

NRF-간행물심의번호

NRF-20141223-1-20

NRF ISSN 2586-1131
ISSUE REPORT

2021_20호

인공지능 시대의 데이터 공급망 관리

- I. 서론
- II. 데이터 공급망
- III. 데이터 수집과 클라우드소싱
- IV. 데이터 품질과 거버넌스
- V. 관련 국내외 동향: 정책과 사례
- VI. 결론

CONTENTS

Ⅰ	서론	1
Ⅱ	데이터 공급망	3
	1. 데이터 공급망 관리	3
Ⅲ	데이터 수집과 클라우드소싱	8
	1. 데이터 수집과 라벨링	8
	2. 클라우드소싱	11
Ⅳ	데이터 품질과 거버넌스	14
	1. 메타데이터와 데이터 품질	14
	2. 데이터 거버넌스	18
Ⅴ	관련 국내외 동향: 정책과 사례	22
	1. EU	22
	2. 미국	24
	3. 중국	26
	4. 한국	28
	5. 인도	31
Ⅵ	결론	32
■	참고문헌	33

표 및 그림 목차

〈표 1〉 해외 클라우드소싱 기업 현황	11
〈표 2〉 DCAT 클래스(2.0 버전 기준)	16
〈표 3〉 데이터 품질 평가 지표	17
〈표 4〉 EU 데이터 거버넌스 법안 주요 내용	18
〈표 5〉 가격책정방법별 장단점 비교	21
〈표 6〉 2014년 이후 EU 데이터 경제 정책 기초	22
〈표 7〉 유럽연합 내 빅테크 기업 제재 현황	23
〈표 8〉 미국 바이든 행정부 정책 기초	24
〈표 9〉 미국 내 데이터 기업 데이터 유출 사례	25
〈표 10〉 2015년 이후 중국 데이터 경제 정책 기초	26
〈표 11〉 중국 금융권 데이터 수집 및 활용 현황	27
〈표 12〉 데이터담 사업을 통한 인공지능 학습용 데이터 구축 참여 사례	28
〈표 13〉 인공지능 학습용 데이터 주요 내용	30
[그림 1] 기계학습 프로젝트 작업별 소요 시간 비중	8
[그림 2] 아태지역 데이터 수집 및 라벨링 시장 규모 및 성장 예상치	10

I. 서론

ICT 기술이 다양한 산업과 융합하면서 디지털 전환(Digital transformation)이 빠르게 진행되고 있다. 산업의 디지털화는 필연적으로 수집되는 데이터의 양을 폭발적으로 증가시켰으며, 대규모의 정형/비정형 데이터를 활용한 인공지능 기술의 발달과 함께 데이터가 실물 경제의 한 축으로 떠오르게 되었다. 2016년 세계경제포럼에서는 이러한 디지털 전환과 데이터 경제로의 이행을 4차 산업혁명이라고 명명하였다. 4차 산업혁명으로 인해 인공지능 기술은 국가 경쟁력을 평가하는 하나의 기준이 되었다. 빅데이터 및 인공지능 기술은 방대한 양의 데이터에서 의미 있는 정보를 추출하고 나아가 경제적인 부가가치를 창출한다. 특히 최근 들어 인공지능 기술의 개발 모멘텀이 모델링 중심에서 데이터 중심으로 옮겨가면서 대량의, 그리고 양질의 데이터를 확보하기 위한 국가 간, 지역 간 경쟁이 심화되고 있다.

ICT 기술 중에서도 IoT(Internet of Things) 기술이 발전하면서 사람 간 연결을 넘어 사람과 사물, 사물과 사물 간 연결성이 강화되는 초연결 사회로 진입하게 되었다. 초개인화된 맞춤형 제품 및 서비스 개발을 위해 기업 간 협업이 활발해지고 있으며, 생산에서 소비에 이르는 전 과정에서 데이터 기반 의사결정이 이루어지면서 제조생산성 및 서비스 효율성이 향상되고 있다. COVID-19 상황이 지속되면서 이러한 기조는 더욱 가속화되고 있는 실정이다. 특히 디지털 전환, IoT, 디지털 경제는 원격/재택근무를 시작으로 비대면 노동시장을 확대시켰다. 특히 데이터의 양과 질이 기업의 핵심 역량으로 주목받게 되면서 클라우드소싱을 통한 데이터 수집 및 가공 작업의 수요가 증가함에 따라 최근 몇 년간 데이터 라벨링 기업들이 크게 성장하였다. 스케일 AI, 라벨박스, 하이브, 클라우드 팩토리 같은 데이터 라벨링 기업들이 꾸준히 투자를 유치하며 성장하고 있으며, 해당 시장 역시 빠르게 성장하고 있다¹⁾.

기계학습, 인공지능을 이용한 데이터 기반 제품 또는 서비스를 개발 및 운영하기 위해서는 기존의 산업과 마찬가지로 생산, 유통, 사후관리 전반에 걸친 종합적인 지식과 경험이 필요하다. 좋은 데이터란 무엇인지 정의하는 것에서부터, 양질의 데이터를 충분하게 준비한 후에도 기술 개발, 테스트, 상용화, 피드백 채널 관리에 이르기까지 정의하고 책임소재를 분명히 해야 한다.

1) 이지현·우창완. (2020). 데이터 라벨링으로 만드는 혁신, 이슈분석. 한국지능정보사회진흥원.

이러한 데이터 제품 및 서비스의 상용화를 위해서 학계와 업계에서는 데이터 품질 평가 기법과 데이터 거버넌스에 대한 논의가 심도 있게 이루어지고 있다.

데이터 제품 및 서비스의 개발 및 운영 전반에 걸친 데이터 생애주기를 종합적으로 관리하기 위해서는 공급망 관리의 관점에서 접근해야 한다. 공급망(Supply Chain)이란 원재료를 획득하고, 이 원재료를 중간재나 최종재로 변환하고, 최종제품을 고객에게 유통시키기 위한 조직 및 비즈니스 프로세스의 전체 네트워크를 가리킨다. 공급망 관리(SCM, Supply Chain Management)란 이러한 공급망에서 필요한 정보가 원활하게 흐르도록 지원하는 시스템을 일컫는다. 전통적인 산업에서 발전해 온 이러한 공급망 관리 기법은 데이터를 원료로 하는 데이터 제품 및 서비스의 개발 및 운영에도 필요한 개념이다.

본고에서는 데이터 기반 제품 및 서비스의 개발 및 운영을 위한 데이터 관리 분야의 현황을 공급망 관리의 관점에서 조망한다. 특히 데이터의 수집 및 가공을 위한 클라우드소싱, 데이터의 품질 측정 및 이를 기반으로 데이터를 유통하기 위한 데이터 표준, 준비된 데이터를 기계학습과 인공지능 등 데이터 기반 제품 개발 단계에서 효율적으로 활용하는 방법들에 대해서 단계별로 짚어보면서 인공지능 시대에서의 데이터 공급망 관리에 관한 시사점을 도출하고자 한다.

II. 데이터 공급망

1 데이터 공급망 관리

데이터는 제품 및 서비스 개발을 위한 새로운 원료가 되고 있다. 기업이 데이터를 통해 가치를 창출하기 위해서는 수집, 생성, 가공, 정리, 저장, 활용을 위한 프로세스를 구축해야 한다. 이러한 맥락에서 기업이 경쟁 우위를 확보하기 위해서는 데이터의 역할과 흐름을 이해하고 관리할 수 있어야 한다. 따라서 제조 공급망의 이론과 경험을 바탕으로 디지털 및 지식 경제의 관점에서 데이터 공급망에 대한 연구 및 적용이 이루어지고 있다. 전통적인 산업에서 데이터는 상품 및 서비스 생산을 위한 공정 과정을 효율화할 수 있는 부차적인 정보를 얻기 위한 수단이었지만, 데이터 공급망 관점에서 데이터는 공정의 입력이자 출력으로, 즉 중간재 또는 최종재이다.

데이터 공급망에서의 교환 요소(element of exchange)는 데이터, 메타데이터, 정보, 지식 등이다²⁾. 정보통신기술의 발전과 더불어 데이터에 대한 접근성과 보편성이 높아짐에 따라 데이터에서 유용한 정보와 지식을 추출하기 위해서는 올바른 접근 방식과 고도화된 기술이 필요해졌다. 센서 기술, 인터넷, 무선통신, 저렴한 메모리 등 새로운 데이터 환경에 의해 데이터 마이닝과 지식 발견을 위한 여러 기술들이 추가적으로 개발되었다. 이러한 상황에서 데이터를 교환요소로 간주하고 관리하려면 데이터의 비정형성을 극복하기 위한 메타데이터의 효율적인 생성 방법, 데이터의 품질을 정량적으로 표현할 수 있는 방법, 개인정보 보호 및 보안 기술들이 뒷받침되어야 한다.

데이터는 단순히 그곳에 존재하고 있는 것이 아니다. 기업 내에서만 하더라도 많은 부분이 기업 내 여러 부서에서 의도적으로 생성되고, 일부는 외부에서 특정 목적을 위해 수입되며, 때로는 제3자(고객 등)가 동의하에 자의로 제공하기도 한다. 이러한 다양한 데이터의 원천이 전체적인 관점에서 관리되지 않으면 여러 가지 문제가 발생하게 된다. 첫째, 중복 데이터 수집의 문제다. 특정 유형의 데이터를 복수의 부서가 필요로 할 때, 관리자 부재 시 동일한 데이터임에도 불구하고 각각의 부서가 독립적으로 수집하는 상황이 발생할 수 있다. 둘째, 데이터의 일관성에 문제가

2) Spanaki, K., Gurguc, Z., Adams, R., & Mulligan, C. (2018). Data supply chain(DSC): research synthesis and future directions. *International Journal of Production Research*, 56(13), 4447-4466.

발생할 수 있다. 팀이 서로 독립적으로 데이터를 수집할 경우, 내용상 같은 데이터임에도 다른 데이터 소스를 활용하게 되면 Single source of Truth 원칙이 훼손될 수 있다. 셋째, 조직 외부, 특히 인터넷에서 수집되는 데이터는 여러 가지로 모호한 측면을 가지고 있다. 수집된 데이터의 출처에서부터 수집 당시의 환경에 대한 불명확한 기록 등 모호성의 원천은 다양하다. 넷째, 의도를 가지고 데이터를 수집할 경우 데이터 자체에 편향이 발생할 가능성이 있다. 자신의 목표를 뒷받침하는 데이터를 선별적으로 수집하거나, 여러 데이터 원천들을 활용하여 작업을 수행한 후 가장 유리한 결과를 얻을 수 있는 원천을 사후에 선택할 수도 있다. 이러한 문제를 극복하기 위해 기업 내부에서 데이터 공급망 관리의 관점에서 일련의 프로세스를 구성할 수 있다. Treder는 데이터가 어떻게 공급망 내에서 관리되어야 하는지 일곱 단계로 나누어 설명하였다³⁾.

첫째, 데이터는 수집한 직후 최대한 검증되어야 한다. 데이터를 사용할 때마다 재검증하는 것보다 원천데이터를 깨끗하게 유지하는 것이 선결과제이다. 모든 데이터가 가능한 한 높은 품질을 가질 필요까지는 없지만, 각 데이터의 품질이 어느 정도인지 항상 파악하고 있어야 한다. 또한 데이터 자체를 검증하고 정리하는 것보다 기존에 보유하고 있는 데이터와의 동기화, 그리고 필요한 경우 익명화, 암호화할 수 있는지가 중요하다. 특히 비정형 데이터의 비중이 높아지고 있는 현재 환경에서 메타데이터는 한층 신경 써서 관리하여야 한다.

둘째, 데이터는 각각 고유한 구조를 가지고 있으며, 이 구조는 조직 내 이해관계자들이 모두 잘 이해할 수 있는 형식으로 기술되어야 한다. 제대로 문서화된 데이터 모형은 이해관계자들이 데이터를 효율적으로 활용하기 위해 매우 중요하다. 정형데이터의 경우 기존의 관계형 데이터베이스 등을 활용하여 목적에 부합하는 데이터 스키마를 설계하고 문서화할 수 있다. 그러나 비정형 데이터의 경우 데이터 원본의 비정형성에 의해 구조화 및 문서화가 어려울 수 있다. 그럼에도 불구하고 데이터를 이해하고 활용하려면 최소한 비정형 데이터에 대한 메타데이터가 잘 정리되어 있어야 하며, 첫 번째 단계에서 메타데이터를 강조한 이유이기도 하다. 즉 데이터를 잘 구조화하기 위해 필요한 메타데이터 수준을 목적에 맞게 설계하는 것이 핵심이다.

셋째, 데이터의 품질을 지속적으로 평가한다. Garbage in, Garbage out은 데이터 공급망에서는 공리와 다름없다. 수집한 데이터의 품질이 낮다면 사용해서는 안 된다. 더욱 나쁜 것은 데이터의 품질이 좋은지 나쁜지조차 알 수 없을 때다. 기계학습이나 인공지능 제품과 서비스, 이를 개발 및 운영하기 위한 데이터 공급망이 다른 전통적인 산업과 가장 큰 차이를 보이는 것은

3) Treder, M. (2020). The Data Supply Chain. In The Chief Data Officer Management Handbook(pp. 35-46). Apress, Berkeley, CA.

출력된 결과에서 입력의 품질을 역으로 추정하기 어렵다는 사실이다.

넷째, 데이터는 끊임없이 정리한다. 처음에는 높은 품질을 지녔던 데이터라고 할지라도 새로운 데이터가 추가되고 구조의 변화가 발생할 때마다 일관성과 완결성, 접근성의 측면에서 품질은 감가상각된다. 데이터의 품질을 유지하기 위해서는 보관할 데이터와 폐기할 데이터를 구분하고, 데이터 생애 주기를 관리하는 규정을 적극적으로 관리하고 적용해야 한다. 최소한 GDPR (General Data Protection Regulation)과 같은 개인정보 보호 규정을 준수하기 위한 관리 체계를 갖추어야 한다.

다섯째, 앞선 모든 단계를 완벽하게 수행하여 품질 좋은 데이터를 대량으로 수집, 정리, 보관할 수 있게 되었다면, 이제는 이해관계자들이 적절하게 데이터를 활용할 수 있는 기반을 마련해야 한다. 데이터 사용자는 이와 관련하여 몇 가지 사실을 알 수 있어야 한다. 우선 데이터 사용자는 필요한 데이터를 어디에서 찾을 수 있는지, 찾은 데이터는 어떻게 사용하는지, 문의 사항이나 변경 요청을 할 수 있는 데이터 소유자는 누구인지, 이 데이터를 이전에 사용했던 이력을 확인할 수 있는지 등을 알 수 있어야 한다. 또한 데이터 사용자는 데이터의 속성에 대한 확신이 필요하다. 예를 들어 데이터 사용자는 자신에게 제공된 데이터가 정확하고, 일관성이 보장되며, 최신 상태이고, 기재된 설명과 일치하며, 충분한 가용성을 가지고 있을 것이라고 확신할 수 있어야 안심하고 데이터를 사용할 수 있다. 데이터를 지속적으로, 그리고 안정적으로 활용할 수 있기 위해서는 특히 해당 데이터와 관련된 이슈나 변경사항이 사전에 사용자에게 투명하게 전달되어야 하고, 데이터 보안 및 개인정보 보호 역시 확실해야 한다.

여섯째, 사용자에게 데이터를 전달하기 위한 문서화는 필수적이다. 데이터 형식, 인터페이스, 접근 방식, 데이터 소유자, 데이터 변경 프로세스 등은 모두 사용자에게 투명하게 공개되어야 한다. 데이터 카탈로그, 검색 사이트 등을 활용하여 사용자에게 친숙한 방식으로 관련 정보를 제공할 수 있다. 또한 이러한 사용자 접점은 피드백 채널을 제공하여 개선 요청을 받을 수 있어야 한다.

마지막으로 일곱째, 단일한 데이터라고 하더라도 데이터의 사용 목적은 다양하다. 예를 들어 이커머스 웹페이지 사용자 로그 데이터의 경우, 해당 사이트에서 각각의 사용자에게 자신의 사용 이력을 보여주는 직접적인 활용에서부터, 해당 사이트에서 발생하는 매출, 클릭률 등 주요 성과 지표 계산, 사이트 개편 등에 대한 A/B 테스트 보고서 작성, 개인화 인공지능 학습, 정기/비정기 감사에 대한 원천데이터로의 활용에 이르기까지 사용처가 다양하며, 이러한 목적들이 순차적으로 달성되는 프로세스를 구성할 수도 있다. 이러한 다양한 사용 목적에 대응하기 위해 최우선적으로

전제되어야 하는 것은 일관성이다. 데이터 사용자는 절대로 원천데이터를 수정할 수 없어야 한다.

이러한 공급망 관리를 위한 데이터 처리 단계를 수행하기 위해서는 공급망 관리 조직 및 프로세스를 구축해야 한다. 하버드 비즈니스 리뷰에서는 여덟 가지 단계로 이를 요약 정리하였다⁴⁾.

1. 관리 책임을 설정한다. 최고 데이터 책임자 또는 제품 관리자는 직원으로부터 데이터 공급망 관리자를 지명하고, 공급망 전체에 걸쳐서 각 부서의 책임자를 모집해야 한다. 그리고 데이터 소유권을 확실히 한다.
2. 데이터 제품을 만들고 유지하는 데 필요한 원천 데이터 및 관련 비용, 품질 및 요구사항을 식별하고 문서화한다.
3. 공급망에 대해 문서화한다. 데이터 수집 및 생성, 데이터 유통, 데이터 분석 및 활용, 데이터 평가 및 피드백에 이르기까지 각 단계를 설명하는 순서도를 개발한다.
4. 데이터 공급망 각 단계를 관측하는 측도를 개발한다. 각 단계별로 소요되는 시간 및 비용을 시작으로, 원천데이터에서부터 각 단계별 부산물, 데이터 제품 전반의 품질 측도를 개발한다.
5. 프로세스 제어 방식을 구축하고 요구사항에 대한 적합성을 평가한다. 단계 4의 측도를 사용하여 공정을 관리하고, 단계 2에서 설정한 요구사항의 충족 정도를 통해 프로세스를 실시간으로 평가한다.
6. 단계 5에서 확인된 요구사항과 현재의 괴리가 발생하는 지점을 단계 3의 순서도를 활용하여 파악한다.
7. 모든 단계를 지속적으로 모니터링하고, 이를 통해 각 단계를 개선한다.
8. 데이터 소스를 “적격화(Qualify)”한다. 품질 좋은 외부 데이터를 얻기 위해 공급업체를 고용하게 되면 이를 평가할 수 있어야 하는데, 기업 내 데이터 공급망에서 측정되는 여러 측도들은 공급업체가 제공하는 데이터의 품질 평가에 활용할 수 있다.

데이터가 경제활동의 생산품이자 최종소비제품으로 기능하게 됨에 따라 비교적 최근에 데이터 공급망에 대한 논의가 이루어지고 있다. 데이터를 공급망을 흐르는 주요한 교환요소로 간주했을 때, 이를 관리하기 위한 방법은 기존 산업의 공급망 관리와 관련한 이론과 실제에서 많은 부분을 참고할 수 있다.

4) <https://hbr.org/2021/06/data-management-is-a-supply-chain-problem>

실제 산업에서 알아낸 인공지능 관련 경험적 지식 또한 데이터 공급망 관리를 위해 중요하다. 맥킨지는 2017년 인공지능 기술 적용 촉진 방안에 대한 제언(Smartening up with AI)에서 인공지능 성능 향상을 위한 경험적 지식을 5가지로 정리하였다⁵⁾. 첫째, AI가 무엇을 할 수 있는지 파악하고, 용례의 우선순위를 정하고, 경제성을 간과하지 말 것. 둘째, 외부의 리소스를 활용하면서 내부적인 핵심 역량을 개발할 것. 셋째, 가능한 한 세분화된 원본 데이터를 적재하고, 비정형 데이터를 가공하여 사용할 것. 넷째, 도메인 지식을 최대한 활용할 것. 다섯째, 파일럿 테스트와 시뮬레이션을 활용하여 빠르게(Agile) 개발할 것. 이 중 원본 데이터의 적재는 반드시 이루어져야 한다. 이러한 원본 데이터는 시간이 지나고 규모가 커짐에 따라 사용성이 떨어질 수도, 아니면 새로운 지식을 제공해 줄 수도 있기 때문에, 최대한 원천데이터를 있는 그대로 보관 및 활용할 수 있는 견고한 파이프라인을 구축하는 것이 중요하다.

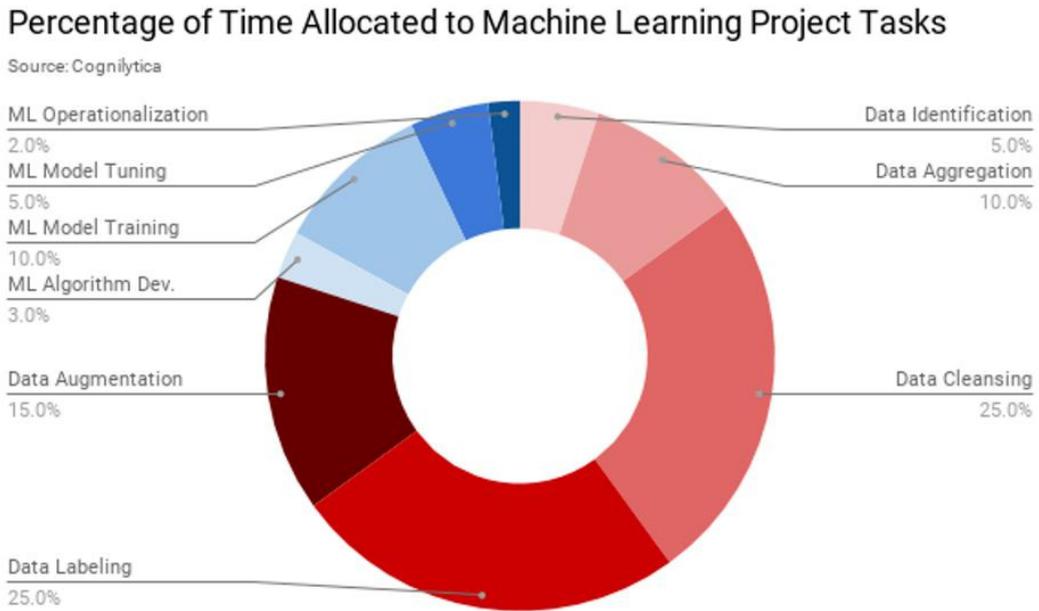
5) Smartening up with artificial intelligence. 2017. 4. McKinsey & Company

Ⅲ. 데이터 수집과 크라우드소싱

1 데이터 수집과 라벨링

데이터 공급망 관리의 첫걸음은 데이터를 수집하는 것이다. Garbage in, Garbage out의 공리 하에서 도입부라고 할 수 있는 데이터 수집 단계는 이어지는 다른 어떤 단계들보다 중요하다고 할 수 있다. 아이러니하게도 기계학습과 인공지능 분야에서 가장 많은 시간이 소요되는 과정이기도 하다.

[그림 1] 기계학습 프로젝트 작업별 소요 시간 비중⁶⁾



6) <https://www.forbes.com/sites/cognitiveworld/2020/02/02/the-human-powered-companies-that-make-ai-work>

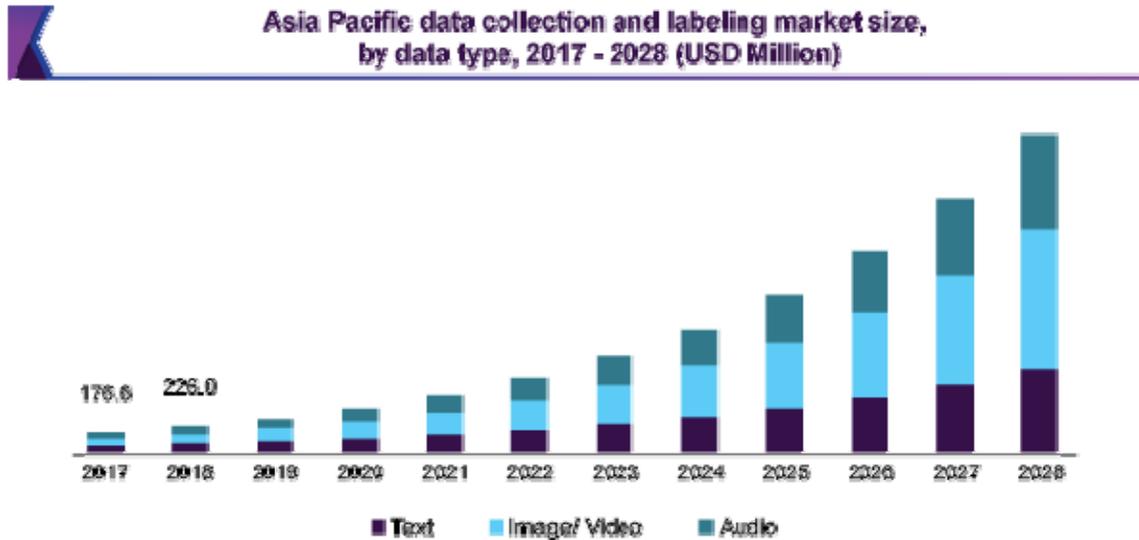
포브스에 따르면 기계학습 프로젝트는 데이터 준비, 기계학습 알고리즘 개발, 모델 훈련, 모델 파라미터 튜닝, 모델 최적화 등의 작업으로 이루어지는데, 전체 작업 수행 시간 중 데이터 준비(데이터 확인, 증강, 정리, 라벨링) 부분이 65%를 차지한다. 특히 최근에는 기계학습이 새로운 응용분야에 적용되기 시작하면서 학습용 데이터가 충분하지 않기 때문에 더욱 많은 시간과 비용이 소요되고 있다. 기계 번역이나 컴퓨터 비전과 같은 기존 애플리케이션들은 수십 년 동안 축적된 방대한 양의 학습 데이터를 향유할 수 있는 것과는 대조적으로, 장애인을 위한 최적 이동경로 탐색이나 제한된 지역의 풍토병과 관련한 전염경로 예측과 같은 새로운 분야에 있어서는 데이터가 매우 적거나 아예 없는 경우도 비일비재하다. 수작업을 통한 데이터 라벨링은 비용이 많이 들고, 특히 해당 분야에 대한 전문 지식이 필요할 가능성이 높기 때문에 기계학습 또는 인공지능 개발 기업이 직접 실행하기 어렵다. 이 문제는 모든 새로운 응용분야에서 발생하고 있는 문제다.

여기서 데이터 라벨링 작업이란 기계 학습 프로그램이나 인공지능 모형을 훈련시킬 수 있게끔 원천 데이터에 추가 정보를 기입하는 제반 행위를 말한다. 간단한 예로는 사진을 보고 고양이, 개, 기타 등으로 구분해서 범주화하는 것이 라벨링 작업이다. 좀 더 어려운 라벨링 작업의 예로는 한영번역을 위한 인공지능을 훈련시키기 위해 서로 대응하는 영어-한국어 텍스트를 준비하는 경우이다. 만약 원천 데이터가 영어라면 라벨은 이에 대응하는 한국어 데이터가 되고, 원천 데이터가 한국어라면 그와 반대가 된다. 이런 경우 라벨링의 품질은 번역가의 실력에 비례할 가능성이 크다.

데이터 인식, 증강, 정리에 있어서는 여러 가지 자동화기법이 적용될 수 있으나, 라벨링 작업에 대해서만큼은 사람이 직접 참여하는 비율이 훨씬 높다. 글로벌 마켓 인사이트가 펴낸 보고서⁷⁾에 따르면 2019년 기준 전체 데이터 라벨링에서 75%가 사람이 직접 참여하는 구조였고, 2026년까지 자동화된 라벨링보다 사람이 개입한 라벨링 작업이 주를 이룰 것이라고 예상했다. 사람이 직접 데이터를 라벨링 하는 경우는 다시 두 가지로 나눌 수 있는데, 기업 내 직원이 직접 참여하는 경우, 그리고 데이터 라벨링을 전문으로 하는 기업에 아웃소싱을 하는 경우이다. 앞서 설명한 바와 같이 새로운 응용분야에 대한 도전이 계속되고 있는 만큼 데이터 라벨링 작업을 아웃소싱하는 사례가 늘고 있다.

7) <https://www.grandviewresearch.com/industry-analysis/data-collection-labeling-market>

[그림 2] 아태지역 데이터 수집 및 라벨링 시장 규모 및 성장 예상치



Source: www.grandviewresearch.com

2 클라우드소싱

데이터 라벨링 수탁 기업은 “클라우드소싱”이라고 불리우는, 다수의 대중이 온/오프라인 플랫폼을 통해 데이터 라벨링 작업을 수행하는 방식을 사용한다. 이러한 클라우드소싱 플랫폼 기업은 작업자 모집, 작업 정의, 라벨링 결과 검수, 작업에 따른 임금 지급 등 라벨링 서비스 공급자측을 관리하는 동시에 라벨링 서비스 수요자 측이 적절한 작업자를 설정할 수 있도록 인터페이스와 상담 서비스를 제공한다. 전 세계 클라우드소싱 시장 규모는 2018년 95억 달러에서 2027년에는 1,548달러까지 연 평균 36.5%의 성장률을 보일 것으로 예상되고 있다(Absolute market insights 2020. 1, Crowdsourcing market 2019-2027). 한국지능정보사회진흥원은 적용 분야, 인력 운영, 수수료 정책의 관점에서 다양한 해외 클라우드소싱 기업의 현황을 정리하였다.

〈표 1〉 해외 클라우드소싱 기업 현황

기업명	인력 운영방식 및 수수료 정책	적용 분야
Amazon Mechanical Turk, Inc. (미국)	<ul style="list-style-type: none"> - 190개국 50만 명 - 태스크 당 임금을 지급하는 모델을 통해 신속하게 인력 소싱 - AWS SDK를 통한 작업 의뢰자의 API 지원 - 작업 의뢰자(고객사)가 설정한 작업별 임금의 20%를 수취 - 원하는 작업자 모집을 위해서는 작업자에게 추가 보상 지불 - 특정 조건에 부합하는 인력 활용 시 132개 조건 별로 상이한 추가 수수료를 부과 	데이터 수집, 주석 달기 등 간단한 작업에 특화
CloudFactory Limited. (미국)	<ul style="list-style-type: none"> - 영국, 미국, 케냐, 네팔에서 1,820명의 핵심 팀원, 5,000명 이상의 클라우드소싱 인력 보유 - 주로 저소득 국가의 인력 소싱 - 구독제 기반 요금 부과, 구독 시간 설정 가능 	데이터 분류, 이미지 주석 달기, 웹서치
Appen Limited. (호주)	<ul style="list-style-type: none"> - 전 세계 140개 이상 국가의 180개 외국어 작업 가능한 40만 명 이상의 인력 보유 - 업무 참가를 위해서는 자격시험을 통과해야 하고, 응시 기회는 2회로 제한 - 프로젝트 매니저가 의뢰자의 타임라인을 설정, 요구사항 상세화 - 임금은 작업 시간에 비례하여 책정 - 클라우드소싱 윤리 원칙 제정 및 적용: 1) Fair Pay, 2) Inclusion, 3) Crowd Voice, 4) Privacy and Confidentiality, 5) Communication, 6) Well-being 	데이터 분류, 주석 달기, 번역, 전사(speech data transcription)

기업명	인력 운영방식 및 수수료 정책	적용 분야
Figure Eight Inc. (미국)	<ul style="list-style-type: none"> - 2007년 설립되어 2019년 Appen Limited. (호주)에 인수됨 - 외주 파트너를 별도로 두고, 이들 인력에 대해 힌디어, 러시아어 사용자 또는 일본인, 독일인과 같은 국적 조건 설정 가능 - 의뢰인은 데이터 보안 유지를 위해 기밀유지협약 하에 특수하게 진행되는 옵션도 선택 가능 	텍스트, 자연어처리, 컴퓨터비전, 데이터증강, 데이터분류
Samasource (미국)	<ul style="list-style-type: none"> - 대표적인 임팩트 소싱 기업으로 케냐, 우간다, 인도, 아이티의 빈민 지역 주민 11,480명을 클라우드워커로 활용 - 소량의 쉬운 작업, 다소 복잡한 작업, 대량의 복잡한 작업(기업용)으로 가격 모델을 분류하여 요금 부과 - 작업자들은 월급 형태로 임금을 받고, 성과에 따라 인센티브도 수령 	컴퓨터비전, 자연어처리, 데이터검증
Scale AI (미국)	<ul style="list-style-type: none"> - 작업 시간당 임금 책정, 월별 최대 작업수를 500개로 제한 - 수작업 → 통계 신뢰도 검증 → 머신러닝 모델 검증 등의 단계를 거쳐 작업한 데이터의 정확성을 제고 	자율주행차량의 사물인식모델 필요 데이터 작업, 드론 및 로봇에 활용되는 데이터 작업 위주
Mighty AI (미국)	<ul style="list-style-type: none"> - 2014년 설립, 2019년 Uber에 인수됨 - 차량 센서 데이터 라벨링을 전문으로 하는 인력 풀 40만 명 	자율주행차량 운행데이터 관련 (사물추적, 의도분류, 교통지표주석) 작업 전문 지원
Clickworker GmbH (독일)	<ul style="list-style-type: none"> - 전 세계 136개국, 190만 명의 대규모 크라우드소싱 인력 보유 - 작업자들은 온라인 테스트 및 훈련을 거쳐 자격 검증 - 자격 급수에 따른 임금 체계 구축 	이커머스, 패션업, 언론, 지식전달 서비스 영역 솔루션 제공
MBH (중국)	<ul style="list-style-type: none"> - 중국 내 경제적으로 낙후된 지역 중심으로 30만 명의 인력 소싱 - 작업자들은 하루 6시간씩 안면 데이터, 의학 촬영본, 도시 촬영 이미지에 대한 태깅 작업 수행 - 아마존의 상품 추천 머신러닝 시스템과 동일한 방식을 채택, 작업자들에게 작업분을 효율적으로 빠르게 배정 	의료, 안전, 도시 등 다양한 분야의 영상 및 이미지 가공 작업 수행, 현재 틱톡의 영상 데이터 중 부적절한 데이터 스크리닝 작업 소싱

기업명	인력 운영방식 및 수수료 정책	적용 분야
Lionbridge AI (미국)	<ul style="list-style-type: none"> - 1백만 명의 준전문가를 통해 300개 이상의 언어 텍스트 데이터와 손글씨 이미지 데이터를 수집 - 번역 콘텐츠 생성, 조정, 감성분석 서비스를 제공 - 한국에서 음성, 손글씨, 얼굴 샘플 데이터를 클라우드소싱 	번역 작업 특화, 언어 데이터 주석 처리, 어휘집 개발, 인공지능 비서 음성인식 모델 고도화
Datapure (미국)	<ul style="list-style-type: none"> - 200명 이상의 풀타임 클라우드소싱 인력 보유 - 작업자들은 약 한 달간의 초기 훈련 후 작업 투입 - 작업 정확도를 높이기 위해 독립된 검수 팀이 데이터를 전수 검사(오류를 찾으면 인센티브 지급) 	자율주행차량에 필요한 데이터 분류, 제거, 라벨링 수행

데이터 수집 수탁 기업들은 클라우드 소싱 인력 풀에 대한 관리와 함께 투입 인력을 최소화하기 위한 자동화된 처리 기법(auto labeling methods) 등의 발전과 함께 수집의 효율성을 높이고 있다. 이러한 클라우드소싱 전문 기업이 성장하고 고객사별로 체계적인 인력 관리가 가능해지면 규모의 경제를 통한 데이터 수집 과정의 효율성 제고와 함께 데이터 품질의 정량적인 평가, 나아가 데이터에 적절한 가격을 산정할 수 있는 기반을 마련할 수 있다.

IV. 데이터 품질과 거버넌스

1 메타데이터와 데이터 품질

2012년 11월 26일 테크크런치(techcrunch), 기즈모도(Gizmodo) 등 몇몇 주요 기술 블로그는 구글이 공공 와이파이 핫스팟 제공업체인 아이코아를 4억 달러에 인수하여 아이코아의 주가가 급등했다고 보도했다. 하지만 이러한 보도는 보도자료 배포업체인 PR Web에 어떤 식으로든 전달된 허위 보도자료에 근거한 것이었다. 나중에 실수가 밝혀졌고, 블로그들은 뒤늦게 그들의 게시물을 수정했다. 이 예시는 잘못된 정보가 네트워크 속에서 얼마나 빨리 전파되는지를 알려준다. 이 블로그들이 실패한 것은 정보 출처의 확인이다. 수집된 최초의 데이터의 품질이 낮을 경우 기계학습과 인공지능 제품 개발을 위한 데이터에 있어서도 심각한 문제를 야기할 수 있다. 학습 데이터가 내재하고 있는 편향성에 의해 발생한 의해 미국의 마이크로소프트사가 공개한 트위터 챗봇 테이(Tay)의 성차별적 발언, 아마존의 초기 인공지능 채용 시스템의 남성 편향, 그리고 데이터 거버넌스의 부재로 인한 최근 국내 스타트업의 챗봇의 개인정보 보호 관련 이슈에 이르기까지 데이터 품질과 관련한 많은 실패 사례가 존재한다.

그렇다면 좋은 데이터란 무엇인지, 특히 본고의 주제인 기계학습과 인공지능 제품 및 서비스를 개발하기 위한 목적에 있어서 좋은 데이터란 무엇인지부터 정리해야 한다. MIT 슬론 매니지먼트 리뷰에서는 다양한 AI 이니셔티브에 대한 분석과 AI 전문가와의 인터뷰를 통해서 고품질 데이터란 무엇인지를 정확성, 완전성, 해석가능성, 그리고 가용성까지 네 가지 차원으로 정리하였다⁸⁾. 정확성이란 데이터 값이 사실과 일치하는지, 표준을 준수하는지를 의미하고, 완전성이란 필수 항목의 누락이 없는지, 사용목적에 맞는 충분한 양이 확보되어 있는지를 의미한다. 데이터의 품질 척도 중 정확성과 완전성이 다른 차원보다 상대적으로 그 중요성이 부각되는 경향이 있지만, 이에 못지않게 중요하면서도 실제 인공지능 서비스의 구현에 있어서 더 큰 병목이 되는 지점은 해석가능성과 가용성 부분이다. 데이터 공급망 측면에서 데이터의 원활한 유통 및 활용을 위해서는 네 가지 차원을 모두 만족하는 양질의 데이터가 필요하다. 관계형 데이터베이스를

8) Vial, G., Jiang, J., Giannelis, T., & Cameron, A. F. (2021). The Data Problem Stalling AI. MIT Sloan Management Review, 62(2), 47-53.

중심으로 축적되어 온 정형 데이터의 경우 잘 정의된 스키마를 통해 데이터를 해석하고 접근할 수 있으나, 현재 생산 및 축적되고 있는 거대한 규모의 비정형 데이터(문헌, 음성, 영상, 사진 등)에 대한 해석가능성과 가용성을 확보하기 위해서는 잘 정의된 메타데이터가 필수적이다.

메타데이터는 “데이터에 관한 데이터”라는 의미로, 한 권의 소설이라는 비정형 데이터가 있다면 책의 제목, 저자, 출판 일시, 출판사 등 소설이라는 데이터에 접근할 수 있는 단서들이 곧 메타데이터라고 할 수 있다. 즉 비정형 원천 데이터의 품질은 메타데이터의 품질과 궤를 함께한다. 이는 앞서 설명한 데이터 수집 단계에서의 데이터 라벨링 작업과도 연관된다. 하지만 메타데이터의 의미에서도 알 수 있듯이, 자칫하면 데이터의 데이터, 그리고 그 데이터의 데이터 같은 식으로 남용하게 되거나, 같은 하나의 데이터임에도 불구하고 해당 데이터에 대한 이해관계자별로 다른 추상화 작업을 거치게 되면 데이터 품질은 가용성의 측면에서 오히려 저하된다.

메타데이터와 관련된 문제를 겪고 있는 대표적인 지역이 EU이다. EU는 수많은 유럽 국가들로 이루어져 있기 때문에 각 국가별로 공공데이터 포털이 존재한다. 이러한 데이터 집합들에 대해 통합적으로 접근하기 위해서는 메타데이터 설계에 있어서 국가 간 합의를 거친 표준이 필요하다. 이를 위해 EU와 W3C는 데이터 카탈로그(Data Catalog) 개념을 제안하였다. 이미 각 국가별로 가지고 있는 메타데이터를 통합하여 수정하기보다는 메타데이터의 메타데이터를 데이터 카탈로그라는 이름으로 구성하는 것을 그 골자로 하는데, 실제 문제를 해결하기 위한 작업과 함께 국제 표준을 구축하는 것이다.

데이터 카탈로그란 “특정 데이터 포털이 어떤 종류의 데이터를 가지고 있고, 그 구조가 어떠한지를 표시”하는, 명명한 대로 카탈로그의 형식으로 표현한 메타데이터이다. 이를 위해 W3C는 2014년에 DCAT(Data CATalogue vocabulary)라는 이름으로 데이터 정보 서술을 위한 어휘 규격 기술표준을 제정하여 발표하였다. 2015년에는 EU 집행위원회 주도로 DCAT-AP(Application Profile)를 추가로 제정하여 보완하였고, 2020년 2월 W3C DXWG(Data Exchange Working Group)에서 2017년부터 작업해온 2.0 버전을 발표하였다.

〈표 2〉 DCAT 클래스(2.0 버전 기준)

명칭	설명
Resource (리소스)	- 데이터 집합과 데이터서비스에 대한 기본/공통 정보(메타데이터)를 서술
DataService (데이터서비스)	- API와 같은 데이터서비스 상품에 대한 정보 서술
Dataset (데이터 집합)	- 파일과 같은 데이터 상품에 대한 관련 정보 서술
Distribution (배포형식)	- 데이터 집합을 유통하기 위한 정보 서술
ConceptScheme (분류체계)	- '주영역-상세영역'과 같이 플랫폼별 데이터 분류체계 표현
Concept (카테고리)	- 플랫폼별 데이터 분류체계 관리 속성 정보 표현
Agent (제공기관)	- 플랫폼 주관사업자 또는 센터 등 데이터 상품을 제공하는 기관에 관한 정보 서술
Catalog (카탈로그)	- 최종 사용자에게 제공되는 데이터 및 관련 서비스의 세부 내역 서술

이러한 메타데이터 기술 표준을 활용하여 데이터의 정확성, 완전성과 함께 최신성, 상호연계성 등의 품질 측도를 메타데이터에 추가하게 되면 데이터의 가치를 산정할 수 있다. 즉 데이터 품질 측도를 표준화하게 되면 품질 관련 메타데이터를 통해 데이터의 품질을 비교할 수 있게 되며, 비교의 결과는 가격의 차이를 설명하는 주요인으로 활용할 수 있다. 즉 적절한 가격을 데이터에 매길 수 있게 되면서 데이터 생산자와 수요자 간 신뢰를 바탕으로 데이터의 거래가 활성화될 수 있다.

한국데이터산업진흥원은 데이터 품질을 평가하기 위한 지표를 완전성(Completeness), 유효성(Validity), 정확성(Accuracy), 일관성(Consistency), 유일성(Uniqueness), 적시성(Timeliness), 활용성(Relevance)까지 총 7개의 차원으로 정리하고 있다.

〈표 3〉 데이터 품질 평가 지표

지표	설명
완전성 (Completeness)	<ul style="list-style-type: none"> - 데이터 생성 시 논리적 설계도인 데이터명세서와 물리적인 데이터가 누락 없이 잘 설계 및 생성되었는지를 평가하는 지표 - 세부항목: 데이터 항목, 레코드, 파일, 물리메타데이터, 관리메타데이터
유효성 (Validity)	<ul style="list-style-type: none"> - 데이터가 정의된 기준에 맞게 유효한 정보의 범위와 형식으로 되어 있는지, 데이터의 기능이 유효하게 서비스되는지에 대한 수준을 평가하는 지표 - 세부항목: 범위, 형식, 목록, 응답, 데이터 기능
정확성 (Accuracy)	<ul style="list-style-type: none"> - 데이터가 실제 메타데이터에서 정의한 대로 정확하게 입력되었는지, 실제 입력된 값이 업무적 요건에 맞게 저장되어 있는지 측정 - 세부항목: 메타데이터, 의미, 계산 및 집계, 선후관계, 파일오류, 내용오류
일관성 (Consistency)	<ul style="list-style-type: none"> - 항목, 레코드, 파일 상호 참조관계에 대한 일관성 확보 수준 및 메타데이터와 데이터 간 일관성이 유지되고 있는지 측정하는 지표 - 세부항목: 참조무결성, 항목 형식, 항목 값, 항목 관계, 메타데이터
유일성 (Uniqueness)	<ul style="list-style-type: none"> - 항목, 레코드, 파일 중복으로 인한 모순의 위험성 확률을 평가하는 지표 - 세부항목: 항목, 레코드, 파일
적시성 (Timeliness)	<ul style="list-style-type: none"> - 사용자가 만족하는 수준의 응답시간으로 데이터가 제공되는지, 데이터 요청으로부터 수집 및 처리되어 제공되기까지의 작업시간이 최적화되어 있는지, 요구된 정보가 최신의 것인지에 대한 수준을 측정하는 지표 - 세부항목: 응답시간, 데이터 제공, 최신 값
활용성 (Relevance)	<ul style="list-style-type: none"> - 사용자가 만족하는 수준의 충분한 정보가 제공되고 있는지, 데이터에 대한 접근이 사용자 편의적인지, 사용자 정보를 유용하게 활용하고 있는지 등을 평가하는 지표 - 세부항목: 친밀도, 효율성, 활용성

데이터 품질을 메타데이터에 표현하기 위한 구체적인 표준 역시 W3C의 DCAT 내의 Namespace 상 DQV(Data Quality Vocabulary)라는 이름으로 정의되어 있다. 따라서 DQV를 준용하여 기술된 품질 메타데이터는 DCAT 표준으로 작성된 메타데이터와 연동할 수 있다. 오해하지 말아야 할 것은, W3C는 품질을 표현하기 위한 언어표준을 제정했을 뿐, 품질 측도 자체를 제시한 것은 아니며, 이는 기본적으로 데이터 집합에 대한 이해관계자들 사이에서 합의를 거쳐야 하는 사안이다. 하지만 데이터 품질과 관련한 정책, 계약, 출처, 품질 인증서 등을 표현할 수 있기 때문에, 해당 정보의 유무 만으로도 상당한 정도의 데이터 품질을 확인할 수 있다.

2 데이터 거버넌스

일반적인 데이터 품질에 대한 논의와 함께, 데이터 기반 제품 및 서비스를 위한 공급망 관리에 있어서 중요한 한 축은 데이터 거버넌스이다. 구글에 따르면⁹⁾ 데이터 거버넌스란 데이터의 보안, 개인정보 보호, 정확성, 가용성, 사용성을 보장하기 위해 수행하는 모든 작업을 가리킨다. 여기에는 사람이 취해야 하는 조치, 따라야 하는 프로세스, 데이터의 전체 수명 주기 동안 이를 지원하는 기술이 포함된다. 2020년 11월 25일 유럽연합집행위원회는 데이터 거버넌스 법안을 발표하면서 공공 부문 데이터의 재사용, 데이터 중개자의 데이터 공유 서비스, 신뢰성 확보, 이타적 목적으로 제공하는 데이터의 수집 및 처리와 관련한 프레임워크 구축을 천명하였다.

〈표 4〉 EU 데이터 거버넌스 법안 주요 내용¹⁰⁾

주요 내용	설명
공공부문 데이터 재사용	<ul style="list-style-type: none"> - 익명화 처리, 영업 기밀 삭제 등을 거친 데이터의 재사용 허용 - 데이터 재사용 요건 및 비용에 대한 내용을 담담할 단일 정보 창구 구축
데이터 공유 서비스 신뢰성 확보	<ul style="list-style-type: none"> - 공공을 위한 개인정보 및 상업적 정보 재사용에 대한 신뢰성 확보 - 데이터 보유자와 이용자 사이의 독립적이고 전문적인 데이터 중개 및 공유 서비스 도입 - 데이터 공유 신뢰성 확보를 위해 유럽연합 회원국마다 데이터 공유 업체 등록 및 관리를 위한 기관 설립과 자격 제한, 규제 감독 등 유럽연합 차원의 규제 프레임워크 구축 - 유럽연합 내 위치와 관계없이 데이터를 사용 가능한 단일 시장 구축
데이터 이타주의	<ul style="list-style-type: none"> - 개인의 동의하에 공익 및 연구 목적으로 유럽연합 회원국 간 국경을 초월한 데이터 풀 구축 - 데이터를 제공하는 비영리 법인이 특정 조건 충족 시 유럽연합에서 승인받은 데이터 이타주의 조직으로 등록 가능토록 법제화 추진
기타	<ul style="list-style-type: none"> - 개인 및 법인의 민원 제기 및 사법적 구제 권리 명시 - 기존 국가 관행이나 정책과 데이터 거버넌스법 간 충돌 조정 및 유럽 내 상호 운용 프레임워크 원칙 준수를 위한 유럽데이터혁신이사회 설립

9) <https://cloud.google.com/learn/what-is-data-governance?hl=ko>

10) Data Economy: Global News Trends in EU. Vol.2. No.4. 2021.4. 한국데이터산업진흥원.

2021년 4월 세계은행이 발간한 글로벌 데이터 규제 진단 보고서(World bank, Mapping Data Governance Legal Frameworks Around the World: Finding from the Global Data Regulation Diagnostic)에서는 이와 관련하여 국가별 데이터 경제 규제 환경을 측정하기 위한 객관적이고 표준화된 지표 개발에 대하여 논하면서, 데이터 경제의 행위주체 간 신뢰를 창출하고 개발을 촉진하기 위한 핵심지표로 보호조치(safeguard)와 촉진조치(enabler)를 소개하였다. 보호조치(safeguard)란 데이터의 오용이나 데이터 침해를 방지함으로써 데이터 경제에 참여하는 개인이나 단체의 권리를 보호하는 것을 목적으로 하는 규범 및 법적 프레임워크이고, 촉진조치(Enabler)란 데이터 이동성 메커니즘, 오픈데이터법 등 데이터의 사용 및 재사용을 촉진하는 규범과 법률을 총칭한다. 이러한 모든 노력은 데이터의 국가 간 이동 및 데이터 경제 주체 간 거래, 데이터 공급망 내 유통을 신뢰를 기반으로 원활하게 만들기 위해서이다.

유럽사법재판소(CJEU)는 미국에 저장된 EU 시민의 개인정보를 충분히 보호하지 못했다는 판단하에 EU에서 미국으로의 개인정보 이동을 규율하는 프레임워크인 EU-미국 프라이버시 실드를 2020년 7월 무효화하였다. 이어서 유럽개인 정보보호이사회(EDPB) 역시 유럽-미국 간 개인정보 이동을 수반하는 활동을 지양하도록 강력하게 권고하는 지침을 2020년 10월 발표하였다. 2020년 11월 12일 유럽연합 집행위원회가 유럽연합 역외로의 데이터 전송을 위한 표준계약서(SCC, Standard Contractual Clauses) 개정 초안을 발표하였다. 올해 1월 유럽데이터보호이사회(EDPB, European Data Protection Board)의 자문을 받은 이 계약서는 유럽연합과 유럽경제지역(EEA, European Economic Area), 유럽연합에 속하지 않은 제3국으로 구분되어 적용 예정이며, 특히 제3국을 대상으로 하는 내용이 주를 이루고 있다. 이 계약서에서는 제3국의 정부 단체 및 민간 기업과 데이터 교류 시 발생할 수 있는 사용자 데이터의 무단 수집 및 유출에 대한 규제가 포함되었다.

2021년 6월 4일 개정안이 정식 발표되었고 27일 공식적으로 발효되었다. 이에 따르면 새로운 계약 체결 시 기존 표준계약조항 사용의 마지막 허용 일자는 3개월 이후인 2021년 9월 27일이고, 기존 표준계약조항의 개정은 2022년 12월 27일까지 완료해야 한다. 미국의 경우 유럽 간 기존 데이터 전송 협의인 프라이버시 실드를 활용해왔으나 작년 CJEU에 의해 무효화되면서 기존 표준계약서가 적용되고 있는 상황이다. 표준계약서 개정안이 적용되면 단순한 요구나 요청으로는 데이터 이동이 금지되며, 법원 영장이나 기타 사법 증거가 제시되는 경우에만 데이터를 이동할 수 있게 된다.

데이터 경제로의 이행이 세계적인 추세가 되어 그 영향력이 확장될수록 데이터 거버넌스에 대한

요구는 지속적으로 증가할 것이다. 이러한 맥락에서 데이터 공급망 관리 체계를 구축함에 있어서 데이터 거버넌스를 준수하는 것은 권장할 사항이 아니라 필수적이라고 할 수 있다.

데이터 거버넌스를 준수하고 데이터의 품질을 표준에 맞게 메타데이터에 기술한 데이터는 시장에서 공개적으로 거래될 수 있는 자격을 비로소 갖추었다고 볼 수 있다. 한국데이터산업진흥원에서는 2020년 1권 데이터가격책정, 2권 데이터품질평가, 3권 데이터법률검토에 대한 종합안내서(이하 안내서)를 발간하여 데이터를 상품화하기 위한 표준을 마련하였다. 1권 데이터가격책정 종합안내서에 따르면, 데이터란 관찰이나 측정값 등의 원천 데이터(raw data), 가공 데이터 및 이를 체계적으로 생산, 수집, 축적한 데이터베이스를 의미하며, 데이터 거래는 공급자(판매자)와 수요자(구매자) 사이에 온/오프라인 방식으로 데이터를 전송/사용/이전하는 행위를 의미한다. 그리고 원칙적으로 정보 주체의 동의를 얻지 못한 개인정보가 포함된 데이터나 국가 안보, 산업기술 관련 중요 정보가 포함된 데이터, 제3자의 명예를 훼손하는 정보가 포함된 데이터, 이용허락을 받지 않은 제3자의 권리가 포함된 데이터 등 데이터 거버넌스 관점에서 적절치 않은 데이터는 거래가 제한될 수 있음을 명시하고 있다.

또한 안내서에서는 데이터의 가치를 “데이터를 활용해 미래에 기대하는 편익을 현재의 값으로 환산한 것”이라고 정의하고, 데이터의 가격은 “데이터의 가치를 투입된 비용에 의해 화폐 단위 가격으로 환산한 것”으로 정의하였다. 데이터의 가치와 가격이 비례하지 않을 수 있음을 명시하고, 1) 데이터 가격 결정의 목표 설정, 2) 수요 예측, 3) 원가 측정, 4) 경쟁환경 분석, 5) 가격책정방법 선택, 6) 데이터 제공방식 결정, 7) 데이터 과금방식 및 할인정책 결정까지 일곱 가지 단계로 데이터 가격 결정 절차를 설명하였다. 특히 가격책정방법에 있어서 원가 기준, 경쟁사 기준, 가치 기준의 세 가지 방식 중에서 소비자(구매자) 관점에서 합리적인 가격 여부를 객관적으로 판단할 수 있도록 가치 기준 가격책정 방법에 초점을 맞추어서 설명을 제공하고 있다.

〈표 5〉 가격책정방법별 장단점 비교¹¹⁾

가격책정방법	장점	단점
원가 기준	- 원가 중심으로 가격을 책정하므로 마진 폭을 쉽게 설정 가능	- 무형가치를 반영하지 않아 큰 마진을 얻을 기회를 놓칠 가능성
경쟁사 기준	- 경쟁 업체에 대한 가격 경쟁력 확보	- 지속적인 경쟁력 확보가 어렵고 가격 변동이 심할 가능성 높음
가치 기준	- 고객 중심으로 가격을 책정, 판매 가능성이 높은 가격으로 접근	- 고객별로 가치는 주관적이므로 가격의 일관성을 확보하기 어려움

가치 기준 가격책정을 위해서는 공급자 관점의 데이터 품질과 함께 수요자 관점의 가치 평가 기준이 추가적으로 설정되어야 한다. 안내서에서는 네 가지 차원을 제시하고 있다.

- 콘텐츠 적합성: 데이터가 내포한 정보 즉, 데이터의 콘텐츠에 대한 가치. 동일한 주제라도 좀 더 다양하고 깊이 있는 정보를 제공하는 데이터의 활용가치가 높음
- 품질 및 공급 신뢰성: 데이터의 물리적인 품질과 데이터 공급에 대한 가치. 데이터의 오류가 적고 향후에도 지속적으로 공급되는 데이터의 가치가 높음
- 기술적 사용성: 구매한 데이터 이용 시 예상되는 기술적 제약사항에 대한 가치. 이용이 편리하고 특정 소프트웨어에 의존적이지 않은 데이터의 가치가 높음
- 경제성: 구매한 데이터 활용의 제도적 제약사항과 대안 가능성에 대한 가치. 활용의 제약이 없고 대안이 적을수록 가치가 높음

데이터 공급망 관점에서 데이터 수집과 라벨링, 그리고 이 과정에서의 데이터 거버넌스를 구축하고 준수하는 것은 가장 중요한 작업이자 비용과 시간을 소비하는 작업이다. 그런 만큼 학계와 업계 전반에 걸쳐 많은 연구가 진행되고 있으며, 데이터 경제 체계의 효율성과 효과성을 높이기 위한 경제 주체들 간 신뢰를 쌓아나가기 위해 노력하고 있다. 최근 주요국들의 데이터 경제 정책과 기업 사례들은 이러한 기초를 잘 보여준다.

11) 성태웅, 변정은, & 박현우. (2016). 데이터베이스 자산 가치평가 모형과 수명주기 결정. 한국콘텐츠학회논문지, 16(3), 676-693.

V. 관련 국내외 동향: 정책과 사례

1 EU

주요 정책

유럽연합은 2014년 이후 본격적으로 데이터 경제 정책을 수립 및 추진하기 시작하였다. 데이터 단일시장을 위해 역내 데이터의 자유로운 흐름과 활용은 촉진하고, 역외 국가와 기업들의 자국 데이터 활용은 엄격하게 규제하는 역내와 역외에 차별적으로 규정을 적용하여 정보 보호와 활용에 대한 노력을 병행하고 있다.

〈표 6〉 2014년 이후 EU 데이터 경제 정책 기초¹²⁾

정책	시기	주요 내용
데이터 주도 경제 결의안 채택	'14.7.	- 유럽집행위원회, 유럽의 경제개발에 데이터 잠재력을 활용할 수 있도록 유럽 회원국의 정책 개발에 “데이터 주도 경제” 결의안을 채택 - 데이터 주도 경제의 특성 및 유럽이 데이터 경제를 주도하기 위한 초기 조치 등을 설명
유럽 디지털 단일시장 전략	'15.5.	- 미국과 중국에 대응한 디지털 경쟁력 제고를 위해 역내 디지털 경제 활동 제약을 제거하고 하나 된 유럽 디지털 시장을 목표로 하는 전략 발표 - 데이터 경제 구축을 목표로 역내 자유로운 데이터 이동 촉진을 위한 유럽 데이터 이니셔티브 및 유럽 클라우드 이니셔티브 제시
유럽 데이터 경제 육성	'17.1.	- EU 내 통합 디지털 플랫폼(Digital European)을 기반으로 데이터 접근, 분석, 활용 강화를 위한 데이터 신사업 창출을 목표로 함 - 보호 강화와 합법적 데이터 유통을 동시에 추구 - 데이터 유형별 정책을 세부적으로 제안
일반정보보호규정 (GDPR)	'18.5.	- 데이터 삭제권, 정보이동권, 프로파일링에 대한 권리와 가명정보 등을 법적으로 규정하여 사용자의 신뢰를 기반으로 하는 데이터 활용과 EU 역내 자유로운 데이터 흐름을 촉진
유럽 데이터 전략	'20.2.	- EU가 데이터 애자일 경제(Data-Agile Economy)의 선두에 서기 위해 취할 수 있는 향후 5년간 EU 데이터 경제의 정책 조치 및 투자 전략을 제시

12) EU Digital Special Report 2편. 데이터 경제의 떠오르는 이슈. 2021. 4. 한국지능정보사회진흥원.

최근 유럽연합이사회는 2021년 1월 5일 전자 프라이버시 규정(ePrivacy Regulation) 개정안 초안을 발표하고, 2월 10일 회원국들의 동의를 얻어 정식으로 발표함으로써 규정의 범위를 공용네트워크 상 M2M(Machine to Machine)까지 확대하였다. 즉 사물인터넷 사용 시 전송되는 데이터도 보호 범위에 포함시킴으로써 데이터 거버넌스의 필수 적용 범위를 한층 더 확대한 상황이다.

최근 이슈 및 사례

유럽연합의 이러한 정책 기조는 빅테크 기업들에 대한 규제 강화로 자연스럽게 연결된다. 이로 인해 빅테크 기업들은 공정 거래, 개인정보 보호 등 데이터 거버넌스 측면에서 집중적으로 제재를 받고 있다. 미국의 아마존, 페이스북, 트위터, 구글, 애플 및 중국의 틱톡까지 사용자 데이터의 무단 수집 및 활용과 관련하여 소송 중이거나 과징금을 부과 받고 있는 상황이다. 데이터 수집 단계에서부터 한층 강화된 데이터 거버넌스가 존재해야 이러한 정책 기조하에서도 원활하게 데이터 경제 주체로서 기업활동을 이어나갈 수 있다.

〈표 7〉 유럽연합 내 빅테크 기업 제재 현황¹³⁾

기업	혐의	제재현황
아마존(미국)	- 소비자 데이터 기반 불공정 경쟁(경쟁업체 방해 혐의)	- 2020년 11월 유럽연합집행위원회가 기소, 최대 연 매출 10%에 달하는 벌금 부과 가능
페이스북(미국)	- 데이터 처리 관행 수정 명령 미준수 - 반복적인 데이터법 위반	- 2021년 2월 이탈리아 경쟁당국이 700만 유로 벌금 부과
틱톡(중국)	- 유해 콘텐츠 제재 미흡 - 불법적인 데이터 처리	- 2021년 2월 16일 유럽 소비자기구가 유럽연합집행위원회에 조사 촉구
트위터(미국)	- 개인정보 위협요소에 대한 미적시 및 신고 누락	- 2020년 12월 아일랜드 데이터보호 위원회가 벌금 45만 유로 부과
구글(미국)	- 데이터 무단 활용 - 디스플레이 광고 독점	- 2020년 10월 이탈리아 경쟁 당국이 관련 혐의 조사 중

13) Data Economy: Global News Trends in EU. Vol.2. No.4. 2021.4. 한국데이터산업진흥원.

2 미국

주요 정책

미국 바이든 행정부는 데이터산업 정책과 관련하여 연방 개인정보 보호법 통과에 초점을 맞추고 있으며, EU의 프라이버시 실드 무효화 행정 명령에 대한 재검토 혹은 수정의 가능성을 타진하며 데이터의 대서양 횡단 이동 프레임워크에 대한 재합의를 준비 중이다. 데이터 활용에 관한 정책이 지속적으로 제정되는 상황에서 2021년 3월 2일 버지니아주에서는 소비자데이터보호법(Consumer Data Protection Act, CDPA)이 통과되었다. EU의 GDPR을 준용한 이 법은 2023년부터 효력을 발휘할 예정이다. 미국의 데이터산업의 특성은 EU와 대조적으로 민간 주도이고 성장이 견인되어 왔으나, 데이터 경제의 규모가 지속적으로 확대되면서 국내외에서 거버넌스에 대한 요구가 증가하고 있는 것에 정책적으로 대응하게 되었음을 알 수 있다.

〈표 8〉 미국 바이든 행정부 정책 기초¹⁴⁾

정책 기초	내용
빅테크 규제 도입, 디지털 시장 질서 개편	- 빅테크 규제와 반독점 기초 유지
플랫폼 책임성 강화 콘텐츠 규율 및 플랫폼 과세	- 통신품위유지법 230조 개정을 통한 책임 강화
인터넷 접근성 확대 통신망 고도화와 망 중립성 강화	- 낙후지역 초고속 통신망 보급 확대 - 망 중립성 강화
중소창업기업 및 소외분야 지원을 통한 신속한 경제 회복	- COVID-19로 인한 경제위기 극복 위한 긴급 지원 - 직접 보조금 지급 및 창업 지원 확대
긱(Gig) 경제 안정성 보장 긱 노동자 보호 조치 확대	- 긱(Gig) 경제 종사 노동자들에 대한 보호 조치 확대
미국 중심의 GVC 재편과 중국 배제 기초 지속	- 미국 중심, 동맹국 협력 형태로 글로벌 가치사슬 개편 - 리쇼어링 인센티브와 오프쇼어링 페널티 실시 - 동맹국 중심 공급망 형성으로 중국 의존 감소
다자주의 회귀와 미국식 규제 확산	- 미국식 통상규칙 글로벌화를 위한 다자협상 중시

14) Data Economy: Global News Trends in USA. Vol.2. No.1. 2021.1. 한국데이터산업진흥원.

최근 이슈 및 사례

미국과 중국, 유럽 간 데이터 분쟁은 계속해서 이어지고 있다. 특히 지난해 중국 베이징 소재 업체인 바이트댄스가 서비스하고 있는 동영상 공유 애플리케이션 틱톡(TikTok)이 사용자 데이터 불법 수집 논란에 휩싸이면서 국가 간 분쟁으로 확대되었다. 미 상무부는 결국 국가 안보에 대한 위협을 이유로 틱톡과 위챗을 미국 내 앱스토어에서 제거하는 결정을 내렸고, 2020년 11월 12일부터 사실상 틱톡 사용 금지 행정명령을 발효하였다.

국가 간 경쟁이 치열해짐과 함께, 글로벌 데이터 경제에서 가장 큰 부분을 미국이 차지하고 있는 만큼 두드러지는 데이터 관련 사고사례가 미국에서 주로 발생하고 있다. 미국의 대표적인 데이터 기업들인 아마존, IBM, 구글 등은 안면 인식 데이터베이스 구축에서 개인정보 침해와 관련된 소송을 당했고, 페이스북은 데이터를 ‘무기화’하여 인스타그램이나 왓츠앱과 같은 잠재적 경쟁사들을 인수한다며 반독점 소송을 당하고 있다. 또한 트위터, 링크드인, 페이스북 등 빅테크 기업의 데이터 유출 피해 사례도 늘어나고 있는 실정이다. 이러한 사고사례는 데이터 공급망 관리 측면에서 공들여 쌓아왔던 공급망 내 이해관계자들 간 신뢰를 크게 무너뜨릴 수 있으므로 개별적인 정보 주체에 끼칠 수 있는 경제적, 심리적 피해와 함께 데이터 공급망의 붕괴를 막기 위한 주의가 필요하다.

〈표 9〉 미국 내 데이터 기업 데이터 유출 사례¹⁵⁾

기업명	피해 규모	주요 내용
페이스북	5억 3,300만 건	- 이름, 연락처, 위치, 이메일 등 개인정보 - 106개국 사용자 정보 유출
링크드인	5억 건	- 이름, 연락처, 아이디 등 개인정보
아스토리아	3,000만 건	- 2021년 1월 26일 피해 사실 확인 - 이름, 연락처, 위치, 이메일 등 개인정보 - 사회보장번호, 은행계좌, 의료기록 등 민감성 정보
파크모바일	2,100만 건	- 2021년 3월 피해 사실 확인 - 이메일 주소, 연락처, 차량 번호, 주소 등
클리어보이스	1,570만 건	- 2021년 4월 17일 피해 사실 확인 - 2015년 이후의 모든 개인정보 - 고객 설문 자료

15) Data Economy: Global News Trends in USA. Vol.2. No.7. 2021.7. 한국데이터산업진흥원.

3 중국

주요 정책

중국은 2017년 6월부터 인터넷 주권과 국가 안전을 명목으로 사이버보안법을 도입하는 등, 중국 내에서는 데이터 공개를 통한 공유를 활발히 할 수 있는 환경을 조성하지만 반대로 국외로 유출은 어렵게 하는 일종의 블록화 전략을 추진하고 있다¹⁶⁾. 2020년에는 데이터를 새로운 생산요소에 포함시키는 생산요소 시장화 배치 메커니즘 구축에 관한 의견을 발표하였다. 2021년 3월 중국 정부는 전국인민대표대회에서 ‘14차 5개년 계획’의 초안을 발표하였다. 이 계획에서 중국 정부는 국가 빅데이터 전략 추진이라는 목표 아래 국가 경제 사회 발전의 핵심 지표 중 하나인 빅데이터 산업의 중요성을 다시 한번 강조하였다. 특히 중국의 대표적인 데이터 산업 정책 기구인 공업정보화부는 이 계획 하에서 데이터 발전 산업 발전 및 데이터 품질 혁신, 데이터 생태계 구축, 데이터 보안 등 4대 주요 목표를 기반으로 한 정책 수립을 계획하고 주요 정책 초안을 작성 중이다.

〈표 10〉 2015년 이후 중국 데이터 경제 정책 기조¹⁷⁾

정책	시기	주요 내용
빅데이터 발전 촉진 행동 요강	'15.8.	<ul style="list-style-type: none"> - 중국 빅데이터 표준 체계 마련, 빅데이터 관련 기초표준, 기술표준, 응용표준, 관리표준 수립 - 정부 정보 수집, 보관, 공개, 공유, 사용, 보안에 관한 기술 표준 수립 - 데이터 표준 검증 및 애플리케이션 시범사업 추진 - 국제 표준 수립 과정에 적극 참여 제안
빅데이터산업발전 계획 2016~20	'17.1.	<ul style="list-style-type: none"> - 빅데이터 산업 발전을 위한 지도이념, 개발목표, 핵심과제, 주요 프로젝트 및 보호 조치 명시 - 빅데이터 기술 개발과 제품 혁신 촉진, 산업 응용 역량 향상 및 생태계 번영 촉진, 산업 지원 시스템 개선, 빅데이터 보증 시스템 통합 및 개선
생산요소 시장화 배치 메커니즘 구축에 관한 의견	'20.4.	<ul style="list-style-type: none"> - 데이터를 새로운 생산요소에 포함시켜 데이터 요소시장을 육성 - 빅데이터가 경제의 고품질 성장을 이끄는 새로운 역동적인 에너지로 작용하도록 유도

16) 포스트 코로나 시대의 미-중 AI 패권경쟁을 바라보는 관전 포인트. 2020. 5. 소프트웨어정책연구소.

17) 데이터산업 동향 이슈 브리프. 중국 국내외 빅데이터 정책과 표준화 동향. 2020. 11. 한국데이터산업진흥원.

최근 이슈 및 사례

중국 역시 미-중-유럽 간 데이터 경쟁에 대응하여 자국 인터넷 기업들을 대상으로 반독점 행위 규제를 강화하고, 데이터 보안 국제 이니셔티브를 발표하여 미국의 클린네트워크 프로그램에 의한 중국 앱 차단에 대응하고 있다. 2021년 6월에는 전국인민대표대회에서 ‘중화인민공화국 데이터보안법’을 발표하여 데이터 보안 및 개발, 데이터 보안 시스템, 데이터 보안 보호 의무 및 법적 책임을 규정하였다. 이러한 중앙정부의 강력한 의지를 토대로 중국의 빅데이터 기업은 빠르게 성장하고 있으며, 한국무역협회에서 발간한 보고서¹⁸⁾에 따르면 현재 중국의 빅데이터 총량은 전 세계의 20% 정도로, 2025년에는 3분의 1 수준으로 증가할 것으로 예상하였다.

특히 최근 들어 중국 금융권에서 데이터 수집 및 활용을 위한 기술 개발이 가속화되고 있으며, 주요 은행들은 데이터 기술을 활용하여 대고객 서비스 및 신규 금융상품 개발을 진행하고 있다.

〈표 11〉 중국 금융권 데이터 수집 및 활용 현황¹⁹⁾

기업명	내용
중국중신은행	- 빅데이터 활용 실시간 마케팅
중국광대은행	- 소셜 미디어 내 고객 행동 데이터 수집 및 데이터베이스 구축 - 은행 내·외부 데이터 통합을 통한 고객 데이터 정확성 향상
중국초상은행	- 빅데이터 분석 기반 소액 대출 서비스 출시 - 고객 거래 기록 기반 금융상품 추천 - 고객 맞춤형 서비스 개발을 통한 신규 고객 확보 및 고객 유출 방지
중국공상은행	- 현장 및 지점 내 빅데이터 응용 프로그램 적용, 고객자산 관리 - 빅데이터 기반 지능형 리스크 관리 시스템 구축
알리바바 파이낸스	- 전자상거래 플랫폼과 신용 서비스 데이터 결합 - 전자상거래 데이터를 활용, 무담보 대출 서비스 개발 및 출시
중국건설은행	- 빅데이터 분석 기반 고객 세분화 마케팅 - 빅데이터 기술 활용 신용 평가 분석
상하이농촌상업은행	- 텐센트 클라우드 빅데이터 제품 기반 플랫폼 구축 - 신용, 금융 중개, 사기 방지, 지불 및 모바일뱅킹에 활용할 수 있는 통합 금융 빅데이터 구축

18) 중국의 빅데이터 시장 트렌드와 시사점. 2020.11. 한국무역협회.

19) Data Economy: Global News Trends in China. Vol.2. No.8. 2021.8. 한국데이터산업진흥원.

4 한국

주요 정책

2019년 12월 시작된 COVID-19로 인해 디지털 전환과 데이터 경제로의 이행은 보다 가속화되었다. 한국 정부는 2020년 7월 14일 ‘한국판 뉴딜’을 발표하면서 포스트 코로나 시대의 경기회복을 위한 국가 프로젝트를 운영하고 있다. 한국판 뉴딜 정책 내 디지털 뉴딜, 특히 데이터댐 과제는 데이터의 수집 및 활용을 목적으로 하는, 데이터 공급망 관리에 가장 관련이 깊은 정부 과제이다. 과학기술정보통신부 보도자료²⁰⁾에 따르면, 2020년 본예산과 추경을 통해 6,449억 원을 투입하여 산업계에서 부족한 양질의 데이터 생산 및 개방을 위해 빅데이터 플랫폼과 센터를 확대 구축하고, 인공지능 서비스 개발에 필수적인 인공지능 학습용 데이터를 대규모로 구축 및 개방 중이다. 2020년 11월 누적 기준 인공지능 학습용 데이터 21종 4,650만 건을 구축·개방하였다. 2021년 올해에는 4월 16일과 6월 4일 인공지능 학습용 데이터 구축 지원사업 1·2차 공모가 완료되었다. 사업 공고문을 살펴보면 데이터 구축 및 공개 정책, 데이터 품질에 관한 사항에 한국데이터산업진흥원의 데이터 품질관리 가이드라인을 참조하게 되어 있으며, 품질관리에 대한 총괄 책임자를 명시함으로써 데이터 소유권을 명확히 하고 있음을 알 수 있다. <표 12>에서 인공지능 학습용 데이터 구축 참여 사례를 통해 국내 클라우드소싱의 주요 내용과 현황을 확인할 수 있다.

<표 12> 데이터댐 사업을 통한 인공지능 학습용 데이터 구축 참여 사례²¹⁾

구분	주요 내용
경력단절여성	<ul style="list-style-type: none"> - 결혼과 동시에 회사를 그만두고 지방으로 이주하여 재택근무가 가능한 데이터 일자리를 통해 다시 업무경력을 이어나가게 됨 - COVID-19로 인해 재취업에 어려움을 겪다가 클라우드소싱 기업이 제공하는 데이터 라벨링 교육에 참여한 후 자유로운 출퇴근으로 가정경제도 회복하고 자녀양육이 가능하여 만족도가 매우 높음
취업준비청년	<ul style="list-style-type: none"> - 학원비 및 생활비 등을 충당하는 아르바이트 매장이 COVID-19 확산으로 매장이 문을 닫게 되어 한국어-영어 번역문장을 올리는 작업에 틈틈이 참여하여 경력도 쌓고 생활비 걱정 없이 취업 준비에 매진

20) 디지털뉴딜의 핵심 축인 데이터 댐 사업 성과보고회 개최. 2020.12.16. 과학기술정보통신부.

21) 국민과 함께 채운 데이터댐, 본격 개방. 2021.6.18. 과학기술정보통신부 정보통신정책실.

구분	주요 내용
실직자	- COVID-19의 여파로 여행사가 폐업하여 실직하게 되었으나, 인공지능 학습용 데이터 구축 참여를 통해 생활에 도움이 되는 새로운 일자리를 얻을 수 있었음
장애인	- 자폐성 장애를 극복하고 사회적 기업에 정식직원으로 채용되어 자율주행차와 관련된 표지판과 신호등 라벨링 작업을 수행
중장년층	- 두 딸을 시집보내고 외로움과 불면증이 있었으나 데이터 라벨링 작업이 적성에 맞아 전업 크라우드워커로 활동하면서 제2의 인생 시작 - 과거 국회 DB 전산화 작업 경력을 살려 데이터 가공 관리자로 재취업에 성공하여 이전 직장보다 더 좋은 환경에서 일할 수 있게 됨
소상공인	- COVID-19 여파로 가게 운영이 어려워졌으나, 가게에서 틈틈이 데이터 가공 작업에 참여하며 수입을 보충
정식직원 성장	- 아르바이트에서 시작한 데이터 라벨링에서 능력을 인정받아 한 달 만에 정식 직원으로 채용되어 크라우드워커들을 교육하고 가이드라인을 만들며 작업한 결과물을 검수하는 등 다양한 업무를 수행 - 여성새로일하기센터에서 라벨러 교육을 받으며 MOU를 체결한 기업 프로젝트에 참여하다가 정식 직원으로 채용 - 정부의 인공지능 학습용 데이터의 다양한 프로젝트에 참여했다가 빠른 시간 안에 전문성을 인정받아 정식 직원으로 채용
전문직 경력 발전	- 지적측량과 항공정비 등 2개 분야의 자격증을 보유하여 드론을 활용한 지적측량 데이터 구축사업에 계약직으로 참여하다가 정규직으로 전환

최근 이슈 및 사례

한국지능정보사회진흥원에서는 데이터댐 사업에서 구축한 인공지능 학습용 데이터를 활용하여 서비스 개발에 성공한 사례를 모아서 2020년 5월 ‘인공지능 학습용 데이터 활용 우수 사례’ 보고서를 발간하였다. NAVER Clova AI, 삼성전자 등 대기업, 스켈터랩스, 인라이플, 로민 등 중견/중소기업, 대학교, 대학생 동아리 등 다양한 집단이 학습데이터들을 사용하고 있다. 현재까지 공개되어 활용할 수 있는 인공지능 학습용 데이터는 8개 분야 170종, 4억 8,000만 건으로 한 달 만에 1만 2,000건이 넘는 다운로드 수를 기록했다. 과학기술정보통신부는 2021년 말까지 190종의 데이터를 추가 구축하기 위해 총 545개 기업과 기관이 함께 사업을 진행할 계획이다.

〈표 13〉 인공지능 학습용 데이터 주요 내용²²⁾

구분	주요 내용
음성·자연어	<ul style="list-style-type: none"> - 대화·명령어·방언 발화 음성, 한국어-외국어 말뭉치 등 39종 - 사람의 언어와 음성을 인식하고, 대화에 내포된 의미와 맥락을 정확히 이해하고 대응할 수 있도록 하는 다양한 한국어 데이터 확보
헬스케어	<ul style="list-style-type: none"> - 암, 뇌질환, 피부, 치과 등 다양한 의료 데이터 32종 - 건강관리, 질병 검진에서부터 예방·예측, 치료에 이르는 전 과정 스마트 의료 혁신을 뒷받침할 임상 중심의 의료영상 데이터 확보
자율주행	<ul style="list-style-type: none"> - 주행영상, 객체·장애물 이미지 등 자율주행 데이터 21종 - 자율주행차, 드론 등 지능형 모빌리티 산업의 핵심 기반이 되는 국내 실제 도로 기반의 대규모 자율주행용 데이터 확보
비전	<ul style="list-style-type: none"> - 방송·광고 영상 및 스포츠 동작 영상 등 15종 - 시각 지능기술의 혁신을 통해 인간 수준으로 사물을 인지하고, 다양한 상황을 판단할 수 있는 밑바탕이 되는 이미지·영상 데이터 확보
국토환경	<ul style="list-style-type: none"> - 토지·산림 위성 이미지, 수질오염 이미지 등 12종 - 국내 생태계 보호, 환경오염 방지활동의 지능화 혁신이 가능하도록 항공·위성사진, 오염원·폐기물 관련 데이터 확보
농축수산	<ul style="list-style-type: none"> - 국내 주요 작물, 가축, 어류의 영상 및 질병 데이터 14종 - 1차 산업인 농·축·수산업의 디지털 전환, 스마트화의 실현에 필요한 작물, 가축, 양식 어류 등에 대한 인공지능 학습용 데이터 확보
안전	<ul style="list-style-type: none"> - CCTV 영상, 이상행동, 교통흐름·재난상황 데이터 등 19종 - 지속적으로 발생하는 재난 및 안전, 보안 사고에 대한 선제적 대응 및 지능화 대응 체계의 구축과 고도화에 필수적인 데이터 확보
기타	<ul style="list-style-type: none"> - 소상공인-고객 질의응답, 패션상품 및 착용영상 데이터 등 18종 - 패션, 교육, 반려동물 등 다양한 분야의 인공지능 기반 혁신서비스 확산을 위한 인공지능 학습용 데이터 확보

22) 국민과 함께 채운 데이터댐, 본격 개방. 2021.6.18. 과학기술정보통신부 정보통신정책실.

5 인도

주요 정책

인도 국가개혁위원회(NITI Aayog)는 2018년 6월 AI 국가정책에 대한 제언에서 국가 AI 마켓플레이스(NAIM: National AI Marketplace)를 구축하여 수요와 공급 간 선순환을 이루어야 한다고 강조하였다. 또한 사회적 요구에 부합하면서 가장 효과적인 AI 기술 분야로 헬스케어, 농업, 교육, 스마트시티, 스마트 모빌리티까지 5개 산업군을 선정하였다²³⁾. 2020년 8월 모디 총리는 국가디지털의료미션(National Digital Health Mission)의 개시를 발표하면서 디지털 의료 데이터베이스와 국제 표준에 맞는 의료 기록 시스템의 구축을 미션의 핵심 목표로 설정하였다. 인도 정부는 이러한 ‘디지털 의료’ 서비스 출범을 준비하면서 데이터 프라이버시에 대한 법제화 역시 동시에 추진 중이다²⁴⁾. 인도 전자통신기술부는 2017년 7월부터 개인정보보호 관련 논의를 위한 위원회를 구성하고 2018년 7월 개인정보보호법안의 초안을 의회에 제출, 2019년 12월 수정안이 의회에 상정되었다. 인도 보건부는 개인정보보호를 위한 전반적인 프레임과 최소한의 개인정보보호 표준 내용을 담은 데이터 관리 정책 초안을 발표하였다. 초안에 따르면 모든 데이터는 개인의 동의하에만 수집될 수 있으며, 언제든지 개인은 이 동의를 취소하여 데이터의 공유를 제한할 수 있다.

최근 이슈 및 사례

인도 역시 중국과의 갈등이 격화되고 있다. 2020년 6월 히말라야 인근 국경에서 인도-중국 양측의 무력 충돌이 일어나면서 중국 앱에 대한 사용자 데이터 외부 전송 우려에 대한 대응으로 틱톡을 포함한 58개 중국 앱을 차단하였다. 이어, 2020년 9월 국가 안보 및 방위에 적대적인 세력에 대한 개인정보의 편집 및 프로파일링 행위로 간주된 118개 모바일 앱 차단을 결정하면서 Baidu, Alipay, WeChat 등의 중국 앱에 대한 차단이 이루어졌다.

23) National Strategy for Artificial Intelligence #AIFORALL. 2018. 6. NITI Aayog.

24) Data Economy: Global News Trends in India. Vol.1. No.6. 2020. 12. 한국데이터산업진흥원.

VI. 결론

데이터·기계학습·인공지능 기술 기반 제품 및 서비스 산업은 ICT 기술의 발전과 모바일 환경의 일반화, 그리고 COVID-19 이후 비대면 기술에 대한 수요 증가에 힘입어 급격하게 성장하고 있다. 4차 산업혁명이라고 불리는 디지털 전환과 데이터 경제로의 이행은 이제 멈출 수 없는 대세가 되면서 데이터는 전통적인 제조업과 서비스업 운영을 지원하는 역할에서 벗어나 데이터 그 자체가 중간재이자 최종재로서 기능할 수 있는 거래 가능한 상품의 모습을 갖추어가고 있다. 원천 데이터의 수집 및 가공, 데이터의 안전한 거래와 유통을 위한 데이터 거버넌스는 데이터 경제의 뼈대이며, 이를 제대로 관리하기 위해서는 데이터를 교환요소로 간주하는 데이터 공급망의 관점이 필요하다.

유럽연합, 미국, 중국 등 주요국은 국가적으로 데이터 경제의 선두에 서기 위해 전략을 수립하고 정책을 실행하고 있다. 현재로서는 실물 원자재와는 다른 데이터의 특성(낮은 배타성, 이질성, 권리와 책임의 모호성 등)으로 인해 보안과 보호에 중점을 두고 규제의 형식으로 신뢰를 쌓아가는 중이다. 현재의 경합이 정리되고 국가 간 신뢰적 데이터 유통이 가능해진다면 데이터 경제를 효율적으로 작동시킬 수 있는 시장 기능이 활성화될 수 있을 것이다. 규칙이 정의된 시장에서 거래되는 데이터는 좋은 품질을 가지고 있어야 하므로 품질을 측정할 수 있는 수단과 이를 표현할 수 있는 표준 언어의 개발 역시 빠르게 진행되고 있다.

우리나라 역시 이러한 흐름에 뒤지지 않고 세계적인 수준의 데이터 공급망을 구축하기 위해 노력을 경주하고 있다. 디지털 뉴딜 정책의 우선과제인 데이터댐 사업은 그 일례로써, 데이터 공급망 관점에서 원천 데이터를 수집하는 가장 중요한 단계에 대한 정책적인 관심이 높은 것은 고무적인 일이다. 데이터댐 사업과 함께 데이터 거버넌스를 고려한 데이터 품질 관리 기술을 개발하고, 나아가 글로벌 데이터 공급망 관리 체계를 구축한다면 국가 경쟁력의 한 축을 담당하게 될 데이터 산업의 기반을 보다 견고하게 다질 수 있을 것이다.

참고문헌

- 국민과 함께 채운 데이터댐, 본격 개방. 2021.6.18. 과학기술정보통신부 정보통신정책실.
- 디지털뉴딜의 핵심 축인 데이터 댐 사업 성과보고회 개최. 2020.12.16. 과학기술정보통신부.
- 성태응, 변정은, & 박현우. (2016). 데이터베이스 자산 가치평가 모형과 수명주기 결정. 한국콘텐츠학회논문지, 16(3), 676-693.
- 이지현·우창완. (2020). 데이터 라벨링으로 만드는 혁신, 이슈분석. 한국지능정보사회진흥원.
- 중국의 빅데이터 시장 트렌드와 시사점. 2020.11. 한국무역협회.
- EU Digital Special Report 2편. 데이터 경제의 떠오르는 이슈. 2021. 4. 한국지능정보사회진흥원.
- Data Economy: Global News Trends in India. Vol.1. No.6. 2020.12. 한국데이터산업진흥원.
- Data Economy: Global News Trends in EU. Vol.2. No.4. 2021.4. 한국데이터산업진흥원.
- Data Economy: Global News Trends in USA. Vol.2. No.1. 2021.1. 한국데이터산업진흥원.
- Data Economy: Global News Trends in USA. Vol.2. No.7. 2021.7. 한국데이터산업진흥원.
- Data Economy: Global News Trends in China. Vol.2. No.8. 2021.8. 한국데이터산업진흥원.
- National Strategy for Artificial Intelligence #AIFORALL. 2018. 6. NITI Aayog.
- Smartening up with artificial intelligence. 2017. 4. McKinsey & Company
- Spanaki, K., Gurguc, Z., Adams, R., & Mulligan, C. (2018). Data supply chain (DSC): research synthesis and future directions. International Journal of Production Research, 56(13), 4447-4466.
- Treder, M. (2020). The Data Supply Chain. In The Chief Data Officer Management Handbook (pp. 35-46). Apress, Berkeley, CA.
- Vial, G., Jiang, J., Giannelia, T., & Cameron, A. F. (2021). The Data Problem Stalling AI. MIT Sloan Management Review, 62(2), 47-53.

저자소개

| 배 금 일

산업및시스템공학 박사, (주)허그랩 기술총괄

| 이 현 규

감수, 정보통신기획평가원 PM(인공지능·데이터)

본 이슈리포트의 내용은 NRF의 공식적인 의견이 아닌 집필진의 견해이며 동 내용을 인용 시 출처를 밝혀야 합니다.

NRF ISSUE REPORT 2021_20호

인공지능 시대의 데이터 공급망 관리

| 발행일 | 2021년 10월 25일

| 발행인 | 이 광 복

| 발행처 | 한국연구재단

본 원 : 대전광역시 유성구 가정로 201

서울청사 : 서울특별시 서초구 현릉로 25

<http://www.nrf.re.kr>

| 편 집 | 정책연구실 정책혁신팀

ISSN 2586-1131