

오픈소스에서 AI 데이터까지, ETRI가 걸어온 오픈소스 거버넌스 확장의 길

• 박정숙 | ETRI 기획본부 정보전략부 품질혁신실장 | jungsp@etri.re.kr



들어가며

오픈소스 거버넌스는 조직이 오픈소스를 활용, 기여, 공개하는 전 과정에서 발생하는 법적/조직적 리스크를 체계적으로 관리하기 위한 정책, 프로세스, 조직 체계를 말한다^[1]. 시간이 갈수록 오픈소스는 매우 중요해지고 있는 반면 AI 기술과의 결합 및 ICT 기술의 발달로 오픈소스 거버넌스는 점점 복잡해지고 어려워지고 있다. 그러므로 기관별로 전담부서를 조직하여 체계적으로 관리하는 것이 매우 효율적이다.

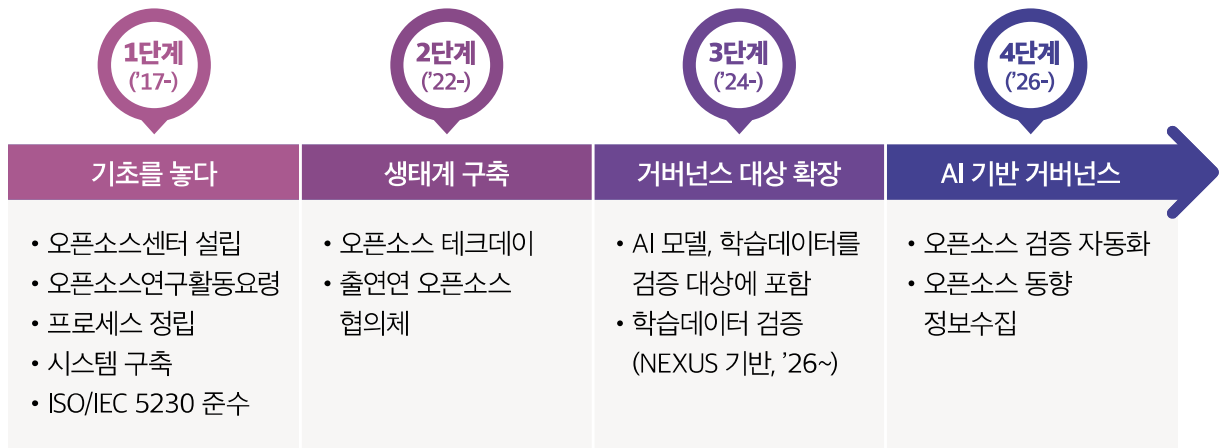
그러나 거버넌스 구축 시 기관별 고유한 특성들을 반영해야 효율적으로 동작하기 때문에 하루 아침에 완성되지 않는다. 따라서 체계 구축은 대상 기관의 상황들을 오픈소스 거버넌스에 제대로 녹여내는 것이 핵심이라고 할 수 있다. 특히 ETRI와 같은 정부출연연의 오픈소스 거버넌스는 R&D 결과물이 바로 상품으로 나오거나 대중들에게 노출되는 것이 아니기 때문에 그 중요성이 피부로 직접 느껴지는 것은 아닐 수 있지만 ETRI의 R&D 결과물이 대외 신뢰도를 가지면서 필요한 부담을 줄이기 위한 작업은 오픈소스 거버넌스 체계 정립에서 매우 중요하다고 할 수 있다.

돌이켜보면 몇 해 전까지만 해도 오픈소스 거버넌스는 비교적 명확한 질문들을 다루었다. “이 라이선스를 써도 되는가”, “배포 시 고지 의무를 충족했는가”, “외부 코드를 가져올 때 어떤 검토 절차를 거쳐야 하는가”와 같은 질문들에 대응하기 위한 체계(라이선스 검토 프로세스, OSS 관리 규정, 배포 전 점검 리스트 등)는 지난 10여 년간 여러 연구기관들과 기업에서 정교하게 다듬어져 왔다. ETRI 역시 2017년 전후로 오픈소스 관리 체계를 제도화하기 시작한 이래, 오픈소스 컴플라이언스를 중심으로 한 거버넌스 인프라를 꾸준히 축적해 오고 있다^[2].

그런데 최근 몇 년 사이, 특히 생성형 AI와 대규모 언어모델이 연구개발의 중심으로 들어오면서 이 질문은 충분하지 않게 되었다. “이 라이선스를 써도 되는가”라는 질문 옆에 “이 학습데이터는 어디서 왔고 누가 어떤 권한으로 썼는가”, “AI가 만들어낸 산출물의 저작권과 책임 소재는 누구에게 있는가”와 같은 질문들을 함께 수반한다. 소스코드의 계보를 추적하던 라이선스 관리 체계만으로는 데이터의 계보, AI 모델의 계보를 추적할 수 없다. 이는 오픈소스 거버넌스가 AI 데이터 거버넌스로 확장되어야 하는 이유가 될 수 있다.

연구기관의 오픈소스 거버넌스는 특정 시점에 설계도를 그리고 그대로 구축하는 것이 아니라, 기술 환경의 변화를 뒤쫓으며 계속해서 다시 그려지는 성격의 것이다. ETRI의 오픈소스 거버넌스 역시 그런 기술 발전과 보조를 맞추어 발전해 나가고 있다. 라이선스 준수를 위한 최소한의 절차에서 출발해, 생태계 참여의 문법을 익히고, 이제는 AI 학습 데이터와 내부 활용 데이터까지 아우르는 확장된 체계로 나아가고 있다. 이 글은 그 궤적을 시간의 흐름을 따라 짚어보며, 각 단계에서 무엇이 축적되었고 다음 단계로의 이행이 왜 필요했는지를 살펴본다. 이를 간략히 도식화하면 다음과 같다.

특히 ETRI는 오픈소스의 특성이 매우 복잡해지고 있어서 자체적인 거버넌스 운영으로는 한계가 있어 다른 출연연들과의 협력을 도모해 오고 있으며, 오픈소스 종사자들이 한자리에 모여 네트워킹을 형성하고 국내의 오픈소스 생태계를 발전시켜 나가는데 협력할 수 있도록 장을 제공하는 것도 중요한 내용으로 추진하고 있다.



1단계 — 제도의 기초를 놓다(2017년~)

오픈소스의 중요성을 일찍부터 인식한 ETRI의 오픈소스 거버넌스는 조직 차원에서 보다 체계적으로 대응하기 위해 2017년 6월 오픈소스 전담 조직인 오픈소스센터를 설립하는 것으로 공식적인 첫발을 내디뎠다. 그 후 2019년부터 제도적 골격을 갖추기 시작했다. 이 시기의 거버넌스 내용은 연구개발 과정에서 오픈소스를 가져다 쓰는 일이 늘어나는 만큼, 그로 인한 법적 리스크를 줄일 수 있는 최소한의 안전장치를 마련하는 것이었다. R&D 결과물에 대한 오픈소스 식별 및 라이선스 검토 절차가 수립되었고, 프로젝트 착수 시 오픈소스 사용 현황을 신고하는 체계가 도입되었으며, 외부 배포(기술이전, 공개 등) 이전에 한 번 더 소스 코드의 법적 리스크가 해소되었는지 확인하고 고지 의무 이행 여부를 확인하는 절차가 자리를 잡았다.

이 시기의 성과라면 오픈소스 거버넌스의 기초를 수립한 것이었다. 원규(오픈소스연구활동요령)를 제정하고, 오픈소스 거버넌스 프로세스(라이선스 검증, 오픈소스 심의, SW 공개)를 정립하고 시스템들을 구축하였다. 매년 프로세스 운영 경험을 오픈소스 리포트 시리즈로 정리하여 연구원에 공유하고 중요한 인사이트는 차년도 업무에 반영했다. 또한 연구원의 오픈소스 인식도를 제고하기 위해 매년 설문 조사를 실시하고 다양한 유형의 교육을 실행했다. 이 과정을 통해 오픈소스 리스크 및 컴플라이언스에 대한 연구원의 인식도가 높아진 점은 부인할 수 없다. 이후 수년간의 가이드 등의 문서화, 교육, 연구원의 환경에 맞는 시스템 개발 및 프로세스 정립은 초기 제도가 조직의 기억으로 축적되는 과정이었다.

2단계 — 안에서 밖으로, 생태계와 만나다(2022년~)

제도 및 시스템이 자리를 잡기 시작하면서 자연스럽게 대외 활동으로 이어졌다. 대외 활동은 국내 오픈소스 생태계 구축에 출연연으로서 기여할 수 있는 좋은 기회임과 동시에 내부의 오픈소스 거버넌스를 고도화하는데 큰 도움을 줄 수 있다. 또한 오픈소스 리스크가 다양한 라이선스 이슈 외에도 보안 취약점, AI 등 새로 대두되는 이슈까지 너무 복잡하기 때문에 배움의 장으로 활용하기 위한 목적도 있었다. 대외 활동의 첫 번째 실행은 오픈소스 테크데이였다. 2020년부터 시작된 본 행사는 처음에는 ETRI 단독으로 오픈소스 관련 국내 전문가들의 발표를 듣는 자리로 마련되었다. 코로나 시기와 겹쳐서 온라인으로 개최했지만, 오히려 온라인 참여자들이 대거 접속하면서 성공적으로 개최할 수 있었다. 2022년부터는 7개의 타 출연연들과 오픈소스 커뮤니티인 출연연 오픈소스 협의체를 결성하면서 오픈소스 테크데이 개최도 더 견고히 이루어질 수 있었다.

오픈소스 테크데이 개최 목적은 1) 출연연들의 오픈소스 R&D 성과 홍보 및 연구결과 확산, 2) 오픈소스 최신 동향 공유 및 활발한 소통·교류를 통한 R&D 혁신 선도, 3) 오픈소스 커뮤니티 활성화 및 상호 협력으로 오픈소스 R&D 생태계 강화, 4) 다양한 오픈소스 활동을 통해 출연연, 대학, 산업체, 기관 간의 협력 체계 구축, 5) 신규 R&D 아이템 발굴 및 협력 수행을 통한 기관 및 국가 경쟁력 강화 등이다.

출연연 오픈소스 협의체는 타 출연연들과 오픈소스에 대한 공동 대응을 도모하고자 조직되었다. ETRI가 2022년 8월 발족을 주도한 ‘출연연 오픈소스 협의체’는 8개 정부출연연구기관이 참여하는 네트워크로 출발했다. 협의체는 오픈소스 이슈 공동 대응, 거버넌스 구축 경험 공유, 활용 사례 확산, 그리고 매년 ‘오픈소스 테크데이’를 통한 성과 공유라는 네 축을 중심으로 운영되어 왔다. ETRI가 축적한 자체 거버넌스 체계와 R&D 관리 노하우가 이 협의체의 기반이 되었다는 점에서, 개별 기관의 실무 경험이 출연연 전체의 표준으로 확산·발전시켜 나갈 수 있는 시초가 되었다는 점은 큰 의미를 가진다.



2026년에 출연연 오픈소스 협의체는 범출연연 체계로 전면 확대된다^[3]. 이는 단순한 참여 기관 확대를 넘어, 오픈소스 거버넌스가 개별 기관의 리스크 관리 이슈에서 국가 연구개발 인프라의 근간을 관리하는 문제로 격상되었음을 보여준다. 확대의 직접적 배경에는 AI 학습데이터 저작권 문제, 라이선스 분쟁, 보안 취약점 등 오픈소스 관련 리스크가 AI·반도체·우주항공·국방 등 국가 전략기술 전반으로 확산되고 있다는 위기의식이 자리한다.

이러한 확장은 같은 시기 추진되고 있는 국가 AI 과학기술 정책, 특히 K-문샷 프로젝트와 국가과학AI연구센터(NAIS)의 문제의식과 정확히 맞닿아 있다^[4, 5]. 2026년 5월 출범한 K-문샷 프로젝트는 휴머노이드, 피지컬 AI 등 8대 분야 12대 국가 미션을 통해 2030년까지 AI 기반 연구 생산성을 2배로 끌어올리는 것을 목표로 한다. 이를 뒷받침하는 NAIS는 학습데이터·컴퓨팅 자원·AI 모델을 연결하는 ‘과학 AI 운영체제(OS) 플랫폼’을 구축해, AI가 연구자를 위한 ‘과학 발견 엔진’으로 기능하도록 지원한다는 구상이다. 여기서 주목할 지점은, K-문샷과 NAIS가 지향하는 “연구 데이터·모델·컴퓨팅 자원의 개방적 연계”라는 비전이, 바로 출연연 오픈소스 협의체가 지난 4년간 다뤄온 라이선스 준수·거버넌스·리스크 관리의 실무 위에서만 지속 가능하다는 사실이다. AI 모델과 학습데이터를 기관 간, 나아가 산·학·연 간에 자유롭게 연계·활용하려는 순간, 그 자원이 어떤 라이선스 조건 하에 만들어졌는지, 저작권 관계는 명확한지, 재배포·재학습 시 어떤 의무가 발생하는지에 대한 질문이 뒤따를 수밖에 없다. 즉 NAIS가 그리는 ‘과학 AI 운영체제’는 기술적 통합 계층인 동시에, 그 아래에서 데이터와 모델의 출처·권리 관계를 정리하는 거버넌스 계층을 필요로 한다.



올해로 7회째를 맞는 오픈소스 테크데이는 국가 AI 연구 인프라 구축과 오픈소스 거버넌스 실무가 실제로 만나는 현장이다. 2025년 테크데이는 정책·동향, 개발자·커뮤니티, R&D, 컴플라이언스, 키노트의 5개 트랙으로 구성되었으며, R&D 트랙에서는 ETRI의 AI 반도체 연계 로봇지능 개발, KISTI의 오픈소스 기반 과학 플랫폼, KAERI의 원자력 시뮬레이션 엔진 공개 사례가, 컴플라이언스 트랙에서는 민간의 법률 준수 체계가 함께 논의되었다. 공공 연구기관의 R&D 성과와 민간의 컴플라이언스 실무가 한자리에서 교차하는 이 구조는, 앞으로 K-문샷이 추진하는 산·학·연 협력형 미션 연구에서도 그대로 요구될 모델이다. 오픈소스 테크데이는 해마다 오픈소스와 AI 거버넌스가 만나는 현장으로서의 역할을 충실히 해나갈 예정이다.

요약하자면, 출연연 오픈소스 협의체의 범출연연 확대는, 국가가 AI 기반 과학기술 혁신을 가속화하려고 할수록 그 인프라의 신뢰성과 지속가능성을 뒷받침하는 거버넌스 체계 또한 국가 단위로 함께 확장되어야 한다는 사실을 보여주는 사례다. ETRI가 이 과정에서 축적해 온 거버넌스 노하우가 NAIS를 중심으로 한 과학 AI 인프라가 흔들림 없이 확산되기 위한 실질적 기반이 될 수 있도록 노력할 예정이다.

3단계 — 대상이 코드에서 데이터로 넓어지다(2024년~)

AI 모델의 가중치 값이 GPL 라이선스의 파생 저작물인지에 대한 소송인 Stockfish v. ChessBase 사건(2021) 및 생성형 AI 학습과 저작권 침해에 관한 대표적인 소송인 Getty Images v. Stability Diffusion 사건(2023)이 발생했을 때 업계는 많은 관심을 보이면서도 판결 결과를 좀 더 지켜봐야 한다는 의견이 우세했다. 그러나 그 후 AI 학습데이터 및 추론의 저작권 침해와 관련한 소송은 매우 빈번히 지속적으로 발생하고 있다. 이제 의미 있는 판결이 나오기 시작하지만, 아직 정립되기까지는 많은 시간이 필요할 것으로 판단된다.

최근 AI가 연구개발의 중심 도구로 자리 잡으면서, ETRI 거버넌스도 새로운 국면을 맞았다. 기존의 코드 중심 관리 체계로는 감당할 수 없는 질문들이 늘어나기 시작한 것인데, 학습데이터셋의 출처 확인, 외부 공개 학습데이터셋 활용의 적법성 여부, AI가 만들어낸 산출물의 권리 관계, 내부 구축 학습데이터셋의 저작권 위반 여부 등이다.

ETRI를 포함한 국내 주요 연구기관들은 오픈소스 활용이 본격화된 2010년대 중후반부터 체계적인 관리 기반을 마련해 왔다. 라이선스 검토 프로세스를 표준화하고, 프로젝트 착수 단계에서 오픈소스 사용 현황을 신고·검토하는 절차를 도입했으며, 배포 전 고지 의무 이행 여부를 점검하는 체크리스트를 운영했다. 이러한 체계는 법적 리스크의 사전 차단, 조직 내 인식 제고, 대외 신뢰도 측면에서 뚜렷한 성과를 남기고 있다. 특히 오픈소스 테크데이 개최, 오픈소스 서밋 참여, 오픈소스 애뉴얼 리포트 발간 등을 통해 ETRI가 오픈소스 생태계에 책임 있게 참여하는 기관이라는 이미지를 축적해 왔다.

그러나 이 체계는 본질적으로 “코드 단위의 저작물”을 전제로 설계되었다. AI 연구개발 환경에서는 학습 데이터의 출처 불투명성, 내부 활용 데이터의 이중적 성격, 산출물의 책임 소재 모호성, 속도의 문제 등에서 기존 체계를 벗어나 새로운 거버넌스 체계를 요구하고 있다. 참고로, 연구기관 내부에서 생성되는 실험 데이터, 로그, 문서 등은 그 자체로는 대외 배포 대상이 아니지만, AI 모델 학습이나 내부 에이전트 시스템의 지식 기반으로 활용되는 순간 별도의 거버넌스 대상이 될 수 있다. 기존 오픈소스 관리 체계는 “외부로 나가는 것”을 통제하는 데 초점이 맞춰져 있어, “내부에서 순환하며 재사용되는 데이터”에 대한 통제 감각이 상대적으로 부족하다. 거버넌스 활동의 성과를 어떻게 정량화하고 조직 내외에 설득력 있게 전달할 것인가는, 제도 설계 못지않게 중요한 과제로 남는다.

기존에는 “소스코드”가 관리 대상의 거의 전부였다면, 이제는 학습데이터, 파인튜닝 데이터, 내부 활용 데이터, 그리고 그 결과물인 모델까지가 관리 대상에 포함된다. 이는 단순히 관리 항목을 추가하는 문제가 아니라, 각 대상이 가진 서로 다른 위험 구조를 이해하고 그에 맞는 절차를 별도로 마련하는 작업을 요구한다. 코드의 위험이 주로 ‘라이선스 조건 위반’에 있다면, 데이터의 위험은 ‘수집 근거의 부재’와 ‘개인정보·기밀 포함 여부’에 있고, 모델의 위험은 ‘학습 이력의 추적 불가능성’에 있다.

ETRI는 거버넌스 대상을 데이터까지 확대하기로 함에 따라 AI 모델 및 학습데이터 거버넌스 체계를 수립하고 있다. LG AI 연구원에서 개발한 AI 컴플라이언스 도구인 EXAONE NEXUS를 활용하여 공개된 데이터셋을 사용하는 경우와 내부 구축 학습데이터셋 모두에 대해서 저작권 침해 이슈가 발생하지 않도록 데이터셋의 출처를 밝히는 프로세스를 올해 안에 확립하고자 한다.

4단계 — 시스템과 결합하다(2026년~)

3단계에서 시작한 ETRI의 AI 거버넌스는 시스템과의 결합을 수반해야 한다. ICT 업계는 AI를 활용하여 업무 자동화를 구축하고 있다. ETRI도 기관 차원에서 활용할 수 있는 인프라, 데이터 품질관리, 연구지원, 보안 등 AI 기반으로 연구원 업무를 자동화하는 작업을 진행 중이다. 이 체계는 각 영역이 독립적으로 존재하는 것이 아니라, 도메인별 에이전트가 서로 다른 데이터와 도구에 접근하며 유기적으로 작동하는 구조를 지향한다.

이 지점에서 거버넌스와 기술이 만나게 된다. 여러 에이전트가 서로 다른 데이터에 접근하는 환경에서는, 접근·호출 이력 자체가 거버넌스의 핵심 증거가 된다. MCP(Model Context Protocol)와 같은 연결 계층은 이런 이력을 표준화된 방식으로 기록할 수 있는 지점을 제공한다. 즉, 라이선스 검토가 사람이 체크리스트를 확인하는 절차였다면, 이 단계의 거버넌스는 시스템이 자동으로 이력을 남기고 정책을 집행하는 방향으로 향해 가고 있다. 이와 함께 ETRI는 Agentic AI 시스템의 지식재산 보호 전략에 대해서도 함께 검토하며, 기술적 확장이 법적·제도적 정합성을 잃지 않도록 병행하고자 한다. 이 외에도 ETRI는 AI 기술을 활용하여 오픈소스의 최신 동향을 수집하는 서비스를 구축하고 있다.



나가며

ETRI의 거버넌스 확장은 크게 세 번의 전환을 네 단계 거쳐왔다고 할 수 있다. 첫째는 “오픈소스는 무엇을 지켜야 하는가” 하는 오픈소스 라이선스 준수 여부에 대한 내부적인 컴플라이언스 체계를 정립하는 것이었고, 둘째는 출연연 오픈소스 협의체 운영과 오픈소스 테크데이 개최를 통한 오픈소스 생태계 구축을 위한 것이었으며, 셋째는 오픈소스와 AI의 결합에 따라 AI 데이터 거버넌스를 설계하고 시스템을 결합하는 작업이다. 각 단계는 이전 단계의 성과 위에 세워졌다. 또한 단계 진화에 따라 거버넌스의 범위는 그만큼 넓어지게 된다. 라이선스 관리의 경험 없이는 데이터 거버넌스의 위험 감각을 가질 수 없었을 것이고, 대외 소통의 경험 없이는 복잡해진 거버넌스 체계를 조직 안팎에 설득력 있게 전달하기 어려웠을 것이다.

거버넌스 확장은 여전히 진행 중이며 특히 AI 거버넌스는 이제 시작한 단계로서, 데이터의 출처를 투명하게 밝히는 일, 내부에서 순환하는 데이터의 재사용 기준을 세우는 일, 에이전트 간 상호작용의 이력을 신뢰할 수 있는 형태로 남기는 일 등은 꽤 많은 시간이 소요될 것으로 예상된다.

다만 분명한 것은, ETRI가 걸어온 이 길이 특정 시점에 완성될 성질의 것이 아니라 기술 환경의 변화에 맞춰 계속해서 다시 그려져야 하는 여정이라는 점이다. 라이선스 관리에서 시작해 생태계 참여를 거쳐 AI 데이터 거버넌스에 이르는 지금까지의 궤적은, 그 자체로 연구기관이 신뢰할 수 있는 AI 시대를 준비하는 방법에 대한 하나의 좋은 레퍼런스 포인트가 될 수 있을 것이다.

덧붙여, ETRI가 지키고자 하는 또 하나의 가치는 출연연 오픈소스 협의체를 안정적으로 운영하여 출연연들, 중소기업들을 포함한 다른 기관들과 협력하며 함께 성장해 나가는 노력을 계속할 것이라는 점이다. 확대된 이슈는 너무 복잡하고 어려워서 단독 대응보다는 협력 대응이 더 효율적이며 또한 사기업이 아닌 정부출연연으로서의 역할에 더 맞기 때문이다.

참고문헌

- [1] TTA 정보통신용어사전, “오픈소스 소프트웨어 거버넌스”, <http://word.tta.or.kr>.
- [2] ETRI 오픈소스 애뉴얼 리포트 2025, https://www.etri.re.kr/kor/bbs/view.etri?keyField=&keyWord=&nowPage=1&b_board_id=ETRI01&year_gubun=1&b_idx=19711
- [3] 헬로디디 외, “국가 R&D도 오픈소스 공동 대응, ETRI 범출연연 협의체로 확대”, <https://www.hellodd.com/news/articleView.html?idxno=112387>, 2026.7.7.
- [4] 과학기술정보통신부, “과학기술 12대 국가미션 해결 ‘K-문샷 프로젝트’ 개시”, 대한민국 정책브리핑, 2026.5.27.
- [5] 국가과학기술연구센터(NAIS), <https://www.nais.re.kr>.