AI를 활용한 지속가능한 건축물 데이터 생태계 구축*

안의순
건축공간연구원 부연구위원

건축물 데이터, 그 가치와 과제

정확한 데이터는 근거 기반 행정의 출발점이다. 건축물은 부동산, 지역 개발, 산업 정책 등 국가 행정 전반을 뒷받침하는 기초 공간 단위이므로, 건축물의 현황을 정확히 파악하는 것이 중요하다. 건축물대장이 「건축법」에 등장한 것은 1975년 개정 「건축법」(법률 제2852호)에서 건축허가 및 신고 접수 내역을 건축물대장에 기재하고 보관하도록 규정한 것이 시초이다. 이후 1992년 개정 「건축법」(법률 제4381호)에서 건축허가 내역이 아닌 '건축물 및 그 대지에 관한 현황'을 관리하고 이를 건축정책의 기초자료로 활용하도록 하면서 현재와 같은 건축물대장이 등장하였다. 현재 건축물대장은 전국 모든 건축물의 소재지, 면적, 용도, 구조 등을 기록한 공적 장부로서 건축통계와 정책 수립의 근간이 되고 있다.

그러나 이 데이터는 과연 얼마나 정확할까? 2022년 건축행정시스템 세움터 점검 결과 전체 점검 대상 데이터 6억 2,000만 건 중 약 855만 건에서 오류가 발견되었다. 전체 오류율은 1.37%에 그친 셈이지만, 특정 항목에서는 25%에 달하는 오류율이 나타나기도 하였다. 건축면적이 대지면적보다 크다고 기재된 경우, 층·동별 면적의 합계가 전체 면적과 일치하지 않는 경우 등이 대표적이다. 서로 같아야 하는 두 수치가 다르게 나타나는 경우, 둘 중 무엇이 맞는지를 판단하여 바로잡기 위해서는 오류 전수에 대한 현장조사가 필요하다. 그러나 이러한 전수 현장조사는 막대한 인력과 비용 및 시간이 요구되기 때문에 현실적으로는 불가능하다.

인공지능(AI) 기술은 이러한 한계를 넘어설 수 있는 현실적 대안으로 주목받고 있다. 현장조사를 완전히 대체하는 것은 아니지만, 사람의 개입을 최소화하면서도 데이터 전수를 검토할 수 있다는 점에서 AI 기반 데이터 품질 진단은 건축물 데이터 관리에 새로운 가능성을 열어 준다. 이 글에서는 건축물(건축행정) 데이터의 역할과 중요성, 현재 데이터의 품질 현황과 오류 양상을 짚어본 뒤에 기계학습을 활용한 면적·용도 데이터의 이상값 탐지 등 AI 방법론의 시범적 적용 결과를 살펴보고, 이를 바탕으로 건축물 데이터 품질 개선을 위한 정책 방향을 제시하고자 한다.

건축물 데이터의 역할과 품질 현황

국가 행정의 기초, 건축물 데이터의 활용 현황

「건축법」 제38조는 시·군·구청장 등 허가권자가 건축물과 그 대지의 현황 정보를 건축물대장에 기재하여 보관하고 지속적으로 정비하도록 규정하고 있다. 이를 근거로 구축된 건축물대장은 건축통계의 작성 기반이자 정책 수립의 기초자료로 폭넓게 활용된다. 건축물통계에서는 시·군·구별 연면적·용도·층수가, 건축허가·착공·준공통계에서는 구조·용도·연면적·허가일 등이 핵심 변수로 쓰인다. 건축물대장 기반 보고통계로 전환하기 이전의 건축통계는 각 허가권자가 작성 규칙에 따라 수기 조사를 실시한 결과를 집계하여 작성되었다. 2001~2002년부터 이 통계들은 건축행정시스템(세움터) 데이터에 기반한 자동 집계 방식으로 전환되었다. 따라서 건축물대장 등 건축행정 데이터의 품질은 곧 국가 건축통계의 신뢰성과 직결되며, 이어 통계에 기반한 행정 정책 수립의 신뢰성에도 영향을 미친다.

건축물 데이터는 통계에 그치지 않고 연구와 산업 현장에서도 폭넓게 쓰이고 있다. 건축공간연구원에서 그동안 수행해 온 연구는 건축물 데이터의 활용 잠재력을 여러 측면에서 실증적으로 뒷받침하고 있다. 건축물의 화재와 침수 발생 및 피해 리스크 분석에 대한 일련의 연구에서 건축물의 용도·구조·연면적·사용승인일 등이 주요 독립변수로 활용되었다(안의순, 박종훈, 2023; 조영진 외, 2022, 2023). 건축물의 범죄예방 환경

* 이 글은 안의순 외(2025)의 일부 내용을 발췌·보완하여 작성하였다.

설계(CPTED) 고도화 연구에서는 건축물에서 발생한 범죄와 물리환경의 관계를 분석하기 위하여 건축물의 건폐율·용적률·건물높이 등과 공간정보를 결합하여 사용하였다(조영진 외, 2024a). 그 외에도 건축물대장 데이터를 사업자등록 정보와 연계하여 빈 건축물을 추정하고 그 특성을 분석하거나(조영진 외, 2024b), 지역 내 건축물 재고 특성을 파악할 수 있는 건축물 연령 지표를 개발하는 등(송유미 외, 2024) 다양한 연구에서 건축물 데이터가 핵심 자료원으로 기능하였다.

이처럼 건축행정 데이터는 그 자체로도 활용 가능성이 높지만 다양한 공공·민간 데이터와 결합함으로써 그 가치를 더욱 높일 수 있다. 최근 건축물(동) 단위 고유식별자인 건물ID가 건축행정 데이터(인허가, 건축물대장)에 도입되면서 건축행정 데이터 간 활용 및 연계 기반도 한층 강화되었다. 앞으로 건축행정 데이터 외 건축물 관련 데이터에 건물ID가 도입되면 건축물 데이터의 활용 및 연계 기반도 더욱 강화될 것으로 기대된다.

건축물 데이터의 신뢰성을 가로막는 품질 문제

이처럼 중요한 데이터임에도 건축물 데이터의 오류 문제는 반복적으로 제기되어 왔다. 국토교통부는 2018년부터 지자체에 건축물대장 정비를 요청해 왔으나, 연도별 정비 완료율은 27~58% 수준에 머물렀다. 2022년 세움터에서는 건축행정시스템에서 관리하고 있는 건축물대장, 건축인허가, 주택인허가 등의 데이터에 대하여 86개 업무규칙을 기준으로 전수 점검을 실시하였는데, 일부 규칙에서 매우 높은 오류율이 존재하는 것으로 드러났다.

건축물대장에서는 대지면적보다 건축면적이 큰 경우(순번 13)의 오류율이 25.02%로 가장 높은 것으로 나타났다. 4건 중 1건이 오류인 셈이다. 대지 내 각 건축물 동의 건축면적을 합산한 결과가 총괄표제부 건축면적과 일치하지 않는 경우(순번 16)는 16.78%로 그 뒤를 이었다. 이 외에도 표제부와 총괄표제부 간 면적·용적률 관련 불일치가 10~17% 수준으로 나타났다.

건축인허가 데이터에서는 건축허가대장과 주차장 테이블 간의 주차대수 불일치(순번 43) 발생이 21.50%로 가장 높았고, 주택인허가 데이터에서는 기본개요와 동별개요 간 세대수 불일치(순번 76) 비율이

2022년 86개 업무규칙 및 정비대상 항목(오류율 상위 일부)

순번	구분	업무규칙명	오류건수	점검대상 건수	점검 당시 오류율
1	건축물대장	(건축면적) 총괄표제부 건축면적과 표제부 건축면적 합계의 일관성	103,141	593,350	17.38%
2		(연면적) 총괄표제부 연면적과 표제부 연면적 합계의 일관성	80,147	593,350	13.51%
3		(용적률산정연면적) 총괄표제부 용적률산정연면적과 표제부 용적률산정연면적 합계의 일관성	78,756	593,350	13.27%
5		(용적률) 총괄표제부 용적률 계산의 정확성	81,647	593,341	13.76%
13		(대지면적) 총괄표제부 내 대지면적의 값이 건축면적보다 작은 경우의 데이터 검증(정확성)	150,276	599,790	25.05%
16		(건축면적) 일반건축물대장 및 표제부의 바닥면적 합이 가장 큰 면적과 표제부의 건축면적이 다른 데이터 검증(정확성)	1,334,538	7,952,243	16.78%
20		(용적률) 일반건축물대장 및 표제부 용적률 계산의 정확성	723,474	5,212,430	13.88%
43	건축인허가	(주차대수) 건축허가대장의 총주차대수와 주차장 테이블의 주차대수 합이 다른 데이터 오류 검증	551,217	2,564,065	21.50%
76	주택인허가	(세대수) 허가대장 기본개요의 세대수와 동별 개요 세대수 합이 상이한 데이터 오류 검증(일관성)	5,861	19,252	30.44%
총합계			8,550,786	623,681,869	1.37%

출처: 세움터 내부자료; 안의순 외(2025, p.37)에서 재인용.

30.44%로 나타나 인허가대장 데이터에서도 오류가 심각한 수준임을 확인할 수 있다.

건축물 데이터를 활용한 선행 연구에서도 실제 데이터 활용 과정에서 드러난 다양한 오류를 보고하고 있다. 건축물의 주 용도 코드에 현행 정상 코드 대신 과거 영문·숫자 혼합 코드가 남아 있거나 아예 '창고'와 '사찰' 같은 이질적 값이 섞여 있는 사례, 소수점 오기로 인해 단일 건축물의 연면적이 여의도 면적의 30배가 넘는 1억m^2에 달하는 값으로 기재된 사례 등이 대표적이다. 이러한 오류는 현행 건축행정 시스템의 입력 내용 검증 기능 미비, 과거 데이터 갱신 및 정비 미비, 지역별 관리 방식의 편차 등 복합적인 요인에서 비롯된다.

내부적으로 완결된 체계를 갖추고 있는 건축행정 데이터의 경우와 달리, 서로 다른 건축물 데이터 간 연계·활용 시에는 데이터 품질 문제가 더 두드러지게 나타난다. 한 예로 건축물대장(795만 1,324동)과 도로명주소대장(1,075만 2,596동) 간 건물 수 차이가 283만 동에 달하고, 양 데이터 간 정보 일치율은 약 70%에 불과하다. 현재 건축물 동 단위 데이터

연계는 고유한 연계키(건물ID 등)나 경위도 좌표 등 잘 정의된 식별체계가 아닌 주소와 동 명칭 등을 활용한 간접적인 방식으로 이루어지고 있기 때문이다. 다만 향후 건물ID의 도입 확산을 통하여 모든 건축물 데이터에 건물ID가 부여된다면 건축물 데이터의 연계·활용이 크게 개선될 수 있을 것으로 기대된다.

AI를 통한 품질 검증 방법론 시범 적용

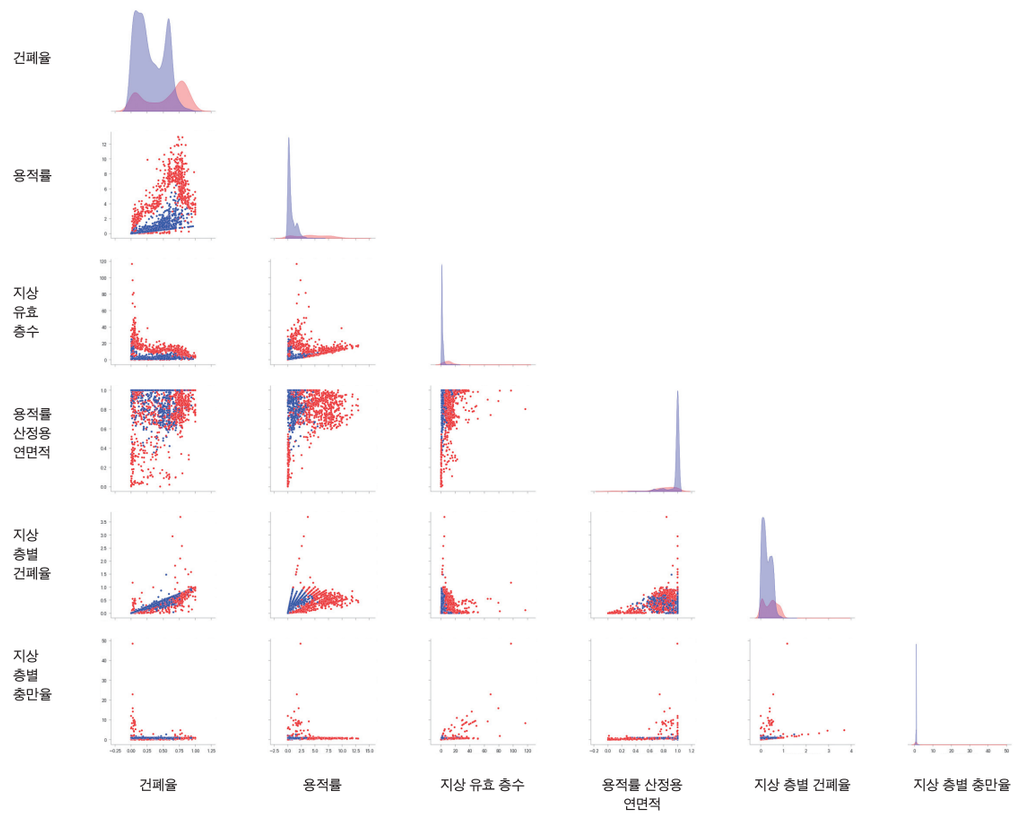
이러한 문제를 해소하기 위해 규칙 기반 진단과 기계학습 이상값 탐지라는 두 가지 접근을 상호 보완적으로 활용하는 다층적 품질 검증 방법론을 시범 적용하였다.

먼저 면적 데이터의 검증은 오류율이 가장 높게 나타난 면적 관련 항목을 대상으로 규칙 기반 방법론과 기계학습 기반 방법론 두 방향으로 검증 방법론을 시범 적용하였다. 규칙 기반 방법론은 연구자가 설정한 규칙에 기반한 오류 탐지로, 크게 두 단계로 진행하였다.

첫 번째 단계는 기존 규칙의 적용이다. 2022년 정비규칙 4개(대지면적 대비 건축면적, 건축면적 합계, 건폐율, 용적률)를 2024년 말 기준 최신 데이터에 적용한 결과, 0.13~5.01%의 오류율이 확인되었다. 그중 건축면적이 연면적보다 큰 경우가 가장 빈번하게 발생한 오류 유형이었다.

두 번째 단계는 기존 86개 규칙이 다루지 않던 신규 검증규칙의 개발이다. 용적률 산정 연면적이 일반 연면적보다 큰 경우(0.26%), 건축물 연면적이 통계적으로 도출한 상한을 초과하는 경우(0.86%) 등을 규칙화하여 기존 규칙에서 다루지 못하였던 빈틈을 채우고 규칙 기반 검증의 범위를 확장하였다.

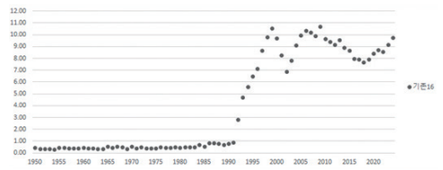
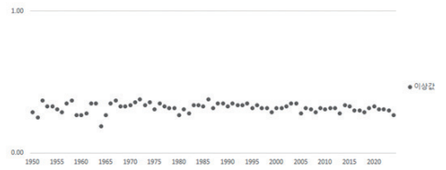
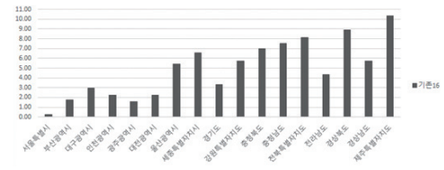
규칙만으로는 포착하기 어려운 잠재적 이상값의 경우는 기계학습으로 탐지하는 방법론을 적용하였다. 이때 건축물 규모의 절대적 크기 차이에 영향받지 않도록 대지면적 대비 건축면적 비율, 연면적 대비 용적률 산정용 연면적 비율, 지상 층별 층만율 등 그 값이 건축물 규모에 영향을 받지 않는 무차원 지표를 설계하였다. 그리고 도출된 지표값을 비지도 기계학습 알고리즘인 'Isolation Forest'와 'One-Class SVM'에 입력하여 일반적인 범위에서 벗어나는 잠재적 이상값을 탐지하였다.



기계학습 기반 이상값 산점도
주: 파란색은 기계학습 모델이 정상으로 분류한 데이터, 빨간색은 이상값 분류 데이터
출처: 안의순 외(2025, p.97)

Isolation Forest는 데이터를 반복적으로 분할하는 과정을 거쳐 쉽게 고립되어 다른 정상값과 구분하기 쉬운 값을 이상값으로 도출하는 알고리즘이다. 다른 기계학습 방법론인 One-Class SVM은 전체 데이터의 분포에서 경계를 찾고, 그 밖에 위치한 값을 이상값으로 판별한다.

서로 다른 방식으로 이상값을 탐지하는 두 알고리즘의 결과를 시기별·지역별로 교차 검토한 결과, 규칙 기반으로 탐지된 오류의 경우 특정 시기에 집중되어 나타났지만, 기계학습 방법론으로 도출된 이상값은 시기별로 고르게 나타나는 대신 서울·부산 등 대도시에 집중되어

	규칙 기반 오류 사례(16. 건축면적 오류)	기계학습 기반 이상값
연도별		
시·도별		

규칙 기반 오류와 기계학습 기반 이상값의 연도별, 시·도별 분포
출처: 안의순 외(2025, pp.80, 82, 96)

나타나는 패턴을 확인하였다. 이는 기계학습 기반 방법론으로 도출한 이상값이 특정 시기의 제도적·행정적 요인에 기인한 단순 입력 오류와 검증 미비로 발생한 것이 아니라, 지역의 규모에 기인한 구조적 오류임을 시사한다.

용도 데이터의 검증은 건축물 데이터의 상당 부분을 차지하는 텍스트 데이터의 품질 검증을 위하여 시도한 것이다. 건축물대장의 경우 자유형식 텍스트로 기재된 항목이 다수 포함되어 있으나, 그 내용에 대한 검증이 충분히 이루어지지 않고 있다. 2022년 세움터 점검 당시 건축물 용도와 구조 등의 분류 코드도 검증 대상에 포함하였으나, 실제 건축물대장 발급 시 표기되는 용도와 구조 기재 내용은 분류 코드와는 별도 항목으로 저장되어 있어 검증 대상에 포함되지 않았다. 용도와 구조 등 분류 코드는 건축통계 작성과 연구 등에 활용되고 있어, 건축물대장 기재 내용과 일치 여부 검증은 건축물 데이터의 신뢰성 확보에 꼭 필요하다.

용도 데이터의 검증은 실제 건축물대장 기재 내용이 포함된 '기타용도' 텍스트에 나이브 베이스(Naive Bayes) 분류 모델을 적용하고, 그 분류 결과를 데이터에 포함된 분류 코드(주 용도 코드)와 비교하는 방식으로 진행하였다. 기타 용도 텍스트를 구성하는 단어를 특성으로 삼아 「건축법」상 용도 코드를 예측하고, 예측값과 실제 코드 간 불일치를 오류 후보로 탐지하는 방식이다.

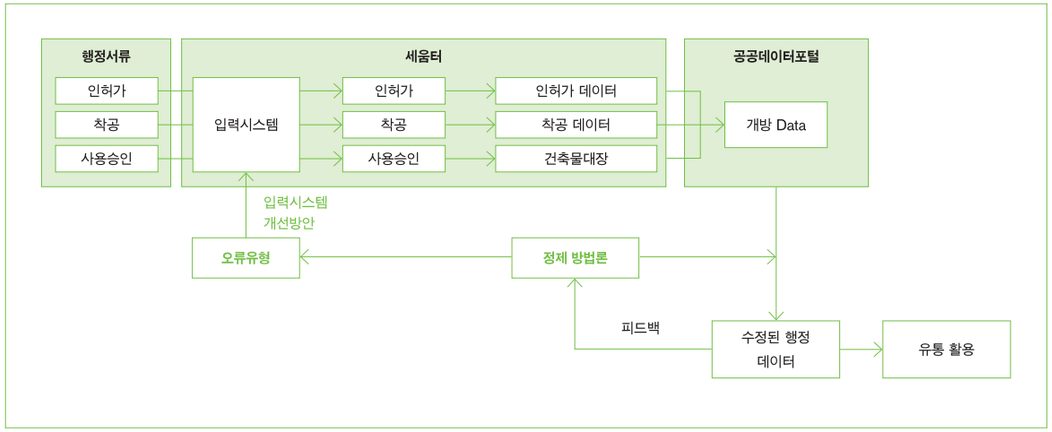
분석 결과 동별 용도 데이터의 예측 정확도는 84.78%, 층별 데이터는 78.97%를 달성하였다. 다만 동별로 여러 용도가 복합된 경우 건축물대장 용도 기재 내용('기타 용도')만으로는 주 용도를 제대로 분류하지 못하는 한계를 보였으며, 건축물대장 층별개요는 해당 층의 주 용도를 기재하지 않은 경우가 많아 개별 층 데이터만으로는 정확한 예측이 어려운 것으로 나타났다. 따라서 용도 데이터의 검증에는 건축물 전 층의 용도를 종합적으로 고려한 검증 체계가 필요함을 확인하였다.

정책 제언: 지속가능한 데이터 생태계 구축

이상의 시범 적용 결과를 토대로, 건축물 데이터 품질을 지속적으로 높여나가기 위한 세 가지 방향을 제안한다.

첫째, 품질 제고 데이터와 방법론의 오픈소스 유통이다. 건축행정 데이터는 공적 장부에 기반하고 있어 직권 정정을 통한 오류 수정이 쉽지 않다. 이를 보완하기 위해 AI 기반 방법론으로 품질을 개선한 데이터를 세움터가 아닌 다른 기관에서 별도로 생산하고 공공데이터포털을 통해 공개하는 방안을 제안한다. 오류 탐지 규칙, AI 알고리즘 코드, 교차검증 프로세스도 문서와 코드 형태로 함께 공개하면 재현성을 높이고 민간 활용을 촉진할 수 있다. 운영 주체를 명확히 하고 버전 관리 체계를 갖춤으로써 이용자 피드백을 반영한 지속적인 품질 개선이 가능하다.

둘째, 데이터 생성 단계에서의 입력 시스템 개선이다. 오류는 입력 시점에서 차단하는 것이 가장 효과적이다. 현행 입력 구조에서는 면적 항목을 담당자가 직접 숫자로 입력하기 때문에 오류 발생 가능성이 구조적으로 높다. 건축허가·신고 단계에서 면적 간 관계식(예: 건축면적 ≤ 대지면적, 연면적 = 각 층 면적 합산)에 기반한 자동 계산을 반영하거나, 입력값이 계산값과 불일치할 경우 경고창을 띄우는 방식을 도입하면 오류를 입력한 즉시 이를 감지할 수 있다. 기타 용도 텍스트 입력 시에도 표준 코드와의 정합성을 실시간으로 확인하는 보조 기능을 추가하면 코드 오기를 줄일 수 있다. 사용승인 단계에서는 건물ID를 연계키로 활용하여 인허가



건축물 데이터 품질 제고 방안
출처: 안의순 외(2025, p.147)

단계 데이터와 건축물대장 생성 시 데이터를 교차 검증함으로써 데이터 무결성을 한 번 더 확인하는 체계를 갖출 수 있다.

셋째, 건물ID의 범용 연계키 기능 강화이다. 건물ID는 다양한 공공데이터를 건축물 단위로 통합하는 핵심 연계키로 기능할 잠재력을 지닌다. 이 잠재력을 실현하기 위해서는 인허가 처리 단계에서 건물ID의 고유성을 시스템 차원에서 보장하고, 데이터 간 연계 성공률을 높이기 위한 정비 작업이 병행되어야 한다. 건물ID가 안정적인 범용 연계키로 기능하게 되면, 건축행정 데이터 내부의 연계는 물론 화재·재해 데이터, 에너지 사용량 정보, 부동산 거래 데이터 등 타 공공데이터와의 연계 활용도도 크게 높아질 것이다. 이는 건축물 데이터가 단순 행정 기록을 넘어 다양한 사회 문제 해결에 기여하는 데이터 인프라로 성장하기 위한 기반이 된다.

정확한 데이터가 만드는 건축·도시의 미래

건축물 데이터의 오류는 오랜 시간에 걸쳐 형성된 시기별·지역별 행정 체계의 차이와 제도적 요인에 기인한 구조적 문제다. 이번 연구는 규칙 기반 진단만으로는 이 문제의 전모를 파악하기 어렵다는 점을 기계학습을 통해 확인하였다. 지역과 시기에 따라 오류 패턴이 다르다는 사실은 앞으로의 품질 정비가 전국 일괄 방식보다 지역별·항목별 맞춤형 접근을 취해야 함을 시사한다. 오픈소스 방법론 공개, 입력 시스템 개선, 건물ID 연계키 기능 강화가 동시에 이루어질 때 건축물 데이터 품질의 지속적 향상이 가능하다. 품질이 확보된 건축물 데이터가 공공·민간에 폭넓게 활용될 때, 정확한 통계에 기반한 과학적 행정과 더 나은 건축·도시 정책이 비로소 가능할 것이다.

참고문헌

1 송유미, 조영진, 안의순. (2024). 건축행정 데이터를 활용한 건축물 연령 지표 개발 연구. 건축공간연구원.
2 안의순, 박종훈. (2023). 건축행정 데이터 기반 재해 취약 지하층 주택 현황 분석. 건축공간연구원.
3 안의순, 허한결, 남기천, 강범준, 이선재, 박동준. (2025). 인공지능을 활용한 건축물 데이터 품질 고도화 방향 연구. 건축공간연구원.
4 조영진, 허한결, 안의순, 류수연, 송유미, 현태환. (2022). 빅데이터 기반 건축물 화재 예측 모델 개발 연구. 건축공간연구원.
5 조영진, 허한결, 송유미, 현태환. (2023). 빅데이터 기반 건축물 화재 및 홍수 리스크 분석 모델 개발 연구. 건축공간연구원.
6 조영진, 안의순, 박성남, 고영호, 권오규, 임보영, 임리사, 김유진, 이정현, (2024a). 범죄예방 환경설계(CPTED) 고도화 및 인증제도 개선 방향. 건축공간연구원.
7 조영진, 유광흠, 박종훈, 안의순, 허한결, 현태환, 송유미, 김효정, 남기천, 김가혜, 박미래. (2024b). 2024년 건축물관리지원센터 업무 위탁. 국토교통부, 건축공간연구원.