# 공공과 민간이 하나 되는 데이터 거버넌스 전략

한국과학기술정보연구원 선임연구원 박강희




위키피디아에 데이터 거버넌스(data governance)의 사전적 정의를 찾아보면 '통상적으로 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시, 보안성, 데이터 품질, 관리 규정 준수를 강조한다'라고 되어 있다. 상당히 난해한 글귀처럼 보이지만, 이를 다시 말하면 데이터 거버넌스는 데이터 관리를 의미한다. 데이터 관리에는 데이터 품질 관리뿐만 아니라 메타데이터 관리, 데이터 인벤토리, 데이터 수명주기 관리, 데이터 액세스 및 권한 부여, 데이터 통합 등의 관리 및 통제 기능들이 포함된다. 데이터의 정의부터 데이터 거버넌스까지, 미래의 데이터 관리 전략에 대해 알아본다.




데이터, 그리고 '데이터의 데이터'라 불리는 메타 데이터

글을 시작하면서, 가장 기본적인 데이터의 정의를 살펴보기로 한다. 데이터란 무엇인가? 사전적인 의미는 제쳐 두고, 일반적으로 많이 사용하는 엑셀 시트를 예시로 살펴본다. 가령 A라는 사람을 설명하는 자료를 엑셀 시트에 표현한다고 하자. A의 키, 몸무게, 학력, 금융 자산, 결혼 유무, 호감 정도, 친한 친구수, 취업한 직장 수 등을 하나의 행에 적을 수 있을 것이다. 이 엑셀시트 행 하나의 전체가 1개의 데이터다. 즉, A의 전체 행이 1개의 데이터라는 의미다. 그럼 데이터를 구성하는 키, 몸무게 등은 무엇일까? 이들은 A라는 데이터를 설명해주는 각각의 속성이다. 그러나 속성이 데이터는 아니다. 키, 몸무게 등 속성은 넓은 의미로 A를 설명하는 메타 데이터다. 사전적으로는 '데이터의 데이터'다. 최근에는 여러 속성 중 데이터를 가장 잘 설명해주는 속성으로 쓰이기도 한다. SNS의 해시태그가 그 예다.




수많은 데이터를 체계적으로 관리하는 데이터 거버넌스

다음으로 A사람의 형제인 B와 직장 동료인 C에 대한 데이터도 있다고 가정한다. A, B, C 사이, 서로 연관된 데이터가 있다면 이것이 바로 데이터 인벤토리다. A, B, C에 대한 데이터는 향후 바뀔 수 있으므로 시간 변동에 따라 업데이트하고 관리해야 한다. 이렇게 업데이트를 관리해주는 것이 데이터 수명주기 관리이다. 또 이 데이터를 어떤 사람들이 열람할지, 어떤 사람들이 관리할지 권한도 부여할 수 있는 분야가 데이터 액세스 및 권한 부여다. 이런 데이터가 다양하며, 여러 파일로 나뉘어 있다면 이를 통합하고 통제하는 일이 통제 기능이다.

위와 같은 이유로 데이터 관리에는 앞서 언급한 다양한 기능들이 필요하다. 그렇다면 왜 이런 작업들이 새롭게 부각되는 것일까? 인터넷과 모바일이 발전하면서 실시간으로 상상을 초월할 만큼 다양한 데이터들이 생산되고 있다. 공공 또는 민간에서는 이런 데이터를 활용해 인사이트를 얻고, 경쟁력을 높여 경쟁자들보다 좋은 정책 또는 높은 수익을 창출하고자 한다. 즉, 데이터를 원동력으로 활용하기를 원하고 있다. 데이터를 활용해 다양한 정책 또는 사업을 추진하려면, 그 재료가 되는 데이터들을 정제, 관리하는 일이 중요하다. 데이터 거버넌스 도입 중요성이 날로 부각되는 이유다.

## 공공기관의 데이터 산업 활성화 전략

| | |
|---|---|
| 데이터 이용 패러다임 전환 | ▪ 개인정보 비식별조치 법제화<br>▪ 정보 주체가 기관에서 직접 자기 정보 받아 활용 |
| 데이터 안심존 구축 | ▪ 데이터는 반출하지 않고 분석 및 AI 개발 결과만 반출 |
| 의료·금융·통신 분야 시범사업 | ▪ 건강 검진 결과 스마트폰으로 다운로드<br>▪ 계좌거래·카드구매내역 통합정보 공개<br>▪ 음성·데이터 사용량 공개해 맞춤형 요금제 추천 |
| AI데이터셋 구축 | ▪ 범용·전문 데이터 1억6,000만건 공급해 인공지능 지원 |
| 국가 데이터맵 | ▪ 공공데이터 전수 조사로 통합 관리체계 구축·개방 |
| 데이터 거래 기반 확보 | ▪ 민간과 공공 데이터 포털 연계한 플랫폼 개발 |
| 국가공인 데이터 자격제도 | ▪ 데이터 분석 국가기술자격제도 신설해 실무인력 양성 |

<출처: 4차산업혁명위원회 '데이터 산업 활성화 전략'>

경계 없이 서로의 데이터를 활용하는 공공과 민간

이미 데이터의 중요성은 공공과 민간 모두가 알고 있으며, 데이터 축적을 위해 노력하고 있다. 공공부문에서 정부 3.0은 데이터를 축적하고 공개한 대표적인 사례다. 서울시와 경기도에서 제공하는 버스 전용 애플리케이션도 축적된 데이터를 활용해 대국민 서비스를 했다. 이외에도 각 공공기관 사이트는 각 분야별, 연도별 통계를 공개했으며, CSV 파일 등의 형태로 다운로드해 활용 가능하도록 제공했다. 민간 역시 데이터를 기반으로 다양한 서비스를 제공 중이며, 관련 API를 공개해 데이터 수집 및 축적에 투자하고 있다.



<출처: 공공데이터 포탈 홈페이지>

공공과 민간의 데이터는 영역 간 경계가 허물어지는 중이다. 서로의 데이터를 활용해 새로운 사업을 하고, 관련 통계를 생산하려는 시도가 활발하다. 가장 보편적인 예는 금융권의 기업 신용평가 분야다. 정부는 기업 재무 등 다양한 데이터를 수집하고, 공공 또는 민간 기업은 이를 축적, 가공해 신용등급을 금융권에 제공한다. 이를 토대로 금융권은 자체적으로 신용평가를 한다. 정부 역시 민간의 다양한 SNS 데이터를 기반으로 공공정책에 귀를 기울이고 있다.

## 공공 민간 서로간의 데이터 결합 활용 사례

| 유형 | 영역 | 사례 |
|---|---|---|
| 정부 데이터를 민간이 활용 사례 | 금융 | ▪ 기업들의 재무제표를 활용하여, 공공 및 민간 기업이 신용평가를 제공 (한국기업데이터, NICE 등)<br>▪ 미국 정부의 주택 데이터베이스를 활용한 벤처기업(Zillow)과 영국 토지 등기소 (UK Land Registry)의 주택 거래 데이터에 기반한 벤처기업(Zoopla)은 온라인 부동산 장터 서비스 제공 |
| | 농업 | ▪ 미국 정부의 작물재배 현황 데이터와 날씨 정보, 토양 데이터를 활용하여 농부들에게 정보를 제공한 벤처기업(Climate Corp) 등장 (The World Bank, 2014) |
| 정부가 민간에게 데이터 협력을 요구 | 거시경제 | ▪ 중국 통계청은 2013년 11월 알리바바, 바이두 등11개 전자상거래업체와 빅데이터 물가 지수 개발 위한 파트너쉽 협약 체결<br>▪ 독일 통계청은 구인구직사이트(job portal) 데이터를 활용, 실업 통계 생산 |
| | 인구통계 | ▪ 벨기에 통계청과 Eurostat은 모바일폰 데이터와 상주인구통계를 비교 분석하여 인구, 관광, 교통, 이사 통계에 활용 |

<출처: KISDI의 「4차 산업혁명 시대의 데이터 거버넌스 개선 방향(2018.08)」 >

### 데이터 패러다임에도 변화가 필요하다

그렇다면 데이터 거버넌스의 주도권은 누가 쥐고 있을까? 점차 공공에서 점차 민간으로 넘어가는 추세다. 구글, 아마존, 페이스북 등 글로벌 기업들은 다양한 검색 플랫폼을 제공하고 있다. 이들은 각 서비스에서 발생하는 다양한 데이터를 다시 수집, 관리하는 방법을 취한다. 네이버, 카카오 등 국내 대표 기업들도 같은 방법을 활용해 데이터 거버넌스를 점차 확장시키는 추세다. 그러나 민간 기업은 이익 창출을 위해 데이터를 편향적으로 활용할 가능성도 존재한다. 광범위한 민간 데이터를 공공재 관점에서 국가통계 생산에 활용하려면 제도화 방안이 필요하다. 더불어 민간 데이터와 정부 데이터를 포괄하는 데이터 거버넌스 체제로 패러다임을 변화시켜야 한다.

데이터 시대에 부합하는 데이터 거버넌스 체계를 강화하려면 어떻게 해야 할까? 국가 핵심 데이터 자원인 국가 통계 거버넌스를 강화해야 한다. 이를 기반으로 데이터와 통계 생산 시스템을 유기적으로 결합하면서, 민간 데이터를 활용을 위해 단계적으로 거버넌스 체계 정립을 고민할 필요가 있다.

> **데이터 거버넌스가 발전**하기 위해서는 **인사이트를 도출하고 이를 다시 데이터화하는 것이 매우 중요**하다.
> 의미 있는 데이터를 잘 축적하고 활용하여 정확한 예측을 할 수 있는 기술에 대한 활용 및 발전도 같이 수반되어야 한다.

## 데이터가 선순환하는 데이터 거버넌스 체계

데이터 거버넌스가 선순환하려면 데이터를 활용해 의미 있는 인사이트를 도출하고 이를 다시 데이터화해야 한다. 이를 위해 확고한 목적에 따라 의미 있는 데이터를 잘 축적해야 한다. 정확히 예측할 수 있는 기술을 활용해 발전시키는 일도 중요하다. 예를 들어 기업의 가치를 추정한다고 하면, 중요한 사항들이 있다. 첫째, 관련 산업의 성장성이다. 둘째, 확장 가능한 분야의 유연성이다. 셋째, 기업이 보유 또는 향후 보유하게 될 기술력 잠재력 평가다. 넷째, 이를 시장에 판매할 수 있는 사업화 능력이다. 다섯째, 투자 가능한 기업 자산 등 각 메타데이터를 정의, 공공 또는 민간에서 축적할 수 있는 데이터베이스를 확보하는 일이다.

## 다양한 메타데이터 수집 전략이 필요할 때

이를 위해 공공, 민간의 다양한 메타데이터 수집 전략이 필요하다. 수집한 데이터를 기반으로 기계학습과 결합해 새로운 데이터를 생성, 유통하는 것도 방법이다. 구글과 같은 글로벌 IT기업들은 이미 유사한 전략으로 오픈 API를 실행하고 있다. 특히, 금융권은 개인과 기업 생성 데이터 중 가장 중요한 데이터인 금융데이터를 가지고 있다. 따라서 데이터 거버넌스 전략을 잘 활용하면 엄청난 부가가치를 이끌 기회를 만들 수 있다.

먼저 목표를 명확히 세우고 공공, 민간에서 생성한 다양한 메타데이터들을 목표에 맞춰 수집한다. 이를 금융데이터에 연결한 뒤 생성된 새로운 데이터를 다시 유통시킨다. 이런 선순환이 모여 가치 있는 플랫폼으로 거듭날 수 있다. 이를 위해 금융권에서도 공공과 민간의 데이터를 선별해 결합해야 한다. 이를 통해 데이터를 관리하고 가치를 생성하는 새로운 데이터 거버넌스 체계, 그 전략 수립이 필요한 시점이 아닐까.