은 분석 결과의 반출을 허용할 수 있다. 1차 삭제 결과와 2차 삭제 결과가 각각 붉은색과 주황색으로 표기되어 있다.

<표 3.4> 분석 결과물에서 개인정보 노출 예제

Industry	Male CEO			Female CEO			Total		
	Counts	Avg.	Var.	Counts	Avg.	Var.	Counts	Avg.	Var.
A	238	122	28,993	76	98	7,440	314	220	36,433
B	①	①	①	②	②	②	349	102	29,222
C	②	②	②	②	②	②	96	216	10,713
Total	304	121	36,633	455	87	39,734	759	101	76,367

따라서, 이용자는 반출물을 수령하기까지 2~3주, 혹은 그 이상의 기간을 기다려야 하고, 위의 예시와 같이 유용성이 매우 낮은 분석 결과를 받아야 할 수도 있다. 정리하면, 이용자는 데이터센터를 방문하기 이전에 공개된 메타데이터 정보만을 가지고 데이터 분석 계획을 수립하고, 필요한 경우 숙박비와 교통비를 지출하며 데이터센터가 위치한 지역을 방문하여 데이터센터 정규 업무시간에만 데이터를 분석하고, 분석 결과의 반출 심사를 오랫동안 기다린 후, 유용성이 낮아 만족하기 힘든 분석 결과를 수령해야 할 수도 있다. 즉, 이용자 편의성이 현저히 낮아 데이터 활용에 대한 동기가 약화할 수 있다.

통계청은 통계데이터센터를 개소한 직후, 6개월 간의 이용자의 반출물 자료를 기반으로 반출물 평가를 위한 가이드라인을 작성했다(박민정 등, 2017b). 또한 지역 변수나 산업 분류 등, 위계 구조를 가진 변수가 포함되어 있는 하나의 마이크로데이터에서 대량의 빈도표를 생성할 때, K-익명성을 유지하면서 정보손실을 최소화하는 알고리즘인 iLBA(information Loss Bounded Aggregation)를 고안하고(박민정 등, 2018b; Park et al., 2024) 특허를 출원했다. 이를 기반으로 데이터센터를 방문한 이용자가 기업통계등록부를 분석하여 그룹별 기술통계를 산출하고 반출을 신청할 때, 자동으로 반출물을 비식별화하는 프로그램인 DISK를 2023년에 시범 개발했다. 이 프로그램은 엑셀을 인터페이스로 활용하여 이용자의 신규 프로그램 습득에 대한 부담을 완화했다. 또한 2024년에는 여러 방식으로 기업통계등록부의 재현자료를 작성하여 베타서비스를 실시했다. 이러한 일련의 프로젝트는 노출위험 제어와 데이터 유용성 확보뿐만 아니라 이용자의 편의성 향상을 비중 있게 다루고 있다.

4. 다자간 피처 스토어 구축 방안

본 논문의 제2절에서는 데이터 연계·활용을 막는 주요 원인을 분석하고, 제3절에서는 이러한 문제를 해결하는 데 사용할 수 있는 최신 기술을 소개했다. 더불어 법과 제도적 측면에서 데이터센터(망분리 보안구역)와 가명결합 전문기관 제도라는 틀 내에서 데이터 연계·활용 문제를 해결해야 함을 확인했다. 이제 현행 제도하에서 데이터의 연계와 활용을 활성화하는 방안으로서, 개인정보보호를 위한 신기술을 기반으로 하는 '다자간 피처 스토어(multy-party feature store)' 구축을 제안하고자 한다. 이때 연계 대상이 되는 데이터란 서로 다른 기관이 보유하고 있으면서 민감성으로 인해 전면 개방이 어려운 미공개 마이크로데이터를 말한다.

본 논문이 제안하는 다자간 피처 스토어 모델을 설명하기에 앞서, 피처 스토어의 역사와 개념을 간단히 소개하면 다음과 같다.

피처 스토어는 머신러닝 모형을 개발하고 운영할 때, 데이터 피처(data feature; 속성, 변수)를 중앙에서 저장·관리·제공하는 플랫폼으로 2017년 우버(Uber) 연구진이 처음 제시한 용어이다(Li 등, 2017). 피처는 머신러닝 모델의 입력값으로 사용되는 데이터를 의미한다. 예를 들면, 통계청 기업통계등록부(SBR)의 여러 변수(컬럼, 속성)의 값들 또는 그것들을 가공한 값들이 SBR로부터 생성한 피처이다. 이때 변수(속성)는 사업체명,

산업 분류 코드, 성별 종사자 수 등을 말하며, 이는 원자료에 있는 값들이다. 가공한 값이란 원자료의 변수, 특성값을 가공하여 생성한 새로운 속성값으로 여성 종사자 비율을 예로 들 수 있다. 피처 스토어는 이러한 피처를 효율적으로 관리하여 데이터 엔지니어링 작업을 단순화하고, 데이터의 일관성과 재사용성을 높이는 기능을 가진다.

일반적으로 피처 스토어는 한 기업 내에서 다양한 팀이 다양한 모형을 다룰 때, 동일한 피처, 특히 가공한 값으로 만들어진 피처를 공유하고 재사용할 수 있도록 지원하는 역할을 한다. 이로써 피처 스토어는 머신러닝 및 인공지능 서비스의 개발과 운영을 손쉽게 하는 소프트웨어적 플랫폼이 된다. 한편 피처 스토어의 스토어(store)는 주로 중앙화된 데이터 저장소(storage)라는 의미로 사용되지만, 다양한 데이터 엔지니어링팀이나 인공지능 서비스 개발팀이 피처를 공유하는 교환소 또는 상점의 의미로 해석할 수도 있다.

본 논문에서 제안하는 다자간 피처 스토어는 위에서 설명한 데이터 저장소 도구로서의 피처 스토어를 다기관 데이터 공유 플랫폼으로 확장하는 개념이다. 다자간 피처 스토어는 서로 다른 기관에서 제공하는 데이터를 안전하게 결합하여, 데이터베이스의 마이크로데이터 또는 세분화된 데이터(granular data) 형태로 데이터를 보안구역에 저장하여 이를 머신러닝 및 인공지능 모델 개발을 위해 활용할 수 있도록 하는 데이터 저장소와 제반 서비스를 지향한다.

제안하는 다자간 피처 스토어는 기존의 데이터 공유 플랫폼 또는 데이터 허브와는 두 가지 큰 차이점이 있다. 첫째, 전통적인 데이터 허브에서는 대체로 집계된 데이터의 형태로 데이터가 저장되고 공유되지만, 다자간 피처 스토어에서는 집계처리 전의 마이크로데이터 형태로 데이터가 공유된다. 둘째, 기존 데이터 공유 플랫폼의 경우, 서로 다른 기관의 데이터가 통합되지 않고 분리된 상태로 존재하는 단점이 있으므로 데이터 연계 및 협업 활용에 제약이 있을 수 있다. 머신러닝이나 인공지능 서비스 개발이 어려움은 물론이고, 단순한 통계처리도 쉽지 않은 플랫폼이라 할 수 있다. 반면, 다자간 피처 스토어에서는 마이크로데이터를 손쉽게 연계하고 안전하게 제공하여 데이터 분석이 더욱 편리하게 된다고 할 수 있다.

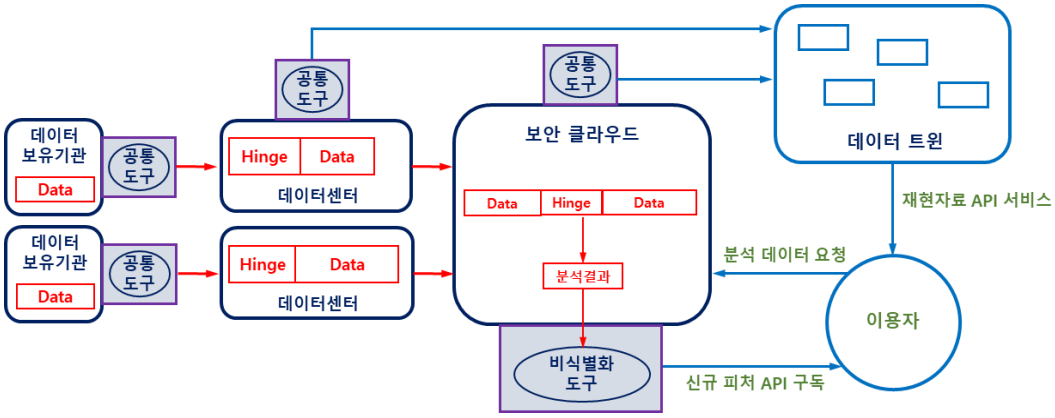
본 논문에서는 다기관 데이터 공유 환경에서 다자간 피처 스토어의 성공적인 구현을 위한 필수적인 요소로서 데이터 연계를 위한 **기준 데이터**(hinge data)와 안전성 담보와 용이한 서비스 구축을 위한 **데이터 트윈**(data twin)을 제안한다.

먼저 서로 다른 원천에서 생성되는 데이터를 효율적으로 연계하기 위해서는 기준 데이터(hinge data)가 필요하다. 기준 데이터는 서로 다른 데이터 세트를 연계하는 역할을 하며, 마치 병풍에서 서로 다른 작품이 걸려있는 여러 장의 판을 통일성 있게 연결하는 경첩(hinge)과 같은 기능을 수행한다. 기준 데이터는 데이터 통합의 일관성을 보장할 뿐만 아니라, 기준 데이터를 구성하는 여러 항목을 활용하여 연계의 효율성을 증진하므로 다양한 기관 간 협력을 위한 데이터 연계의 핵심 요소로 작용한다. 최근 행안부에서 국가기준데이터 관리지침을 제정(‘22.3)하고, 몇몇 주제별로 국가기준데이터 목록을 지정하여 각 기관이 해당 속성을 관리하도록 했다. 그러나 개별 기관의 데이터가 공통 변수를 가져서 일부 사례에서 데이터가 연결될 수도 있는 수준을 넘어서는, 효율적인 기준 데이터를 운영할 필요가 있다. 통계청의 통계등록부와 같이 중앙에서

관리하는 모집단에 대한 마이크로데이터를 기준 데이터로 활용하는 것이 바람직하다.

다음으로 데이터 트윈은 실제 데이터를 기반으로 만들어진 재현자료로 구성되며, 재현자료는 원본 데이터의 통계적 특성과 구조를 유지하면서도 보안 위협을 최소화한다. 다자간 피처 스토어에 저장된 원본 데이터는 강력한 보안 및 프라이버시 보호 기술로 인해 데이터 접근과 분석에 시간과 노력이 필요할 수 있다. 데이터 트윈은 이러한 한계를 극복하여 이용자가 자신의 공간에서 자신의 분석도구를 사용해 원본 데이터 대신 재현 자료를 사전에 분석해 볼 수 있도록 하는 기능을 가진다. 이를 통해 연구자와 개발자의 데이터 접근성을 높이고, 이들이 효율적으로 데이터를 분석하게 지원할 수 있다.

다음 <그림 4.1>은 다자간 피처 스토어와 데이터 트윈에 대한 개념도이다. 다자간 피처 스토어는 데이터 트윈을 제외한 구성요소들을 보안구역 안에 구축하고, 민감한 원본 자료를 보유한다. 데이터 트윈은 안전하면서도 원자료를 대신하여 사전 분석용으로 활용하기에 적절한 수준의 유용성을 가지는 재현자료로 구성된다. 이용자는 재현자료를 사전 분석하여 필요한 원본 데이터 속성을 구체적으로 결정하고, 이를 다자간 피처 스토어에 요청할 수 있다. 이용자는 다자간 피처 스토어에서 요청한 원본 자료를 직접 분석할 수 있으며, 원하는 신규 피처를 생성하여 구독을 요청할 수 있다. 이용자가 구독하는 신규 피처는 프라이버시 측면에서 안전성이 확보된 피처들이다. 플랫폼에서 통계를 구독하는 개념에 관한 논의는 변준석 등(2020)을 참고할 수 있다.



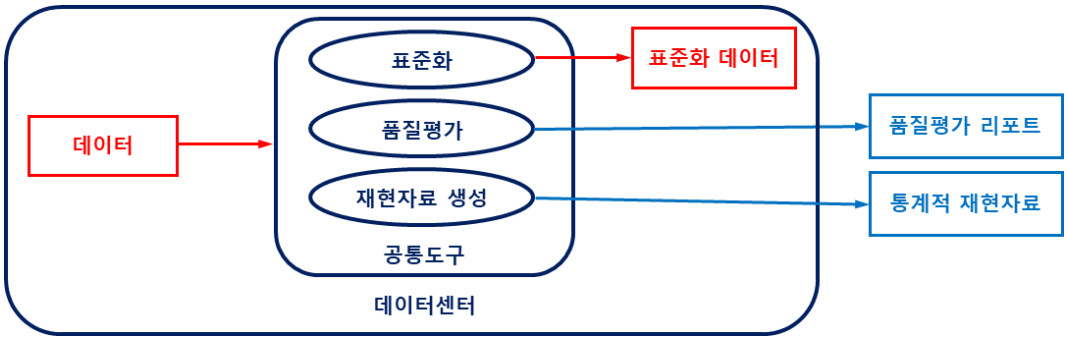
<그림 4.1> '다자간 피처 스토어' 개념도

이제 '다자간 피처 스토어'와 '데이터 트윈'을 구성하는 기술적인 요소들을 하나씩 살펴보도록 한다. 먼저 모든 데이터 보유기관이 사용해야 하는 '공통도구', 그리고 데이터센터가 사용하는 '분석결과 비식별화 도구' 및 '기준 데이터(hinge data)'를 설명한다. 이 도구들은 향후 개발해야 하는 것들로 본 논문에서는 기본 개념을 제시한다. 다음으로 이러한 기술적인 요소를 활용하여 구축하는 '데이터 트윈' 및 '다자간 피처 스토어'를 제안하고, 마지막으로 '다자간 피처 스토어에서 보상 체계', '다자간 피처 스토어의 한계점' 등을 논의한다.

4.1 공통도구 개발 및 활용

피처 스토어에서는 데이터센터를 포함한 모든 데이터 보유기관이 활용할 수 있는 '공통도구'를 개발하여 사용한다. 이 도구의 기능은 데이터 표준화, 품질평가, 재현자료 생성이며, 이를 위해 각 기능을 수행하는 독립적인 프로그램을 모듈 형식으로 가진다. 따라서, 각 기능별로 업데이트된 프로그램을 플러그인 할 수 있다. '공통도구'의 각 기능을 설명하면 다음과 같다.

첫 번째 기능은 표준화이다. 2.4절에서 언급한 대로, 표준화는 실제로 이행이 이루어지도록 관리하기 어려운 작업이다. 각 데이터 담당자가 매뉴얼 책자를 참고하여 주어진 표준을 직접 적용하는 방식으로는 표준화 실행의 품질을 담보하기 어려운 경우가 많기 때문이다. 반면 통계청의 데이터융복합관리시스템 등은 자동화된 표준화 솔루션을 운영하여 서로 다른 기관에서 입수하는 데이터에 대한 표준화를 수행하고 있다. 공통도구는 이와 유사한 방식의 데이터 표준화 프로그램을 첫 번째 구성요소로 가진다.



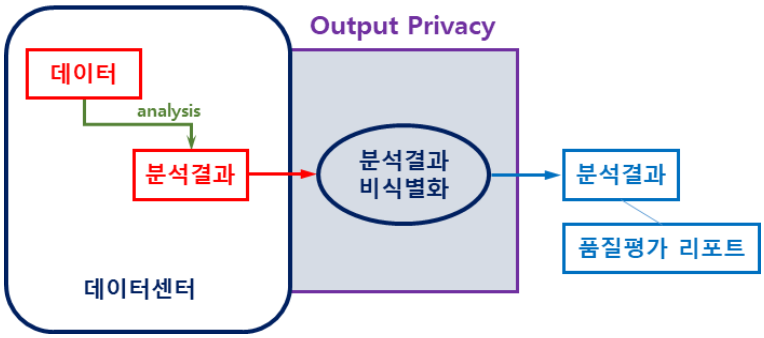
<그림 4.2> 표준화, 품질평가, 재현자료 생성을 수행하는 공통도구

두 번째 품질평가 기능을 수행하는 메뉴는 원데이터에 대한 품질진단 결과, 간단한 통계분석(EDA) 결과, 노출위험 측정 결과 및 각종 메타데이터 등을 리포트 형식으로 출력한다. 마지막으로 재현자료 생성 프로그램은 유용성 측면에서 다양한 수준의 통계적 재현자료를 생성할 수 있고, 각 재현자료의 노출위험과 유용성을 제시한다. 표준화, 품질평가 및 재현자료 생성 기능은 통계청의 최근 연구·개발 사업 성과를 직접 활용하거나 이를 고도화하여 새로 개발할 수 있다.

세 가지 기능을 수행하는 공통도구는 망분리 상태에서도 사용할 수 있도록 망간 전송할 수 있는 형식의 프로그램으로 제작해야 하며, 이용자가 이 도구를 다루기 쉽도록 편리하게 제작해야 한다. 또한 데이터센터는 공통도구로 생성한 표준화 데이터만을 수용하도록 운영하여 데이터 표준화가 실행되도록 한다. 즉, 각 데이터 보유기관은 공통도구를 이용하여 자신의 데이터를 처리해야 데이터센터에 자신의 데이터를 제출할 수 있다. 연계된 데이터에 대한 재현자료 생성 등, 필요한 경우 데이터센터도 이 공통도구를 사용한다. 공통도구를 활용하여 다자간 피처 스토어 전체적으로 데이터와 리포트의 통일성을 유지하는 것은 다자간 피처 스토어의 효율성 확보 측면에서 중요하다.

4.2 분석결과 비식별화 도구 개발 및 기준 데이터 (Hinge Data) 활용

이 소절에서는 데이터센터가 수행하는 두 가지 주요 역할을 설명한다. 먼저 각 데이터센터는 이용자에게 편리한 분석결과 반출 체계를 운영해야 한다. 이를 위하여 데이터센터의 반출물 비식별화를 수행하는 기능을 가진 도구를 개발한다. 분석결과에 포함된 각 통계량이 충분한 자유도를 가진다면 비식별화 처리를 하지 않고 반출할 수 있고, 악의성이 의심되며 반복적으로 요청되는 통계량에 대해서는 차등정보보호를 적용하는 전략을 취할 수도 있다. 유용성이 중요한 상황이라면 통계청이 특허출원한 '빈도값 비밀보호 알고리즘 iLBA' 및 통계청이 시범 개발한 '그룹별 기술통계 비식별화 프로그램 DISK'를 활용할 수 있다. 위계 변수가 없는 간단한 경우에 대해서는 해외에서 제안된 또 다른 빈도값 비밀보호 알고리즘인 CKM(cell key method)을 활용할 수도 있다(www.ons.gov.uk). 이 '분석결과 비식별화 도구' 역시 최신 기술을 반영하여 업데이트하면, 플러그인 방식으로 손쉽게 교체할 수 있도록 제작해야 한다. 더불어, 분석결과에 대한 유용성 및 노출위험 정도를 설명하는 리포트를 출력하는 기능을 가져야 한다. 다음 <그림 4.3>은 '데이터 분석결과 비식별화 도구'의 개념도를 나타낸다.



<그림 4.3> 데이터 분석결과 비식별화 도구

한편, '분석결과 비식별화 도구'를 사용할 수 없는 새롭고 특수한 유형의 분석결과에 대해서는 기존의 방식대로 심의절차를 운영할 수 있다. 반출 횟수가 누적되어 새로운 반출유형으로 정형화할 수 있으면, 이를 반영하여 '분석결과 비식별화 도구'를 업데이트하고 새로운 유형의 분석결과 반출을 자동화할 수 있다. 참고로 현재 통계청에서 시범 개발한 DISK는 그룹별 기술통계 유형을 자동 반출하는 기능을 가진다.

다음으로 데이터센터는 기준 데이터(hinge data)를 운영하고, 이용 빈도가 높은 원자료 마이크로데이터를 통계등록부 등의 기준 데이터에 연계하여 보관하는 역할을 수행할 것을 제안한다. 통계등록부란 통계청장이 개인, 법인 또는 단체 등에 관한 분야별 모집단의 기본정보를 수록하여 관리하는 '주제별 모집단 자료'이며, 각 기관에서 공표하는 승인통계의 모집단으로도 활용되고 있다. 현재는 인구·가구·주택·기업 총 4종의 기본통계등록부와, 취업활동·아동가구·청년 총 3종의 정책맞춤형 통계등록부가 있다. 데이터센터에서 이용 요청이 많은 데이터와 해당 데이터의 모집단과 관련 있는 통계

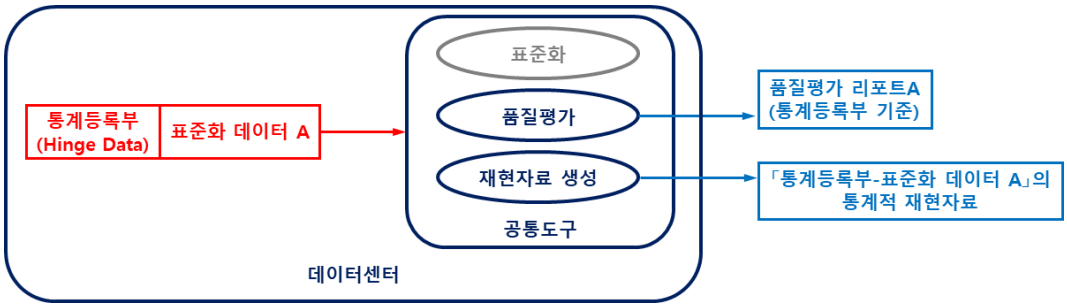
등록부를 사전에 연계해 놓으면, 서로 다른 데이터를 통계등록부를 기준으로 연계할 수 있어 연계 효율성이 강화된다. 통계등록부에 연계된 데이터들을 연계할 수 있는 것은 물론이고, 통계등록부에 포함된 특정 변수를 기준으로 작성된 데이터도 빠르게 연계할 수 있다. 통계등록부에 포함된 다양한 변수를 연계키로도, 집계키로도 손쉽게 이용할 수 있기 때문이다.

예를 들어 어느 데이터 보유기관이 지역변수를 가지지 않는 데이터 A를 '공통도구'에 입력하여 표준화 데이터 A를 생성하고 데이터센터에 제출했다고 하자. 데이터센터는 표준화 데이터 A와 통계등록부를 가명결합하여 보관한다. 통계등록부에는 ID(가명), 지역, 위치, 성별, 연령 등의 변수가 존재한다. 한편 다른 기관이 격자 단위로 작성한 새로운 표준화 데이터 D를 표준화 데이터 A에 연계하고자 한다고 하자. 이때 표준화 데이터 D는 통계등록부에 연계되는 상세한 수준이 아니며 격자 단위로 집계한 자료임을 주의하자. 이제 지역 정보가 없었던 표준화 데이터 A를 통계등록부의 위치 변수를 이용해 격자 수준에서 집계하여 데이터 D와 연계할 수 있다. 즉, 통계등록부에 포함된 모든 변수를 집계키로도 활용할 수 있으므로, 동일한 수준의 마이크로데이터뿐만 아니라 다양한 수준의 집계 데이터에 대해서도 보다 빠르고 광범위하게, 요청받은 데이터들을 효율적으로 연계할 수 있다.

<그림 4.4>는 데이터센터에 제출된 서로 다른 기관의 표준화된 데이터 A 및 D가 통계등록부에 연계되어 있음을 나타낸다. 이때 연계 요청이 있기 이전에는 각 데이터를 기존 방식대로 서로 다른 컴퓨터에 보관할 수 있다. 데이터 보유기관에게 기존 방식과 다른 데이터 이관을 요구하지 않음을 주지할 필요가 있다.

<그림 4.4> 통계등록부 등의 기준 데이터(hinge data)의 활용

한편 데이터센터도 통계등록부와 연계된 표준화 데이터를 공통도구에 입력하여 통계적 재현자료를 생성할 수 있다. 더불어 통계등록부를 기준으로 데이터를 설명하는 리포트를 작성하여 배포할 수 있다. 통계등록부를 기준으로 데이터를 설명하면, 서로 다른 데이터를 비교하기에 편리한 장점이 있다. <그림 4.5>는 연계된 데이터에 대한 재현자료 생성을 나타낸다.



<그림 4.5> 통계등록부와 연계된 표준화 데이터의 재현자료 생성

지금까지 공통도구 개발·활용, 분석결과 비식별화 도구의 개발·활용, 그리고 기준 데이터의 활용을 설명했다. 이들은 다음 두 소절에서 제안하는 데이터 트윈과 다자간 피처 스토어에 필수적인 도구이다.

4.3 데이터 트윈 (Data Twin) 구축

데이터를 연계하여 활용할 때, 우리는 개인정보보호 수준과 데이터의 유용성을 심각하게 고려한다. 그러나 이 두 목적을 동시에 달성하는 것이 무척 어려워서, 이용자의 편의성 확보라는 문제는 간과되어 온 경향이 있다. 반면, 민간 부문의 많은 IT 서비스는 유사한 기능을 가지더라도 이용자가 얼마나 편리하게 이용할 수 있는지에 따라 서비스 성패가 달라지는 경우가 많다.

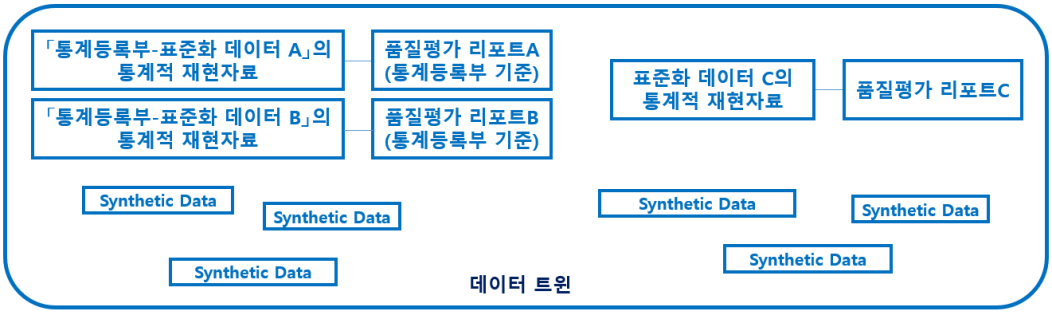
현재의 데이터센터는 망분리를 통한 개인정보보호와 원자료 제공을 통한 데이터 유용성 확보를 달성하지만, 시간적·공간적 제약이 커서 이용하는 것이 편리하지는 않다. 실제로 데이터 활용이 활발하지 않은 이유로 쓸모 있는 데이터의 부족과 더불어 데이터 이용의 불편함이 지적되고 있다.[6] 즉, 국내에서 유용성이 가장 높은 원자료를 제공하는 거의 유일한 수단이 데이터센터임에도 불구하고, 데이터센터의 이용 실적이 저조한 이유는 업무시간 중에 물리적 장소를 방문해서 주어진 환경에서만 작업해야 하는 불편함 때문이라고 할 수 있다.

앞 소절에서 제안한 '분석결과 비식별화 도구'는 반출 절차에 있어 이용자 편의성을 향상하지만, 데이터 접근성이 낮다는 데이터센터의 한계를 해결하지는 못한다. 이 소절에서는 데이터 트윈 구축을 통하여 데이터 접근성을 포함한 이용자 편의성을 확보할 것을 제안한다. 데이터 트윈은 재현자료로 구성되어 미공개 마이크로데이터에 대한 이용자의 데이터 사전 탐색 및 사전 분석을 가능하게 한다.

앞에서 설명한 공통도구는 재현자료를 생성하도록 고안되어 있다. 재현자료를 생성하는 통계적 모형은 다양하며, 모형에 따라 생성 난이도, 노출위험, 유용성이 다르다. 노출위험을 낮추기 위해 유용성을 쉽게 포기하면, 생성 난이도가 낮아 비용이 적게 발생하는 재현자료를 생성할 수 있다. 노출위험을 낮추면서도 유용성을 높게 유지하려면, 재현자료의 생성이 어려우므로 생성 비용이 많이 발생할 수 있다. 본 논문에서는

6) 디지털플랫폼정부위원회 실현계획(dpg.go.kr)

원자료와 비교하여 자료의 유용성이 적절한 수준으로 확보되도록 생성한 통계적 재현자료를 API 형식으로 외부에 공개하는 데이터 트윈을 구축할 것을 제안한다. <그림 4.4>는 데이터 트윈 개념도를 나타낸다. 데이터 트윈은 '공통도구'를 사용하여 표준화 데이터 C와 같이 개별 기관이 생성한 재현자료와 리포트, '통계등록부-표준화 데이터' A 및 B와 같이 데이터센터가 생성한 통계등록부에 연계된 재현자료와 리포트 등으로 구성된다.



<그림 4.6> 통계적 재현자료로 구성된 데이터 트윈

데이터 트윈은 디지털 트윈(digital twin)의 개념을 빌려온 것이다. 시간적·공간적 제약이 있는 데이터센터를 방문하여 실제 자료를 직접 분석하기 이전에, 이용자가 집이나 사무실에서 자신에게 친숙한 분석 자원을 활용하여 재현자료를 편리하게 분석할 수 있도록 하는 것이 목적이다. 이는 데이터 연계·활용을 활성화하기 위한 핵심적 전략이라 할 수 있다.

또한 재현자료의 품질 수준을 다양하게 생성할 수 있으므로, 데이터 도메인의 상황에 따라 재현자료 제공 API를 완전히 공개하는 방식, 등록한 이용자에게만 공개를 허가하는 방식 등에 대해 다양하게 고려할 수 있다. 재현자료의 품질에 따라 서비스에 대한 가격 책정을 다양하게 할 수도 있으며, 이에 관해서는 4.5절에서 더 논의하도록 한다.

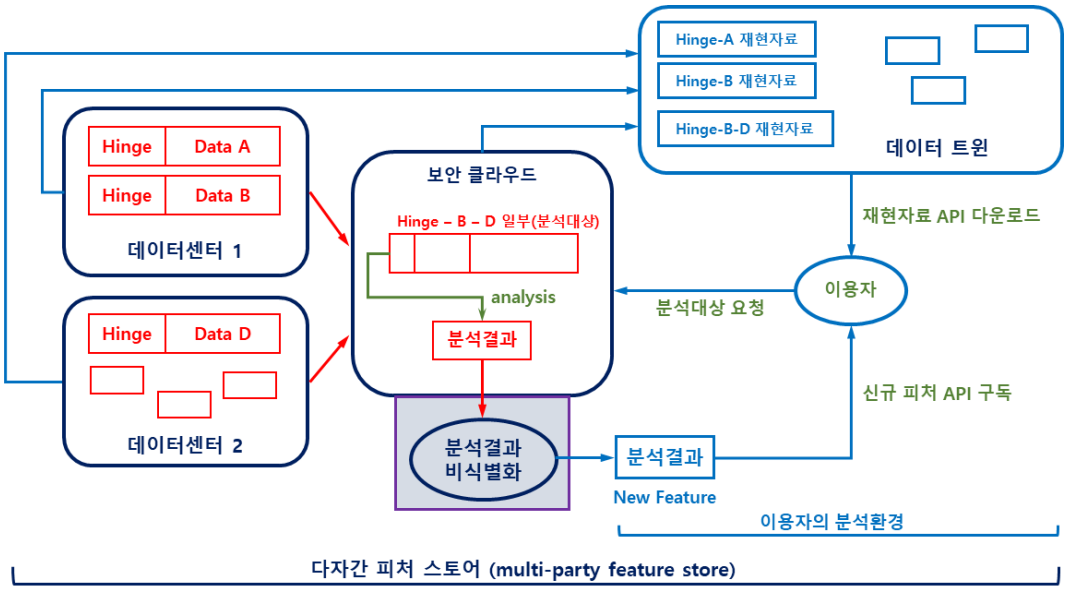
4.4 다자간 피처 스토어 (Multi-Party Feature Store)

지금까지 데이터 트윈은 공통도구를 활용해 생성한 재현자료로 구성되며, 데이터센터는 각 기관이 공통도구를 활용해 생성한 표준화 데이터를 제출받아 이를 기준 데이터에 연계하여 보유함을 설명했다. 이는 현재 데이터센터 운영 방식에 공통도구와 기준 데이터의 이용을 추가한 것으로, 데이터 자체는 기존의 방식대로 보관되며 다른 장소로 이관되지 않는다. 다자간 피처 스토어를 구성하는 마지막 요소는 이용자가 선택한 속성만 포함한 데이터를 업로드하여 데이터 연계와 분석이 이루어지는 보안이 유지되는 클라우드이다. 이 클라우드는 데이터센터를 방문하여 접속할 수 있다. 초창기에는 클라우드 접근을 위해 데이터센터 방문을 요구하는 데서 시작할 수 있으며,

향후 완전한 온라인 접근방식으로 확대해 갈 수 있다.

클라우드의 기능을 설명하면 다음과 같다. 이용자는 사전에 데이터 트윈을 통해 재현자료를 충실하게 탐색하여 데이터 분석 코드를 작성할 수 있다. 또한 원본 데이터 전체를 요구하지 않고, 분석 목적에 맞는 속성을 구체화할 수 있다. 이용자는 다자간 피처 스토어 운영자에게 구체화된 범위의 분석 대상 데이터를 요청한다. 그러면 각 데이터센터는 요청받은 범위의 데이터를 클라우드에 업로드하고, 이 데이터들은 기준 데이터를 이용하여 연계된다. 이제 데이터센터를 방문한 이용자가 작성해 둔 코드를 활용해 클라우드에서 연계된 데이터를 분석하고, 분석결과 비식별화 도구를 통해 빠르게 안전성을 확보한 최종 피처를 반출한다.

다음 <그림 4.7>은 다자간 피처 스토어의 각 구성 요소를 표현한다. 참고로 <그림 4.7>에서 최종 피처를 생산하는데 사용한 연계된 자료 Hinge-B-D에 대해서도, 공통 도구를 이용해 생산한 재현자료를 데이터 트윈에 올려 다수의 이용자에게 새로운 거래 대상으로서 홍보할 수 있다.



<그림 4.7> 다자간 피처 스토어 및 이용 방식

지금까지 설명한 피처 스토어의 데이터 연계·분석 과정에서 보안구역 내의 클라우드조차 신뢰하기 어려운 매우 민감한 속성을 포함한 데이터가 있을 수 있다. 그런 경우 데이터 제공기관의 요구에 따라 해당 속성을 동형암호 처리하여 연산할 수 있다. 즉, <그림 4.7>에서는 데이터 B 또는 D에서 일부 컬럼을 동형암호 처리할 수 있는 것이다. 한편, 이용자가 자신이 작성한 분석 코드에 대한 기밀성(confidentiality) 보장을 요구할 수 있는데 그러한 경우에는 신뢰 실행 환경(TEE)을 도입할 수 있다. 참고로 <그림 4.7>에서 붉은 색으로 표시된 사항은 모두 물리적 보안 환경에서 수행 혹은 Input Privacy에 해당한다.

4.5 보상 체계에 대한 제언

지금까지 보상에 대한 논의는 데이터의 가치 측정의 관점에서만 진행된 경향이 있다. 그러나 데이터 연계·활용은 데이터 분석 서비스, 새로운 피처의 개발, 피처 구독 등으로 이어지므로, 보상 체계를 논의할 때 다른 일반적인 서비스 재화의 가격 결정처럼 다양한 요소를 고려할 필요가 있다. 즉, 데이터 제공에만 기준을 두지 말고 서비스 재화라는 측면도 고려하여, 데이터를 이용해 생성되는 최종 피처의 가치, 피처를 생성(데이터 엔지니어링)하기 위한 인프라에 대한 보상, 피처를 고안하는 과정의 데이터 과학자가 발휘하는 데이터 분석의 창의성, 피처 이용의 편리함 등의 요소를 종합적으로 고려할 필요가 있다. 피처 이용의 편리함이라는 요소에는 다자간 피처 스토어를 이용할 때 발생할 수 있는 불편함을 해소하기 위한, 예를 들면 보안 구역 방문 대행과 같은 여러 서비스가 포함될 수 있다.

더불어 데이터 접근성을 높이는데 중요한 역할을 하는 데이터 트윈에서 통계적 재현자료의 수준을 보상기준으로 고려할 수 있다. 서비스 재화라는 측면에서 쉽게 만든 재현자료와 원자료와 유사성이 높은 통계적 재현자료는 서비스 수준이 다른 것으로 판단할 수 있기 때문이다.

종합하면, 데이터 자체에 대한 보상이라는 기존의 관점보다는, 다자간 피처 스토어의 각 기능이라는 서비스에 대한 보상이라는 차원으로 접근의 방향을 바꾸는 것을 제안하는 바이다. 이를 통해 상당히 오랜 기간 적절한 해법을 찾기 어려웠던 데이터에 대한 보상체계 문제를 새롭게 풀어가는 길이 열릴 수 있다.

다자간 피처 스토어에서 보상의 대상이 되는 주요 공헌자와 역할은 다음과 같다. 먼저 스토어 인프라 전체를 운영하는 중앙기관이다. 이 기관이 하는 역할은 공통도구 개발·운영·개선, 시점별 업데이트를 포함한 기준 데이터 운영, 보안 클라우드 운영, 분석결과 비식별화 도구 개발·운영·개선, 신규 피처 API 구독 제공, 데이트 트윈 운영, 기타 데이터 엔지니어링 역할 등이다.

두 번째는 데이터 과학자이다. 다자간 피처 스토어라는 플랫폼이 효율적으로 활발하게 운영된다면 데이터 생태계로 기능할 수 있다. 이 생태계에서 데이터 과학자는 플랫폼의 주요 이용자로서 전문성을 가지고 새로운 재화인 신규 피처를 발굴하는 역할을 수행한다. 구독 요청이 많은 신규 피처를 고안하여 판매하거나, 과학이나 정책 측면에서 어떤 현상을 설명하는 새로운 피처를 발견할 수도 있다. 이러한 행위에 대한 보상이 충분해야 창의적인 데이터 활용이 확대될 수 있다.

세 번째는 데이터 제공기관이다. 데이터 제공기관에게도 적절한 보상이 주어져야 표준에 충실한 표준화 이행, 데이터 품질관리 등을 통해 좀 더 사용하기 편리한 형태로 다양한 마이크로데이터를 제공할 것이기 때문이다. 다만 이전처럼 데이터 용량을 보상의 기준으로 삼아서는 안 된다. 신규 피처 생산에 활용된 데이터의 컬럼을 기준으로, 그리고 피처 구독 수량에 따라 적절한 보상 방안이 결정되어야 할 것이다. 또한 기관이 보유한 데이터를 홍보하기 위하여 데이터 트윈에 배포하는 재현자료의 수준에 따른 보상도 주어질 필요가 있다.

마지막으로 데이터 오너이다. 불특정 다수의 데이터 오너들에게 적절한 보상 방안이

필요하다. 이를 위해서는 발생하는 수익의 일정 비율을 공적 기금 등으로 활용하는 방안 등을 제안하고자 한다. 이는 데이터 거래가 바람직하지 않다는 윤리적 관점을 반영하여, 데이터 활용으로 발생하는 이익이 다시 공공의 이익으로 환원될 수 있도록 하는 조치이기도 하다. 다만, 피처 서비스에 수반되는 노동과 투자에 대한 보상은 공공의 이익과 분리하여 논의할 대상임을 주의해야 한다.

플랫폼을 통한 데이터 연계·활용에서 발생하는 경제적 이익의 규모가 어느 정도일지는 아직 알기 어렵다. 다만 지금까지는 데이터를 제공하는 자체에 대한 보상을 먼저 수행하는 것이 전형적인 사고방식이었다면, 다자간 피처 스토어라는 새로운 플랫폼에서는 최종 서비스 재화가 활용되는 결과에 따라 보상 규모가 결정되도록 한다는 것이 기존 관점과 큰 차이이다. 다자간 피처 스토어에서는 두 지점에서 최종 서비스 재화가 생성된다. 먼저 최종 피처를 구독하는 지점에서 피처라는 서비스 재화가 생성되며, 구독 규모에 따라 이익의 규모가 결정된다. 다음으로 데이터 트윈에서 통계적 재현자료를 다운로드하는 지점에서 재현자료라는 서비스 재화가 생성되고, 재현자료의 수준에 따라 이익이 발생할 수 있다.

재현자료는 데이터 이용 활성화를 위하여 무료로 제공하는 것이 바람직하다. 그러나 재현자료를 다양한 수준으로 생성할 수 있기 때문에, 많은 노력을 투입하여 원자료와 유사한 분석 결과를 재현하도록 생성한 통계적 재현자료의 경우 적절한 보상이 이루어지는 것이 생태계 활성화에 바람직할 수 있다. 이는 넷플릭스 등의 동영상 구독 서비스에서 낮은 화질의 동영상에 대해서는 적은 비용을 지불하고, 고품질로 동영상을 시청하게 하는 서비스에 대해서는 많은 비용을 지불하는 것과 유사하다. 여타의 IT 서비스처럼, 데이터 생태계에서도 재화 서비스의 편의성에 대한 보상 개념이 반영되고 강화될 필요가 있다.

4.6 다자간 피처 스토어의 한계점

본 논문에서 제안하는 다자간 피처 스토어는 여러 기관 간의 데이터를 안전하게 연계하여 그 활용성을 극대화하는데 초점을 맞추었다. 본 소절에서는 다자간 피처 스토어를 구축할 때 발생할 수 있는 여러 실제적인 문제와 한계점에 대해 간단히 논의한다.

먼저 다자간 피처 스토어를 활용한 데이터의 연계와 활용은 개인정보보호법 등의 다양한 법적 제약을 받는 것이 사실이며, 데이터 소유권 및 책임 소재에 대한 명확한 규정이 필요하다. 예를 들어, 기준 데이터(hinge data)를 각 데이터센터 내에서 표준화 데이터와 결합(join)하여 연계의 유용성과 확장성을 높이는 방안을 실행하기 위해서는 기준 데이터가 될 통계등록부와 표준화 데이터의 연계에 법적인 제약 사항이 있는지 등을 확인해야 한다. 또한 클라우드 서비스를 이용하여 망분리된 서로 다른 데이터센터의 데이터를 결합할 때, 신뢰할 수 있는 클라우드 서비스를 구축하는 것도 관건이라고 할 수 있다. 그리고 다자간 피처 스토어는 공통도구, 기준 데이터, 보안 클라우드 등 복잡한 기술 인프라가 필요하며, 이를 구축하고 유지하는 데 드는 비용이 발생할 수 있다. 클라우드를 제외하면 그간 통계청의 사업 성과를 활용할 수 있으므로, 개념

검증 사업 등을 통해 예상 비용을 구체화할 필요가 있다. 마지막으로, 다자간 피처 스토어의 성공적인 운영에는 높은 수준의 데이터 과학자와 데이터 엔지니어링 자원이 필요하며, 이러한 전문 인력이 부족한 경우에는 효율적인 운영이 어려울 수 있다. 이러한 전문인력 확보를 위해서라도 적절한 보상체계가 정립될 필요가 있다.

현재 IT 업계에서는 내부 피처 스토어를 구축하여 데이터 기반 서비스의 효율을 높이고 있다. 우버(Uber)는 실시간 운행 데이터와 고객 행동 데이터를 미켈란젤로(Michelangelo)라는 기업 고유의 피처 스토에에서 관리하여 도착시간 및 수요 예측 모델의 성능을 크게 향상시켰다.[7] 에어비앤비는 숙소 추천 및 가격 예측 모델에서 예약 데이터와 사용자 리뷰 데이터를 집라인(Zipline)이라는 이름의 기업 고유 피처 스토어 플랫폼을 이용해 관리하고, 학습과 추론 단계에서 데이터 일관성을 유지하며 재사용성을 강화하였다.[8] 또한 아마존 AWS, 구글 Cloud 등의 클라우드 컴퓨팅을 이용하는 기업들을 위해, 아마존과 구글은 각각 Amazon SageMaker Feature Store 및 Vertex AI Feature Store라 불리는 피처 스토어 플랫폼을 운영 중이다.

한 기업 내에서 운영되는 플랫폼으로서의 피처 스토어는 AI 모델 학습뿐만 아니라 서비스 제공 과정에서도 활용된다. 예를 들어, 영화 추천 알고리즘을 학습하기 위해 필요한 다양한 피처를 생성하고 학습하는 과정에서 피처 스토어에 저장된 큐레이션된 데이터를 사용하며, 학습된 알고리즘이 새로운 소비자에게 적합한 영화를 추천할 때에도 피처 스토어에 저장된 데이터를 실시간으로 활용한다.

단일 기업 내에서 구축된 피처 스토어는 실시간으로 데이터를 생성하고 제공하며, 인공신경망 등의 훈련과 추론에 동일한 피처를 사용할 수 있는 등의 장점이 있다. 이와 다르게, 다자간 피처 스토어는 데이터 분석 및 모형 학습에 연계된 데이터를 이용한다는 장점이 있지만, 실시간 데이터 제공은 데이터 보유기관과 데이터센터의 데이터 현행화가 동시에 이루어져야 하는 등, 상대적으로 복잡한 측면이 있다. 따라서 초창기에는 일정한 주기를 가지는 통계의 생산 및 구독 지원, 정적인 모형 구축 및 분석에 더 적합한 플랫폼이라고 할 수 있다.

5. 결론

여러 기관 사이의 데이터를 연계하고, 연계한 데이터에 대한 전통적인 통계적 분석뿐만 아니라 이를 활용하여 차세대 인공지능 서비스를 개발하는 데까지 나아가기 위해서는, 기존의 데이터 공유 플랫폼 형식을 넘어설 필요가 있다. 이를 위해 본 논문에서는 데이터 연계·활용 저해 요인을 분석하고, 최신 프라이버시 강화 기술을 소개하였다. 또한 국내에 분야별로 조금씩 다르게 이해되고 있는 용어와 개념에 대하여 정리하였다. 본 논문의 4장에서는 핵심 주제인 신기술을 활용한 데이터 연계·활용 방안으로서 다자간 피처 스토어(multi-party feature store)를 제안했다. 주요 요소는 공통

7) https://www.uber.com/en-KR/blog/michelangelo-machine-learning-platform/
8) https://www.slideshare.net/slideshow/zipline-airbnbs-machine-learning-data-management-platform-with-nikhil-simha-and-andrew-hoh/102213782

도구 개발, 재현자료(synthetic data)를 활용한 데이터 트윈(data twin) 구축, 기준 데이터(hinge data) 활용, 데이터 제공 자체가 아닌 서비스 재화의 활용성에 따른 보상 체계 등이다. 이러한 요소들은 개인정보보호와 자료 유용성 확보는 물론이고, 사용자 편의성을 향상해 데이터 이용 규모를 확대하는 것을 목표로 고안되었다.

본 논문에서는 다자간 피처 스토어의 개념과 얼개를 제시하였다. 특히 데이터 생태계 구축이라는 문제에 다룸에 있어, 데이터 거래가 아니라 피처 서비스라는 측면에서 접근하자는 관점의 변화를 제안하였다. 이를 구체적으로 실현하기 위해서는 개념검증 사업을 수행하여 세부 계획을 수립할 필요가 있다.

민간의 데이터 거래 플랫폼에서도 기존의 참여 기업들을 중심으로 다자간 피처 스토어를 구축할 수 있으나, 통계청과 같은 정부기관이 다자간 피처 스토어를 구축·운영한다면 더욱 바람직할 것이다. 전문성과 풍부한 경험을 갖춘 통계청과 같은 기관은 공공데이터 과학자(public data scientist, specialized data steward)로서의 역할을 더욱 효율적으로 수행하면서, 동시에 신뢰할 만한 데이터 혹은 피처 생태계 관리자가 될 수 있기 때문이다. 앞으로 다자간 피처 스토어에 대한 논의가 활발하게 이루어지고, 이를 발판으로 편리하고 효율적인 생태계 조성을 향해 나아갈 수 있기를 기대한다.

감사의 글

본 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(0769-20240034)

(2024년 12월 27일 접수, 2025년 2월 24일 수정, 2025년 3월 5일 채택)

참고문헌

김학래 (2021). 국가 데이터의 의미적 표현과 연계를 위한 데이터맵 지식 모델, <디지털콘텐츠학회논문지>, 22(3), 491-499.

김현태, 장가영 (2023). 데이터 가명·익명처리 기법의 현황과 대안: 재현데이터를 중심으로, 보험개발원 연구보고서.

관계부처 합동 (2016). 개인정보 비식별 조치 가이드라인 - 비식별 조치 기준 및 지원·관리체계 안내.

개인정보보호 위원회 (2022). 가명정보 처리 가이드라인

박민정, 김항준 (2016). 마이크로데이터 공표를 위한 통계적 노출제어 방법론 고찰 (Statistical Disclosure Control for Public Microdata: Present and Future), <응용통계연구>, 27(1), 1-19.

박민정, 김정연 (2017a). 재현자료 작성 방법론 검토, 통계청 통계개발원.

박민정, 박영옥, 황영자 (2017b). 통계데이터센터 반출 결과물의 통계적 노출제어 가이드라인(안), 통계청 통계개발원.

박민정, 이용희, 권성훈 (2018a). 차등정보보호에 관한 연구, 통계청 통계개발원.

박민정, 신우람, 박천영, 서영오 (2018b). 그리드별 통계서비스의 비밀보호 방안, 통계청 통계개발원.

박민정, 한정임, 박노성 (2020). 통계데이터센터 DB를 활용한 재현자료 생성 방법 연구, 통계청 통계개발원.

변준석, 박민정, 천정희 (2020). 데이터 거래 활성화를 위한 구독통계 플랫폼 구축 사전 연구, 통계청 통계개발원.

안성빈, 트랑 도안, 이주희, 김지우, 김용재, 김윤지, 윤창원, 정성규, 김동하, 권성훈, 김항준, 안정연, 박철우 (2023). 유용성과 노출 위험성 지표를 이용한 재현자료 기법 비교 연구 (A comparison of synthetic data approaches using utility and disclosure risk measures), <응용통계연구>, 36(2), 141-166.

통계청 (2023). 통계 작성 및 통계자료 제공을 위한 비식별화 가이드라인.

Dwork, C. (2006). Differential Privacy, In 33rd International Colloquium on Automata, Language and Programming, Part II (ICALP 2006), Springer, Venice, Italy, 1-12.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P. P. (2012). *Statistical Disclosure Control*, John Wiley & Sons Ltd.

ISO/IEC (2018). ISO/IEC FDIS 20889:2018(E) Privacy Enhancing Data De-Identification Terminology and Classification of Techniques.

Li, L. E., Chen, E., Hermann, J., Zhang, P., and Wang, L., “Scaling Machine Learning as a Service,” Proceedings of The 3rd International Conference on Predictive Applications and APIs, PMLR 67:14-29, 2017.

Little, R. J. A. (1993). Statistical Analysis of Masked Data, *Journal of Official Statistics*, 9(2), 407-426.

OECD (2023), "Emerging Privacy-Enhancing Technologies: Current Regulatory and Policy Approaches", OECD Digital Economy Papers, No. 351, OECD Publishing, Paris (https://doi.org/10.1787/bf121be4-en)

Park, M-J., Kim, H. J. and Kwon, S. (2024). Disseminating Massive Frequency Tables by Masking Aggregated Cell Frequencies, *Journal of the Korean Statistical Society*, 53, 328-348.

Rubin D.B. (1993). Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 9(2), 461-468.

UN (2023). *The UN Guide on Privacy-Enhancing Technologies for Official Statistics*.

Wasserman (2012). Minimaxity, Statistical Thinking and Differential Privacy, *Journal of Privacy and Confidentiality*, 4(1), 51-63.

Design and Operation of Multi-Party Feature Store (MPFS) for Secure and Privacy-Preserving Data Integration and Utilization

Min-Jeong Park[9], Sungkyu Jung[10]

Abstract

This paper proposes the concept, design and operational framework of the Multi-Party Feature Store (MPFS) as a platform for data linking, utilization, and trading. The MPFS is a platform that incorporates the latest Privacy-Enhancing Technologies (PET) to reflect the current constraints in South Korea and supports microdata sharing by linking data from multiple institutions. This enables advanced statistical analysis and supports the development of artificial intelligence (AI) models. The platform aims to significantly enhance user-centric data accessibility and efficiency, with four key characteristics. First, users can conduct preliminary analyses conveniently and without restrictions by utilizing synthetic data through a data twin before directly accessing sensitive original data at data centers, thereby saving time and costs. Second, by efficiently linking multi-institutional data using hinge data (such as the National Statistical Office's statistical registers) based on population standards, the platform supports data integration, advanced analysis, and machine learning model development. Third, after completing the preceding processes, users can safely analyze the linked data repeatedly in a secure cloud environment and subscribe to the results in API format. Fourth, for data providers and platform operators, a reward mechanism is established based on the value and demand of the shared data and features, fostering a sustainable data platform ecosystem. This paper first discusses various factors that hinder data linking, integration, and utilization, and the latest PETs that address these issues. It then presents the blueprint of the multi-party feature store platform utilizing the resources and the latest technologies currently held by the National Statistical Office, before discussing the operational framework and expected benefits.

Key words : privacy-enhancing technologies, data center, data twin, synthetic data, hinge data, statistical register, multi-party feature store

9) (Corresponding author) Senior Deputy Director, Research Planning Division, Statistics Research Institute, Statistics Korea, Hanbatdaero 713, Seo-gu, Daejeon 35220. E-mail: mjstat@korea.kr

10) Professor, Dept. of Statistics, Seoul National University, Gwanak-ro 1, Gwanak-gu, Seoul 08826, E-mail: sungkyu@snu.ac.kr