국가 데이터 주권 확립을 위한 AI 학습 데이터 표준

장시환 한국전자통신연구원 콘텐츠연구본부 선임연구원

1. 머리말

AI 기술이 국가 경쟁력의 핵심으로 부상하며, AI 개발 기반이 되는 데이터의 중요성이 그 어느 때보다 강조되고 있다. 특히, 특정 국가나 기업에 종속되지 않고 자체적으로 AI 기술 및 인프라를 개발·운영·통제하는 '소버린 AI(Sovereign AI)'는 21세기 기술패권 시대의 핵심 지정학적 의제로 떠올랐다[1, 2]. 소버린 AI의 성공적 구현은 결국 '데이터 주권(Data Sovereignty)' 확보에 달려 있으며, 이는 국가가 자국 데이터에 대한 완전한 통제권을 가지고 생성·저장·처리·활용할 수 있는 권리를 의미한다[3]. 그러나 현재 국내 AI 생태계는 핵심 인프라인 DBMS(데이터베이스 관리 시스템, Database Management System)와 클라우드 서비스가 높은 해외 의존도를 갖고 있다는 구조적 문제에 직면해 있다[4]. 이러한 상황은 데이터 주권 침해 위험을 가중시키고, 장기적으로 기술 종속을 심화시킬 수 있다. 따라서 이번 원고에선 소버린 AI 시대 데이터 주권을 확립하기 위한 선결 과제로 'AI 학습 데이터 표준화 및 개방형 생태계'의 중요성을 역설하고, 이를 통해 국가 AI 경쟁력을 강화하기 위한 방안을 논하고자 한다.

2. 소버린 AI와 데이터 주권의 부상

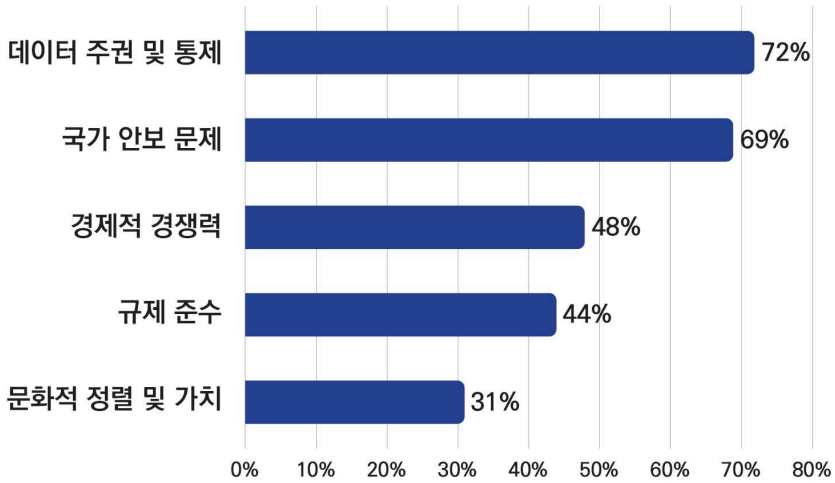
소버린 AI는 단순히 AI 모델을 개발하는 것을 넘어, AI 기술 전 주기에 걸쳐 국가적 통제권을 확보하는 것을 목표로 한다. 리눅스 재단(Linux Foundation) 2025년 보고서에 따르면, 전 세계 조직 79%가 소버린 AI를 중요한 전략적 과제로 인식하고 있으며, 데이터 주권 및 통제(72%)와 국가안보 문제(69%)를 주요 동인으로 꼽았다[5]. 이는 데이터가 단순한 자원을 넘어 국가안보와 경제적 자율성을 결정하는 핵심 자산으로 간주되고 있음을 시사한다.

이러한 배경 속에서 각국은 서로 다른 데이터 전략을 택하고 있으며, 크게 데이터 주권을 최우선하는 전략과 데이터 개방성을 강조하는 전략으로 나눌 수 있다[1, 4]. 전자는 국가안보 보호와 문화적 정체성 보존을 장점으로 하지만, 과도한 폐쇄성으로 인해 기술 고립이나 혁신 속도 저하라는 단점을 안고 있다[2]. 반대로 개방성을 중시하는 전략은 글로벌 협력과 기술혁신을 가속화할 수 있으나, 해외 의존도 심화와 문화적 획일화라는 위험을 수반한다[5, 6].

소버린 AI 전략은 두 극단 중 어느 한쪽을 선택하는 것이 아니라, 주권과 개방성 사이에서 균형을 찾는 과정이라고 볼 수 있다. 국가가 데이터를 통제·보호하는 동시에, 국제협력과 표준화 논의에 참여해 글로벌 경쟁력을 확보하는 것이 핵심 과제다[7]. 이러한 균형적 접근이 이뤄질 때만이, 데이터 주권을 지키면서도 기술·경제적 성장을 동시에 달성할 수 있을 것이다[8].

<표 1> 데이터 주권 vs. 데이터 개방성 전략 비교

구분	데이터 주권 중심 전략	데이터 개방성 중심 전략
정책 목표/목적	국가안보 보호, 데이터 주권 확보, 디지털 주권 강화	글로벌 AI 경쟁력 확보, 해외 기술 융합, 혁신 촉진
데이터 수집/관리 방식	국내 규제와 표준에 따라 엄격한 통제, 지역 내 저장·관리	국제협력 통한 데이터 공유, 클라우드 기반 접근성 중심
데이터 품질/ 다양성 확보 전략	자국 내 데이터 다양성 확보(지역, 사회 계층, 언어 등)	글로벌 데이터 풀 활용, 다국어·다문화 데이터 포함
데이터 접근 및 유통 방식	내부망/폐쇄형 유통, 접근 권한 엄격 제어	API 및 데이터 허브 기반 개방, 상호운용성 중심
보안 및 개인 정보보호 전략	강력한 암호화, 익명화, 내부 감사체계 강화	차등 접근 제어, 프라이버시 보존 기술 적용
기술 의존/자립성	국산 DBMS, 클라우드 인프라 우선 사용	해외 기술 활용 병행, 국제 플랫폼과 연계
혁신 및 협력 방식	내부 중심 연구개발, 자체 AI 모델 개발	다자간 협업, 오픈소스 프로젝트 참여
장단점 요약	(장점) 안보 강화, 주권 확보 (단점) 고립 가능성, 혁신 둔화	(장점) 협력 확대, 혁신 가속 (단점) 데이터 종속, 규제 충돌
주요 리스크	국제표준 미비, 비용 부담, 기술 갱신 지연	외국 기업 의존, 데이터 주권 침해, 보안 노출
정책 제언 방향	내부 표준 강화, 기술 국산화, 보안 강화	데이터 연맹 구축, 국제 규범 참여, 상호호환 표준 마련



출처: [5] M. Gerosa, et al., "The State of Sovereign AI: Exploring the Role of Open Source Projects and Global Collaboration in Global AI Strategy", Linux Foundation Research, Aug 2025.

[그림 1] 소버린 AI의 주요 관심 동인

3. 국내 데이터 인프라의 현실과 과제

소버린 AI와 데이터 주권의 중요성에도 불구하고, 국내 데이터 인프라 현실은 녹록지 않다. 2025년 행정안전부 통계에 따르면, 국내 공공기관의 외산 DBMS 점유율은 80.34%에 달하며, 이는 국가 주요 데이터가 해외 기업 기술에 의해 관리되고 있음을 의미한다[4]. 이러한 높은 외산 의존도는 데이터 관리의 투명성 저하, 높은 유지보수 비용, 그리고 잠재적인 데이터 유출 및 안보 위협으로 이어질 수 있다. 데이터 주권 확보의 첫 걸음은 데이터가 저장·관리되는 기반 인프라 자립에서 시작된다. 국산 DBMS 활용을 촉진하는 동시에 국내 데이터센터 및 클라우드 서비스 생태계를 강화하는 것은

외산 종속을 줄이고 데이터 통제권을 회복하기 위한 필수적 과제다. 국산 DBMS는 한국어 데이터 처리의 특수성을 살리고, 국내 법규 및 규제를 준수하며 안정적인 AI 학습 데이터 기반을 제공하는 데 강점을 가질 수 있다[3].

4. AI 학습 데이터 표준화 및 개방형 생태계의 필요성

데이터 주권을 지키는 것이 국가 AI 경쟁력 확보의 핵심 과제이지만, 동시에 글로벌 환경 속에서 개방형 생태계에 적극 참여하는 것도 중요하다. 이를 실현하기 위한 핵심 수단이 바로 AI 학습 데이터 표준화다. 데이터 표준화는 데이터 형식, 구조, 의미 등을 일관된 기준으로 정의해 상호운용성과 품질을 보장하는 과정이다. 이는 단순히 국가 내부 데이터 관리 차원을 넘어, 국제사회 협력과 교류 속에서 신뢰할 수 있는 데이터 생태계를 구축하는 기반이 된다. 결국, 데이터 주권을 확립하면서도 국제표준을 수용하고 개방형 데이터 생태계에 참여하는 것이 소버린 AI 전략의 핵심 요소로 작용한다. 즉, 표준화는 한편으로 외부 종속을 최소화해 국가 데이터 주권을 견고히 하고, 다른 한편으로 글로벌 생태계와의 호환성을 보장해 국제 경쟁력을 강화하는 '이중 전략'의 수단으로 기능한다.

4.1 데이터 품질 및 신뢰성 확보

표준화된 데이터는 AI 모델의 성능과 직결되는 데이터 품질을 체계적으로 관리할 수 있게 한다. 즉, 데이터 수집·정제·가공·검증 등 AI 파이프라인 전 과정에 걸쳐 일관된 품질 기준을 적용함으로써 모델의 편향성을 줄이고 신뢰성을 높일 수 있다[7, 8]. 특히 데이터 출처, 이력, 소유권 등을 명확히 하는 데이터 거버넌스 체계 구축에 있어 표준화는 필수적이다.

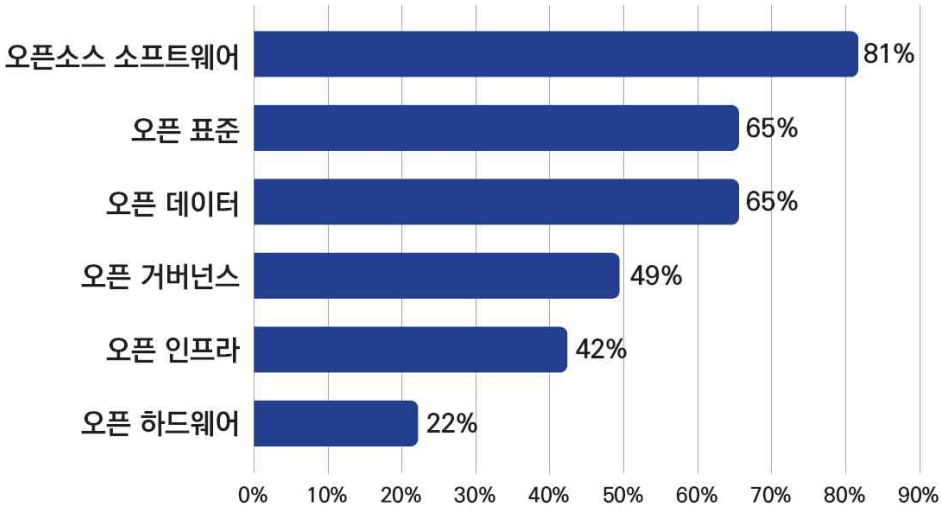
4.2 데이터 상호운용성 및 활용성 증대

데이터가 특정 플랫폼이나 시스템에 종속되지 않고 자유롭게 공유·활용되기 위해선 표준화가 전제돼야 한다. IDS(국제 데이터 공간, International Data Spaces) 원칙과 같이 데이터 주권을 보장하면서도 상호운용성을 높이는 프레임워크는 표준화된 데이터 모델을 기반으로 한다[8]. 이를 통해 국가 간, 산업 간 '데이터 동맹'을 구축해 고립주의를 넘어 데이터 풀을 확장하고, 공동 AI 모델을 개발하는 등 글로벌 협력 기반을 마련할 수 있다[9, 10].

4.3 개방형 생태계 구축 촉진

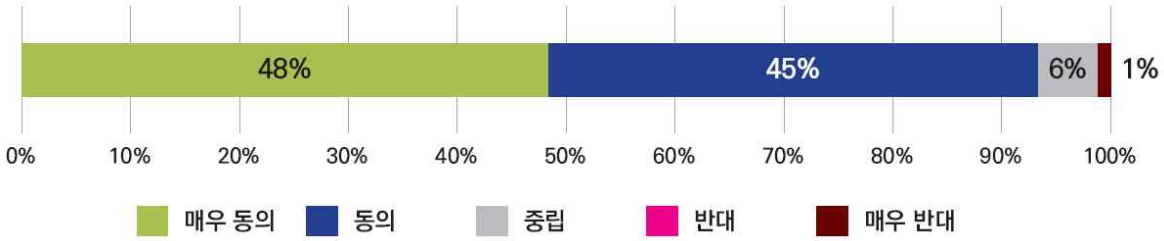
리눅스 재단 보고서는 소버린 AI 개발에 있어 오픈소스 소프트웨어(81%)가 가장 중요한 접근 방식으로 꼽혔다고 밝혔다[5, 10]. 데이터 표준화는 이러한 개방형 혁신 생태계를 촉진하는 핵심 동력이다. 표준화된 데이터는 특정 기업에 종속되지 않고 누구나 활용할 수 있는 '오픈 데이터'로의 전환을 용이하게 하며, 이는 다양한 AI 스타트업과 연구기관의 참여를 유도해 국가 AI 생태계 전반의 경쟁력을 강화하는 효과를 가져온다.

더 나아가, 이는 데이터 주권을 지키는 틀 안에서 개방형 생태계를 구축함으로써, 국가 핵심 데이터는 보호하면서도 비핵심 영역에서 개방과 공유를 통해 혁신을 가속화할 수 있다.



출처: [2] S. B. Chetty, et al., "Sovereign AI for 6G: Towards the Future of AI-Native Networks", arXiv:2509.06700, Sep 2025.

[그림 2] 소버린 AI 발전을 위한 선호 접근 방식



출처: [2] S. B. Chetty, et al., "Sovereign AI for 6G: Towards the Future of AI-Native Networks", arXiv:2509.06700, Sep 2025.

[그림 3] 안전하고 문화적으로 정합된 소버린 AI를 위한 기반으로서의 개방형 협력

5. 국가 데이터 주권 확립을 위한 정책 제언

국가 데이터 주권 확립과 AI 경쟁력 강화를 위해 다음과 같은 정책적 노력이 필요하다.

첫째, 범부처 차원의 국가 데이터 표준화 전략을 수립하고, 이를 총괄·조정할 수 있는 중앙 데이터 거버넌스 기구를 설립해 정책 일관성과 지속성을 확보해야 한다.

둘째, 국산 DBMS 및 주권형 클라우드 인프라에 대한 연구개발 투자와 정책적 지원을 확대해 외산 플랫폼 의존도를 줄이고, 데이터 주권의 기술적 기반을 강화해야 한다.

셋째, 데이터 주권의 사회적 인식 제고를 위해 국민교육과 공론장을 확대하고, 필요하다면 디지털 주권을 국가 핵심 가치로 간주하며, 이를 뒷받침할 법·제도적 기반을 강화해야 한다.

넷째, 고립주의적 접근을 지양하면서도 데이터 주권을 확고히 지키는 원칙 아래, 민주적 가치와 신뢰를 공유하는 국가들과 데이터 동맹을 구축해야 한다. 이를 통해 국제협력 속에서도 핵심 데이터 자산은 국내 통제 하에 유지하면서, 글로벌 데이터 표준과 AI 거버넌스 논의를 주도할 수 있다.

6. 맺음말

소버린 AI 시대 도래는 국가에게 있어 위기와 기회를 동시에 내포한다. 데이터 주권을 상실할 경우 기술 종속과 글로벌 경쟁력 약화라는 결과를 피하기 어렵다. 반대로, 선제적으로 데이터 주권을 확

립한다면 AI 강국으로 도약할 수 있는 전략적 발판을 마련할 수 있다.
그 핵심은 AI 개발 근간이 되는 학습 데이터의 표준화에 있다. 표준화를 통해 데이터 품질과 신뢰성을 체계적으로 보장하고, 상호운용성을 확보함으로써 국내 혁신 생태계뿐 아니라 개방형 글로벌 협력 생태계로의 참여를 동시에 실현할 수 있다. 또한 국산 데이터 인프라를 강화하고, 민주적 가치를 공유하는 국가들과 데이터 동맹을 주도함으로써, 고립을 피하면서도 주권을 견고히 유지할 수 있다.
궁극적으로 대한민국은 데이터 주권을 기반으로 한 균형 잡힌 표준화 전략을 통해 기술적 자립과 국제협력을 동시에 추구해야 한다. 이러한 노력이 병행될 때, 우리는 다가오는 소버린 AI 시대에 단순한 추종자가 아니라 글로벌 AI 질서를 선도하는 주권국가로 자리매김할 수 있을 것이다.

※ 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024년도 문화체육관광 연구개발사업으로 수행되었음(과제명 : 중소 게임 기업의 게임 제작 검증 효율화를 위한 AI 기반의 대규모 게임 자동검증 기술 개발, 과제번호 : RS-2024-00393500, 기여율: 100%)

[참고문헌]
[1] "AI 주권 전쟁: 소버린 AI와 데이터 주권의 부상", mediaCORPUS, 2025.07.18.
[2] S. B. Chetty, et al., "Sovereign AI for 6G: Towards the Future of AI-Native Networks", arXiv:2509.06700, Sep 2025.
[3] "[커버스토리] 소버린 AI 시대, 데이터 주권 확보 첫걸음은 '국산 DBMS'", 컴퓨터월드, 2025.08.29.
[4] "소버린 AI, 데이터 주권을 잃으면 기술 주권도 없다", TmaxTibero, 2025.09.09.
[5] M. Gerosa, et al., "The State of Sovereign AI: Exploring the Role of Open Source Projects and Global Collaboration in Global AI Strategy", Linux Foundation Research, Aug 2025.
[6] "이재명 정부 디지털 주권, 데이터 주권부터 확립하자", 서울이코노미뉴스, 2025.06.08.
[7] M. Altendeitering, et al., "Data Sovereignty for AI Pipelines: Lessons Learned from an Industrial Project at Mondragon Corporation," 2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN), 2022.
[8] Y. Hirota, et al., "International Testbed Data Sharing Framework with Data Sovereign Features for Network AI/ML Empowerment," 2025 Optical Fiber Communications Conference and Exhibition (OFC), 2025.
[9] "[데이터 주권] AI 경쟁력, 국가 협력 통한 데이터 개방이 핵심", ZDNet Korea, 2025.08.22.
[10] "[9월 월간브리핑] 소버린 AI 발전의 핵심 기반, 오픈소스와 글로벌 협력", 오픈소스 소프트웨어 통합지원센터(OpenUP), 2025.09.22.

[주요 용어 풀이]
• 소버린 AI(Sovereign AI): 특정 국가나 조직이 AI 개발·운영 전 과정에서 자율적 통제권을 보유하는 AI 체계. 데이터·인프라·모델에 대한 독립성과 주권을 강조하며, 국가 안보 및 기술 자립과 직결됨
• DBMS(데이터베이스 관리 시스템, Database Management System): 데이터를 효율적으로 저장·관리·검색하는 소프트웨어 시스템. AI 학습 데이터 기반을 안정적으로 제공하는 핵심 인프라

• GDPR(일반개인정보보호법, General Data Protection Regulation): EU가 제정한 개인정보보호 및 데이터 주권 강화를 위한 법률
• Gaia-X(Gaia-X European Association for Data and Cloud): EU 주도로 추진되는 데이터·클라우드 상호운용 생태계 프로젝트
• SNS JU(Smart Networks and Services Joint Undertaking): EU 차원의 6G 및 차세대 네트워크 연구개발 공동 기구
• Bhashini(Bhasha Interface for India): 인도의 다국어 AI 음성·텍스트 플랫폼으로, 22개 공용어를 포함한 지역 언어 디지털 격차 해소를 목표로 함
• TELUS Factories: 캐나다 통신사 텔러스(TELUS)가 운영하는 AI·데이터 혁신 허브
• SEA-LION(South East Asian Languages in One Network): 싱가포르가 주도하는 LLM(거대언어모델, Large Language Model) 프로젝트로, 동남아 다국어 지원을 통한 지역 특화 AI 인프라 구축을 지향
• AIRR(AI Research Resource): 영국이 구축 중인 국가급 AI 연구 인프라 및 슈퍼컴퓨팅 자원

※ 출처: TTA 저널 제221호