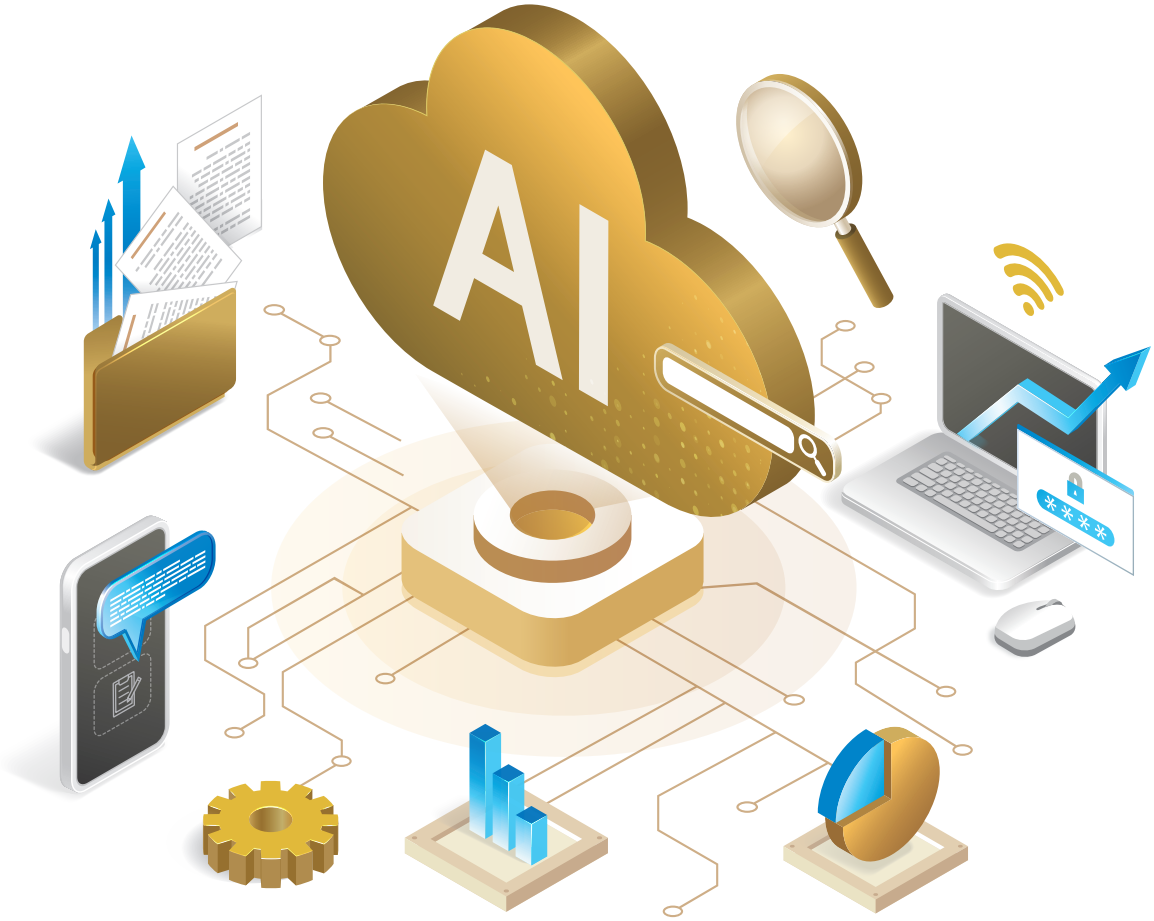
공공데이터의 인공지능 친화적 관리 가이드라인

2026. 03



행정안전부

NIA 한국지능정보사회진흥원

가이드라인의 개정 이력

연번	버전	발간일	주요 내용
1	V1.0	2026.03.31	▶ 최초 발간 공공데이터의 인공지능 친화적 관리 가이드라인 제시

가이드라인 활용 안내

1. 목적

- 행정안전부는 인공지능 활용을 고려한 공공데이터 관리체계를 구축하기 위해 본 가이드라인을 마련했습니다. 공공부문 전반에 적용 가능한 공통 기준을 제시하여, 데이터의 상호운용성과 활용성을 높임으로써 데이터 기반 행정과 정책 활용의 효율성을 제고하는 동시에 인공지능의 공공데이터 활용을 촉진하고자 합니다.
- 본 가이드라인은 공공데이터를 단순히 공개하는 수준을 넘어 인공지능 학습·추론·분석에 활용 가능한 형태로 관리할 수 있도록 데이터 구조, 메타데이터, 문서화, 품질관리, 접근 방식, 책임성과 윤리 등 전반적인 관리 요소를 체계적으로 제시하고 있습니다. 이를 통해 공공데이터의 신뢰성과 활용 가능성을 동시에 확보하고, 다양한 인공지능 서비스 및 정책 활용의 기반을 마련하고자 합니다.

2. 가이드라인 적용 대상 및 범위

- **적용 대상:** 공공데이터 업무·실무 담당자 등 「공공데이터의 제공 및 이용 활성화의 법률」 (이하 '「공공데이터법」') 상 공공기관 데이터를 생산·관리·제공하는 담당자
- **적용 범위:** 「공공데이터법」에 제2조(정의) 제1호에 해당하는 모든 공공기관의 데이터 중 공공기관 간 공유되거나 국민에게 제공 또는 제공될 수 있는 형태로 처리된 자료 일체

3. 가이드라인 구성

구분	구성요소		
데이터 구조 및 접근	오픈 포맷 및 기계 이해 가능 구조		
	데이터 부분 조회 및 적시 접근 방식 제공		
속성정보 제공	메타데이터		
	문서화		
품질지표	완전성	정확성	유효성
	일관성	적시성	유일성
표준 코드			
관리체계	데이터 책임 주체 및 관리체계 수립		
	이용자 접근 편의성 인프라 구축		
	이용자 관점 데이터 평가 및 피드백 체계 구축		
	데이터 접근권한 및 활용체계		
	데이터 거버넌스와 책임 구조		
	법적 충돌 해소 및 데이터 보호 체계		

※ 가이드라인 내용의 이해도 향상을 위하여 '참고', '참조' '예시' 박스를 통해 참고할 수 있는 내용 제공

구분	내용
참고	▶ 권고 방향, 참고할 수 있는 타 기관 또는 기업의 사례, 방법론 등 내용 관련 가이드라인 이해와 이행을 위해 도움이 될 수 있는 내용 제시
예시	▶ 가이드라인 적용 예시로 공공데이터 담당자의 실무 적용 용이성을 위한 내용 제공
참조	▶ 관련 지침, 매뉴얼과 연계되는 내용 제시 - 「공공데이터 관리지침」, 「공공데이터 품질관리 매뉴얼」, 「공공데이터 제공관리 실무 매뉴얼」, 「공공데이터베이스 표준화 관리 매뉴얼」 등

4. 활용시 유의사항

- 본 가이드라인은 공공데이터의 인공지능 활용 기반을 마련하는 데 필요한 기준을 제시하지만, 모든 데이터와 공공문서가 인공지능에 의해 즉시 활용 가능한 상태로 전환되는 것을 보장하지는 않습니다.
- 행정안전부는 공공데이터의 인공지능 활용을 대비한 출발로서 기술적으로 접근하기 용이한 구조화된 데이터 중심의 기준을 일차적으로 마련하였습니다. 비정형 데이터에 대해서도 파일 형식, 해상도 등 일부 기술적 조건을 제시하고 있으나 문서의 의미를 인공지능이 직접 이해할 수 있도록 하는 구조화, 텍스트화, 의미 단위 분해 등에 대한 구체적인 기준은 충분히 포함하고 있지 않습니다.
- 따라서 생성형 인공지능이 공공문서를 읽고 이해하기 위해서는 본 가이드라인이 제시하고 있는 기준 외에 독점 포맷의 변환, 텍스트 추출, 데이터 구조화, 활용 목적에 맞는 재가공 등 추가적인 처리 과정이 필요할 수 있습니다.
- 아울러, 본 가이드라인은 법령 또는 별도 규정에 따른 의무사항이 아니므로 대상 기관이 일률적으로 준수해야 하는 것은 아닙니다. 각 공공기관은 개별 데이터의 특성, 운영 여건 및 정책 환경을 고려하여 본 가이드라인을 참고 기준으로 삼아 합리적인 범위에서 자율적으로 적용할 수 있습니다.
- 본 가이드라인은 특정 시점의 기술과 정책 환경을 반영하여 작성한 한계점이 있습니다. 행정안전부는 하루가 다르게 변화하고 있는 인공지능 기술과 데이터 활용 방식에 대응하기 위해 지속적으로 본 가이드라인을 보완하고 개정해 나갈 계획입니다.



목 차

CHAPTER 01. 개 요 08

1.1 배경 및 필요성 08
1.1.1 인공지능 확산과 데이터 패러다임 전환 08
1.1.2 공공데이터의 인공지능 친화적 관리의 당위성 08
1.1.3 가이드라인 필요성 09
1.1.4 가이드라인 목적 10
1.2 적용 대상 및 범위 10
1.3 기본 개념 11
1.3.1 인공지능 친화적 공공데이터 개념 11
1.3.2 공유 데이터 자격 기준 13
1.4 가이드라인의 개발 방법론 15
1.5 가이드라인의 구성 16

CHAPTER 02. 인공지능 친화적 공공데이터 세부기준 17

2.1 데이터 구조 및 접근 체계 17
2.1.1 오픈 포맷, 기계 이해 가능 구조 17
2.1.2 데이터 부분 조회 및 적시 접근 방식 제공 19
2.2 속성정보 제공 21
2.2.1 메타데이터 21
2.2.2 문서화 31
2.3 품질 지표 34
2.3.1 적용범위 34
2.3.2 인공지능 친화적 공공데이터 품질 지표 34
2.3.3 AI 모델 기반 데이터 품질 평가 44
2.4 표준 코드 45

CHAPTER
03. 인공지능 친화적 공공데이터 관리원칙 47

3.1 데이터셋 책임 주체 및 관리체계 수립 47
3.1.1 데이터 검증 절차 및 공유 품질 기준 48
3.1.2 데이터 유지관리 절차 50
3.1.3 데이터 생애주기 및 버전 관리 50
3.1.4 데이터 보안 및 윤리 절차와 기준 마련 51
3.2 이용자 접근 편의성 인프라 구축 52
3.3 이용자 관점 데이터셋 평가 및 피드백 체계 구축 52
3.4 데이터 접근 권한 및 활용 체계 54
3.5 데이터 거버넌스와 책임 구조 55
3.6 법적 충돌 해소 및 데이터 보호 체계 56
3.6.1 법률 간 충돌 조정 체계 56
3.6.2 데이터 보호 및 권리 준수 57

부록1 인공지능 친화적 공공데이터 관리 원칙 60
부록2 주요 용어 정의 61
부록3 메타데이터 예시 64
부록4 인공지능 친화적 공공데이터 관리 체크리스트 66
부록5 인공지능 친화적 공공데이터 데이터 카드 상세 69

CHAPTER 01

개 요



1.1 배경 및 필요성

1.1.1 인공지능 확산과 데이터 패러다임 전환

- 인공지능(AI) 전환이 세계적으로 본격 가속화됨에 따라, 데이터는 국가 경쟁력을 좌우하는 국가의 차세대 핵심 인프라로 인식
  - 각국 정부는 AI·데이터 기반 행정·정책을 국가 전략 차원에서 추진하고, AI를 공공서비스 혁신과 산업 경쟁력 강화를 위한 핵심 수단으로 적극 활용
  - 우리나라도 「대한민국 인공지능 행동계획」 수립 추진을 통해 AI 혁신 생태계 조성, 범국가 AI 기반 대전환, 글로벌 AI 기본 사회 기여 등을 핵심 정책으로 제시[1)]
- 인공지능 모델의 성능과 신뢰성 확보를 위해, 학습 데이터와 참조 데이터의 품질, 구조, 대표성, 거버넌스를 체계적으로 확보·관리하는 것이 중요
  - 대규모 언어모델(LLM)과 생성형 AI의 등장으로, 공공분야에서도 민원 응대, 정책 설명, 법령 질의응답, 행정문서 초안 작성, 데이터 분석 지원 등 다양한 활용 사례가 빠르게 확산되는 양상
  - 공공데이터 또한 단순 개방의 단계를 넘어 모델 학습·추론에 활용 가능한 수준으로 준비된 '인공지능 친화적 공공데이터'를 체계적으로 구축·관리하고 제공하기 위한 가이드라인 확립 필요

1.1.2 공공데이터의 인공지능 친화적 관리의 당위성

- **혁신 생태계 조성 및 효율성 증대:** 인공지능 친화적 공공데이터를 확충하여 공공 및 민간의 AI 서비스 개발·검증 환경 지원 및 공공부문 내부 효율성 제고 등 국가 AI 혁신 생태계 기반 조성
- **정책 책임성 및 설명 가능성:** 공공부문 AI 의사결정에 활용되는 데이터의 출처, 품질, 편향성, 한계점 등을 체계적으로 관리하고, 설명함으로써 행정 결정의 책임 소재를 명확히 하고 국민의 기본권과 안전을 도모
- **개인정보 보호 및 신뢰성 강화:** 공공데이터를 AI에 활용 시 개인정보의 비식별화·가명 처리를 철저히 수행하고, 이용 목적·범위와 라이선스 체계를 사전에 정립하여 사회적 신뢰성을 제고하고 국민의 기본권을 보호
- **공정성 및 포용성 제고:** 데이터 수집·정제·가공 단계에서 특정 집단이나 계층이 과소·과대 대표되지 않도록 데이터 대표성을 관리하고 편향성을 최소화하여 공정성과 포용성 확보

1 국가인공지능전략위원회. (2025.12). 대한민국 인공지능 행동계획(안)

1.1.3 가이드라인 필요성

- 공공데이터 정책은 기존의 개방·표준화 중심에서 AI 모델이 쉽게 활용할 수 있는 '인공지능 친화적 공공데이터' 구축 중심으로 확장 필요
- AI 활용 관점의 필수 정보가 체계적으로 관리·제공될 수 있는 범정부 차원의 관리 가이드라인 필요
  - 기관별로 관리체계, 품질 지표, 메타데이터 관리 수준의 차이로 인해 발생하는 데이터 전처리 비효율을 해소하고, 데이터 관리 책임·역할 분담을 명확히 하기 위한 공통 기준 필요
  - 공공데이터 개방 정책 및 관련 지침을 기반으로, 현행 체계를 AI 환경에 부합하도록 확장·고도화 필요
- 미국, 영국, OECD, 국제 데이터 기관(ODI, Open Data Policy Lab 등)은 인공지능 친화적 데이터 중요성을 제시하며, 데이터를 AI 활용에 용이한 형태로 전환하기 위한 정책·가이드라인[2]을 마련하는 추세
- 국제 동향과의 정합성을 확보하는 동시에, 공공데이터 영역에서 인공지능 친화적 수준의 관리 가이드라인을 선제적으로 구축함으로써 글로벌 AI 경쟁력을 확보

[표 1] 인공지능 친화적 관련 국외 정책 및 가이드라인

구분	기관/국가	문서명	핵심 내용
미국	상무부 (DoC)	Generative Artificial Intelligence and Open Data: Guidelines and Best Practices (2025)	▸AI에 최적화된 개방형 데이터 게시를 위한 구체적 가이드라인(문서화, 메타데이터, 저장·배포, 라이선스, 품질·무결성 등) 제시
미국	상무부·연방정부	AI and Open Government Data Assets RFI (2024)	▸인공지능 친화적 개방형 데이터 자산 정의 및 구현 방안에 대한 정부·민간·학계 의견 수렴
영국	영국정부	AI Opportunities Action Plan (2025)	▸정부의 데이터가 AI 발전 핵심 자원이며, 데이터의 수집·구조·품질을 AI 활용에 적합하도록 고도화 해야 할 필요성 시사
국제 연구기관	ODI	The UK government as a data provider for AI (2025)	▸정부의 데이터 공급자 역할 강조 및 AI 학습·활용을 전제로 한 데이터 준비도(Readiness) 고도화
		A framework for AI-ready data (2025)	▸인공지능 친화적 데이터의 핵심 요소와 구현 프레임워크 제시
국제기구	OECD	Government at a Glance 2025 -Open Government Data (2025), Digital government and Innovation(2025)	▸개방형 데이터와 AI를 결합한 디지털 정부 고도화, 인공지능 친화적 데이터 생태계 구축 필요성 강조
국제 연구기관	Open Data Policy Lab	A Blueprint to unlock new data commons for AI (2025)	▸공익 목적 AI를 위한 데이터 공동 활용 체계 제시, 고품질 인공지능 친화적 공공데이터의 공동 거버넌스 체계 구축

2 U.S. Department of Commerce. (2025). Generative Artificial Intelligence and Open Data: Guidelines and Best Practices ; U.S. Department of Commerce. (2024). AI and Open Government Data Assets - Request for Information (RFI) ; UK Government. (2025). AI Opportunities Action Plan ; Open Data Institute. (2025). The UK Government as a Data Provider for AI ; Open Data Institute. (2025). A Framework for AI-ready Data. ; OECD. (2025). Government at a Glance 2025: Digital government and Innovation (2025), OECD Publishing ; OECD. (2025). Governing with AI, OECD Publishing ; Open Data Policy Lab. (2025). A Blueprint to unlock new data commons for AI

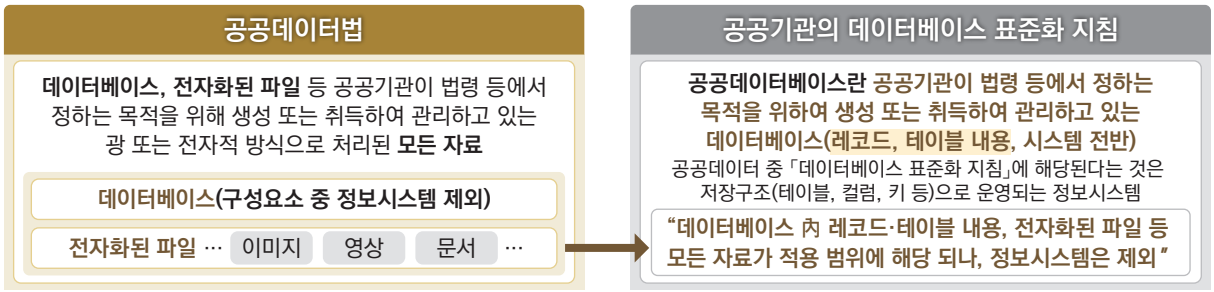
1.1.4 가이드라인 목적

- 공공부문 전반에 적용 가능한 공통 관리 기준을 마련하여 데이터가 기관별로 상이한 기준으로 구축·관리되어 발생하는 연계·통합·재사용의 문제 해소
- 인공지능 친화적 공공데이터의 개념·기준을 명확히 정의하고, 품질·문서화·메타데이터·거버넌스·보호·윤리 기준 등을 제시함으로써, 공공부문 전반의 인공지능 친화적 공공데이터 관리 역량을 체계적으로 향상

1.2 적용 대상 및 범위

적용 대상	공공데이터 업무·실무 담당자 등 「공공데이터의 제공 및 이용 활성화의 법률 (이하 「공공데이터법」) 상 공공기관 데이터를 생산·관리·제공하는 담당자를 본 가이드 주요 활용 대상으로 설정
적용 범위	「공공데이터법」에 제2조(정의) 제1호에 해당하는 모든 공공기관의 데이터 중 공공기관 간 공유되거나 국민에게 제공 또는 제공될 수 있는 형태로 처리된 자료 일체

[그림 1] 인공지능 친화적 공공데이터 관리 가이드라인의 적용 범위



참조 공공데이터의 제공 및 이용 활성화에 관한 법률 제2조(정의)

제2조(정의) 이 법에서 사용하는 용어의 뜻은 다음과 같다. <개정 2016. 1. 6., 2020. 6. 9.>

1. "공공기관"이란 국가기관, 지방자치단체 및 「지능정보화 기본법」 제2조 제16호에 따른 공공기관을 말한다.
2. **"공공데이터"란 데이터베이스, 전자화된 파일 등 공공기관이 법령 등에서 정하는 목적을 위하여 생성 또는 취득하여 관리하고 있는 광(光) 또는 전자적 방식으로 처리된 자료 또는 정보**로서 다음 각 목의 어느 하나에 해당하는 것을 말한다.
   가. 「전자정부법」 제2조 제6호에 따른 행정정보
   나. 「지능정보화 기본법」 제2조 제1호에 따른 정보 중 공공기관이 생산한 정보
   다. 「공공기록물 관리에 관한 법률」 제20조 제1항에 따른 전자기록물 중 대통령령으로 정하는 전자기록물
   라. 그 밖에 대통령령으로 정하는 자료 또는 정보 … (생략)

1.3 기본 개념

1.3.1 인공지능 친화적 공공데이터 개념

- 인공지능 친화적 공공데이터란 인공지능 모델의 학습·추론·평가에 쉽게 활용될 수 있도록 위험 최소화 및 기본권 보호를 전제로 기계 이해성 및 품질이 다음과 같이 확보된 수준의 공공데이터를 의미
  - 본 가이드라인에서 제시하는 인공지능 친화적 요건은 공공데이터가 AI 활용을 전제로 구조·메타데이터·문서화·품질 정보를 갖추도록 준비된 관리 수준을 의미
    - '인공지능 친화적' 상태는 접근성과 상호운용성(표준 포맷, API, 메타데이터), 문서화와 거버넌스(출처·처리 과정·이용 권한 범위), 데이터 품질, 특정 AI 활용 목적에 대한 적합성 등을 갖춘 수준을 의미
- 인공지능 친화적 공공데이터는 다음과 같은 핵심 특성을 포함

[표 2] 인공지능 친화적 공공데이터의 핵심 특성

핵심 특성	내용
기계가공 및 상호운용 가능 기술 구조	▸ CSV, Parquet, JSON, GeoJSON 등 기계 이해가 가능한 개방형 포맷으로 제공하고, 명시적인 스키마(필드 정의, 단위, 코드값)를 제공 ▸ API를 통해 대용량 데이터에 안정적으로 접근할 수 있는 체계를 확보하고, 버전·배포 정보 및 식별자(영구 식별자, URI 등) 부여를 통해 데이터의 연속성과 상호운용성 보장
메타데이터와 문서화	▸ 데이터셋 및 속성 단위의 상세 설명, 생성·수집·정제·가공 과정, 어노테이션 기준, 품질 진단 방법과 지표를 메타데이터 및 부속 문서로 제공 ▸ AI 시스템 개발 시 해당 데이터의 적합성을 사전에 진단할 수 있도록 활용 목적, 기술적 제한사항, 권장·비권장 활용 사례 명시
내용 품질과 대표성	▸ 정책·서비스에서 대상이 되는 인구·지역·현상을 균형 있게 반영하여 데이터의 대표성을 확보하고, 이와 동시에 원천 데이터와의 정보 일치성 및 데이터의 최신성, 무결성을 확보 ▸ 데이터의 한계·불균형·잠재적 편향을 메타데이터와 문서에 명시하여 AI 모델의 편향·차별·유해 결과를 최소화
권리·윤리·보안·책임 거버넌스	▸ 라이선스, 저작권, 개인정보·민감정보 처리, 재사용 조건 등을 명확히 표시하고, 관련 법령 및 AI 윤리 원칙에 부합하도록 설계 ▸ 데이터 변경 이력과 책임 주체의 추적 가능성을 확보하고, 데이터 오남용·위험 발생 시 대응 절차와 연락 창구를 지정

- 인공지능 친화적 공공데이터 개념은 기존 공공데이터의 관리 기준에서 수준이 확장된 개념으로, 확장된 사항은 아래와 같음

[표 3] 기존 공공데이터 관리 기준에서 확장된 인공지능 친화적 공공데이터 개념

구분		내용
공개 목적	기존	▸ 투명성 제고, 국민 알 권리 보장 및 데이터 2차 활용 촉진
	확장	▸ AI 모델의 학습·추론·평가에 쉽게 활용 가능한 상태로 데이터를 제공
데이터 품질	기존	▸ 데이터 오류 검수, 최신성 및 원천데이터와의 정합성 관리에 중점
	확장	▸ AI 모델의 학습 적합성 및 추론 성능 제고를 위해 정확성, 완전성, 대표성 등을 핵심 품질 요소로 설정하여 데이터 활용 역량 강화
데이터 구조	기존	▸ CSV, XML 등 기계 이해가 가능한 포맷 준수, 공공데이터 개방 표준에 따른 스키마 정보 제공
	확장	▸ 표준 코드 적용, 대량 처리 최적화 구조 확보 등 기계 이해성 및 처리 효율성 극대화를 위한 구조 설계 강화
접근 방식	기존	▸ 공공데이터포털 등 국가 플랫폼을 통한 개별 파일 다운로드 및 API 제공
	확장	▸ 대량 다운로드 지원, 버전 관리 및 영구 식별자 부여를 통한 접근 효율성 및 데이터 재현성 확보
메타 데이터	기존	▸ 제목, 요약 설명, 일부 필드 설명 등 기본적인 데이터 속성 정보 제공
	확장	▸ 생성·수집·정제·가공 과정, 품질 점검 방법·지표, 편향·한계 등 상세 데이터 정보 제공
권리·라이선스	기존	▸ 공공누리 등 표준 라이선스 표기 및 소유권 중심 이용 권한 관리
	확장	▸ 라이선스, 저작권, 개인정보·민감정보 처리 원칙을 명확히 명시
위험·윤리 관리	기존	▸ 법령 및 보안 규정 준수 중심 관리
	확장	▸ AI 편향·차별·유해성 방지를 위한 사전 위험 평가 및 관련 가이드라인 제시

1.3.2 공유 데이터 자격 기준

- 데이터 개방에 따른 리스크를 최소화하고 기관 간 신뢰 기반의 공유 환경을 조성을 위해 인공지능 친화적 공공데이터를 플랫폼에 등록하고 공유 시 데이터의 민감도와 법적 제한 사항을 사전에 검토할 수 있는 절차 필요 (* 공공데이터 공유의 의미는 데이터 개방까지 고려한 용어 확대 개념)
- 모든 공유 데이터는 등록 전 데이터에 포함된 데이터의 성격에 따라 아래와 같이 3단계로 등급을 분류하고 관리 가능
  - ▸ 데이터의 등급 분류는 향후 국정원의 보안 가이드라인(N2SF) 기준에 따라 변경 가능

예시 데이터 등급별 공유범위

등급	구분	정의	공유범위
1등급(O)	공개공유	개인정보 및 보안상 문제가 없음	모든 기관 및 공공데이터포털
2등급(S)	제한공유	비식별 처리를 통해 안전성 확보	승인된 기관 간 한정
3등급(C)	내부전용	비식별 처리 후라도 보안상 외부 공유 부적절	해당 기관 내부 한정

- **공유 불가 데이터 명시:** 데이터의 성격이 다음에 해당하는 경우, 공공데이터의 안전한 관리와 국민의 기본권 보호를 위해 플랫폼 등록 및 공유가 엄격히 제한
  - **고유식별정보 포함:** 「개인정보 보호법」상 주민등록번호, 여권번호 등 개인을 고유하게 식별할 수 있는 정보가 포함된 경우
  - **국가보안사항:** 「국가보안법」에 따라 비밀 또는 비공개로 분류되어 보안 유지가 필요한 경우
  - **제3자 제공 금지:** 의료법, 금융정보법 등 개별 법률에서 데이터의 제3자 제공을 명시적으로 금지하고 있는 경우
  - **재식별 위험 잔존:** 비식별 처리를 수행하였음에도 불구하고 기술적으로 개인 식별 가능성이 남아있는 경우

참조 개인정보 보호법 제28조 4,5[2025. 10.]

「개인정보 보호법」 제28조의4(가명정보에 대한 안전조치의무 등)

1. 개인정보처리자는 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 경우에는 원래의 상태로 복원하기 위한 추가 정보를 별도로 분리하여 보관·관리하는 등 해당 정보가 분실·도난·유출·위조·변조 또는 훼손되지 않도록 대통령령으로 정하는 바에 따라 안전성 확보에 필요한 기술적·관리적 및 물리적 조치를 하여야 한다.
2. 개인정보처리자는 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 경우 처리목적 등을 고려하여 가명정보의 처리 기간을 별도로 정할 수 있다.
3. 개인정보처리자는 제28조의2 또는 제28조의3에 따라 가명정보를 처리하고자 하는 경우에는 가명정보의 처리목적, 제3자 제공 시 제공받는 자, 가명정보의 처리 기간(제2항에 따라 처리 기간을 별도로 정한 경우에 한한다) 등 가명정보의 처리 내용을 관리하기 위하여 대통령령으로 정하는 사항에 대한 관련 기록을 작성하여 보관하여야 하며, 가명정보를 파기한 경우에는 파기한 날부터 3년 이상 보관하여야 한다.

「개인정보 보호법」 제28조의5(가명정보 처리 시 금지의무 등)

1. 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 자는 특정 개인을 알아보기 위한 목적으로 가명정보를 처리해서는 아니 된다.
2. 개인정보처리자는 제28조의2 또는 제28조의3에 따라 가명정보를 처리하는 과정에서 특정 개인을 알아볼 수 있는 정보가 생성된 경우에는 즉시 해당 정보의 처리를 중지하고, 지체 없이 회수·파기하여야 한다.

- **비식별 처리 수준 기준:** 다음과 같은 비식별 처리 기술 적용과 적정성 평가를 통해 데이터를 공유 가능하도록 처리(※ 개인정보보호위원회 「가명정보 처리 가이드라인」(2026.3.) 참고)
  - **기술적 보호조치:** 준식별자 범주화, 직접식별자 마스킹, k-익명성 확보 등을 통해 충분한 수준의 비식별성 확보
  - **적정성 검토:** 「개인정보 보호법」 및 같은 법 시행령에 따라 가명정보 처리 시 기관은 비식별 처리의 적절성 및 재식별 위험 가능성을 확인하기 위한 자체 적정성 검토를 수행

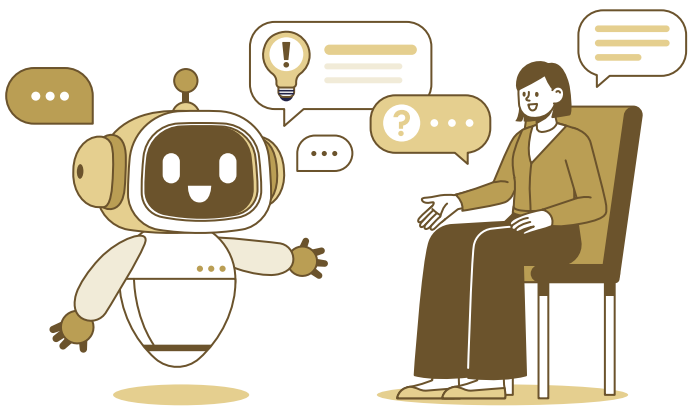
참조 가명정보 처리 가이드라인[2026. 3.]

「가명정보 처리 가이드라인」 (제2장 가명처리 절차)

- 가명정보 특례 제도에 따라 수행하는 가명처리는 단순히 개인정보 일부를 가리는 기술적 처리만을 의미하는 것이 아니라, 목적 설정, 가명처리, 검토, 사후 관리까지 재식별 위험을 낮추고 높아지지 않도록 적정 수준으로 처리·관리하는 일련의 과정을 의미
- 개인정보의 가명처리 절차는 다음과 같은 5단계 절차를 거쳐 이루어짐

1. **사전준비:** 왜 가명정보를 활용하려고 하는지를 정리하는 단계- 처리 목적을 설정·검토하고, 해당 목적에 필요한 가명처리 대상을 선정
2. **위험성 검토:** 가명처리 전, 가명정보 처리의 위험성을 미리 검토하는 단계- 가명정보 활용 시의 위험성을 먼저 검토하여 어떤 방식과 수준으로 가명처리를 수행할 것인지를 판단
3. **가명처리:** 위험성 검토 결과를 바탕으로 실제 가명처리를 수행하는 단계- 처리 목적을 달성할 수 있으면서도, 재식별 위험을 통제 가능한 수준으로 낮추는 방향으로 가명처리 수행
4. **적정성 검토:** 가명처리가 적절하게 수행되었는지 검토하는 단계- 처리 목적이 적절한지, 가명처리 이용항목이 잘 선정되었는지, 위험도 판단 및 가명처리가 적정하게 수행되었는지, 해당 가명정보로 목적 달성이 가능한지 등을 종합적으로 검토
5. **안전한 관리:** 가명정보를 활용하면서 재식별 위험을 지속 관리하는 단계

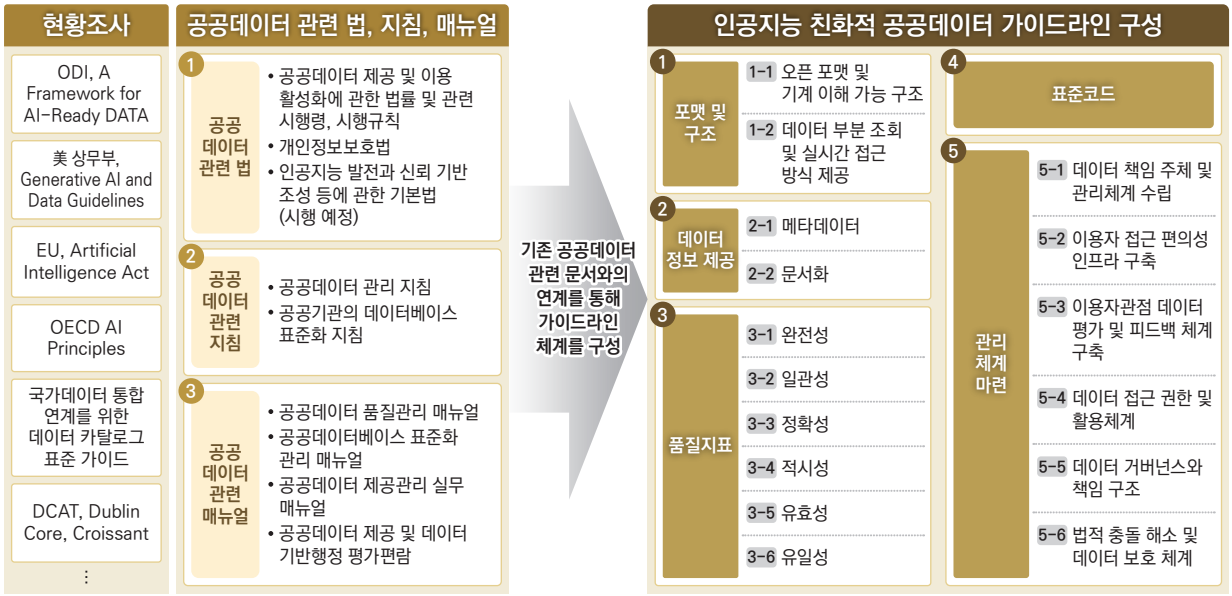
- **기관 판단 보조 도구:** 담당자의 신속한 의사결정을 돕기 위한 보조 도구를 도입
  - **자가진단 체크리스트:** 플랫폼 내 탑재된 체크리스트를 통해 공유 가능 여부를 실무자가 직접 1차 판정 가능
  - **사전 유권해석:** 자격 판단이 곤란한 데이터에 대해서는 행정안전부에 공식적인 유권해석 요청 절차 운영
  - **평가 환류:** 기관별 연간 데이터 등록 및 거부 현황을 환류하여 적극적인 데이터 공유 유도



1.4 가이드라인 개발 방법론

- 공공데이터 개방 정책 및 관련 지침 등 현행 기반을 AI 환경에 부합하도록 확장하기 위하여 가이드라인을 마련하고, 국제 표준과 정합성을 유지하면서 국내 공공데이터 실무자가 적용할 수 있도록 내용을 구성
- 인공지능 친화적 데이터 관련 국외 정책 동향, 국제 표준 및 학술 연구 현황을 검토하여 가이드라인 주요 구성 내용을 도출하였으며, 국내 공공데이터 관련 법령, 지침, 매뉴얼과 연계될 수 있도록 구성
  - **현황 분석:** 미국, EU, 영국 등 주요국에서 시행 중인 AI 관련 법령 및 지침과 ODI, OECD 등 국제기구의 가이드라인을 분석하여, '기계 이해가 가능한 포맷 및 구조', '상세 메타데이터를 통한 데이터 정보 제공' 등 핵심 구성 요소를 도출
  - **표준 연계:** DCAT, Dublin Core, Croissant 등 데이터 정보 제공과 관련된 국제 표준과 과학기술정보통신부에서 발간한 「국가 데이터 상호운용성을 위한 데이터 카탈로그 표준 가이드」 등을 준용하여 인공지능 친화적 공공데이터 가이드라인이 국제 표준 및 국내 정책 가이드와 유기적으로 연계될 수 있도록 설계
  - **공공데이터 관련 매뉴얼과의 연계:** 「공공데이터법」 및 관련 지침·매뉴얼과의 정합성을 확보하고, 기존 관리체계와 인공지능 친화적 요건 간의 연계성을 명확히 파악할 수 있도록 주요 항목별로 상세 참조 정보를 병행 표기
- 가이드라인 체계는 글로벌 현황 분석 결과와 국제 표준에서 공통적으로 제시하는 핵심 요소를 주요 축으로 구성하고, 각 항목에 대한 상세 설명과 실무 적용 예시를 함께 제시
  - 데이터 구조 및 접근 체계, 속성정보 제공, 품질 지표, 표준 코드 등 기술적 요소와 운영 및 관리를 위한 관리 기준, 윤리 및 법적 책임 관련 사항을 주요 축으로 구성
- 가이드라인의 주요 영역별로 진단 체크리스트를 제공하여, 기관별 인공지능 친화적 공공데이터 수준을 점검하고 품질 개선 및 환류 체계를 마련할 수 있도록 구성

[그림 2] 가이드라인 개발 방법론



1.5 가이드라인의 구성

✔ 인공지능 친화적 공공데이터 관리 가이드라인은 데이터 구조 및 접근 체계, 속성정보 제공, 품질지표, 표준 코드 등 영역별로 공공데이터의 기술적 기준과 이를 위한 기관의 관리원칙을 제시

[표 4] 가이드라인 구성

구분				원칙
데이터 구조 및 접근	오픈 포맷 및 기계 이해 가능 구조			▶ 데이터는 개방형 방식으로 제공되어야 하며, 기계 이해가 가능한 형식으로 구조화되어야 한다.
	데이터 부분 조회 및 적시 접근 방식 제공			▶ 데이터는 부분 조회가 가능하여야 하며, 실시간 또는 준실시간 접근 및 활용을 지원하여야 한다.
속성정보 제공	메타데이터			▶ 데이터의 구조와 의미를 설명하는 포괄적이고 기계가 이해 가능한 메타데이터를 제공하여야 한다.
	문서화			▶ 데이터 구조, 활용 목적, 편향성 및 한계점, 버전 및 변경 이력을 투명하게 문서화하고 공개하여야 한다.
품질 지표	완전성	정확성	유효성	▶ 인공지능 활용에 적합한 공공데이터 품질을 지속적으로 점검하고 개선하여야 한다.
	일관성	적시성	유일성	
표준 코드				▶ 기관 간 데이터 연계성과 상호운용성을 확보하기 위하여 표준 코드를 일관되게 적용하여야 한다.
관리 체계	데이터 책임 주체 및 관리체계 수립			▶ 데이터의 품질·보안·윤리·업데이트에 대한 책임 주체와 관리 체계를 명확히 지정하여야 한다.
	이용자 접근 편의성 인프라 구축			▶ 사용자 중심의 데이터 인프라를 구축·운영하여 데이터에 대한 접근성을 강화하여야 한다.
	이용자 관점 데이터 평가 및 피드백 체계 구축			▶ 사용자 평가 및 피드백 체계를 운영하고, 그 결과를 데이터 품질 개선 과정에 반영하여야 한다.
	데이터 접근 권한 및 활용체계			▶ 데이터 접근 권한을 차등 관리하고, 데이터에 대한 보안을 강화하여야 한다.
	데이터 거버넌스와 책임 구조			▶ 데이터의 생산부터 활용까지 전 주기에 대한 책임 소재를 명확히 하여야 한다.
	법적 충돌 해소 및 데이터 보호 체계			▶ 데이터 생애주기 전반에 윤리적 법적 기준을 준수하고, 데이터 무결성과 신뢰성을 확보하여 공공의 이익을 보호하여야 한다.

✔ 가이드라인 내용의 이해도 향상을 위하여 '참고', '참조' '예시' 박스를 통해 참고할 수 있는 다양한 내용을 제공

[표 5] 가이드라인 내용 구성

구분	내용
참고	▶ 권고 방향, 참고할 수 있는 타 기관 또는 기업의 사례, 방법론 등 내용 관련 가이드라인 이해와 이행을 위해 도움이 될 수 있는 내용 제시
예시	▶ 가이드라인 적용 예시로 공공데이터 담당자의 실무 적용 용이성을 위한 내용 제공
참조	▶ 관련 지침, 매뉴얼과 연계되는 내용 제시 - 「공공데이터 관리지침」, 「공공데이터 품질관리 매뉴얼」, 「공공데이터 제공관리 실무 매뉴얼」, 「공공데이터베이스 표준화 관리 매뉴얼」 등

CHAPTER 02

인공지능 친화적 공공데이터 세부 기준

2.1 데이터 구조 및 접근 체계

2.1.1 오픈 포맷, 기계 이해 가능 구조

비독점적 오픈 포맷을 기반으로 기계가 데이터를 자동으로 해석·처리할 수 있도록 표준화되어 구조화된 데이터 형식

- **오픈 포맷:** 국제 표준에 따라 기술 명세가 공개된 비독점적 포맷을 적용하여, 다양한 시스템 및 AI 프레임워크 간의 기술적 상호 운용성이 향상된 형태[3)]
  - 오픈 포맷은 비용이나 사용상의 제약 없이 다양한 소프트웨어로 처리할 수 있는 데이터 포맷 (공공데이터 관리지침 제2조제4호)
  - 유료 소프트웨어가 필요한 독점 포맷(예: .hwp, .xlsx)은 클라우드나 리눅스 기반의 GPU 서버에서 자동화된 처리가 어렵거나 별도의 변환 비용이 발생

참고 데이터별 권장 오픈 포맷

구분		형식
권장 포맷 (오픈 포맷)	텍스트	ODT, CSV, JSON, XML, Parquet, ORC 등
	지리정보	GeoJSON, GeoPackage, GeoTIFF 등
	이미지	PNG, JPG, JPEG, TIF, TIFF, SVG 등
	건설/건축	IFC, OBJ, GLTF, GLB 등
비권장 포맷 (벤더 종속)	텍스트	HWP, HWPX, DOC, DOCX, XLS, XLSX, PPT, PPTX 등
	이미지	AI, PSD 등

* **비권장 포맷:** AI 활용을 위한 1차 제공 형식으로 적합하지 않음을 의미하며, 원본 보존 또는 내부 행정 목적의 사용까지 제한하는 것은 아님

3 Open Data Institute. (2024). A Framework for AI-Ready Data. ODI Research

참고 [Parquet] 고성능 AI 분석을 위한 열 지향(Columnar) 저장 규격

Parquet

- Apache Parquet은 '컬럼형 저장(Columnar Storage)', '메타데이터 첨부', '다양한 데이터 타입(벡터 데이터 포함) 지원 기능', '고유한 압축' 등의 기능을 보유
- 컬럼형 저장방식은 AI/ML 시스템의 데이터 접근을 최적화하는 방식으로 AI/ML 모델은 학습 과정에서 데이터의 모든 행(Row)을 읽을 필요 없이, 특정 피처(Features)에 해당하는 열(Column)만을 선택적으로 접근할 수 있음
- 필요한 열만 선택적으로 읽어 들임으로써 입출력(I/O) 작업을 줄여 지연 시간을 감소시키고, 특히 클라우드 기반 환경에서의 학습 효율성을 극대화

기계 이해 가능 구조: 컴퓨터가 데이터의 구조와 의미(Semantics)를 해석하여 데이터의 '의미(뜻)'를 자동으로 연결, 식별, 분석하여 자동으로 처리할 수 없다면, AI 모델의 데이터 파이프라인 자동화에 부적합[4)]

- 데이터 제공 시 지리정보·관계형데이터·그래프 구조 데이터 등 구조적 의미 보존을 위해 계층형 구조를 유지하되, AI 활용을 위한 별도의 분석용 파생 데이터셋 제공을 권장
- 다만 기계가 해석하기 쉽도록 하기 위해 플랫형(Flat) 구조 선택도 가능, 명확한 스키마 검증과 AI 모델의 데이터 해석 효율성 극대화를 위해 속성 간 관계를 단순화한 2차원 구조로 구성

예시 동일한 데이터 적용 시 계층형(Hierarchical) 구조와 플랫형(Flat) 구조 차이

계층형 구조	플랫형 구조
▶ 각 개체는 여러 속성을 가질 수 있으며, 하위 데이터로 연결되는 참조를 포함할 수 있음 - 한 개의 속성 아래에 다른 속성 데이터가 연결되는 형식	▶ 각 행은 하나의 고유한 관측값을 의미 ▶ 각 열은 하나의 고유한 속성을 가짐 ▶ 각 셀에는 하나의 값만 들어감 ▶ 리스트, 배열 같은 복합 구조를 가지 않으며, 다른 데이터로 연결되지 않음
주문ID : A001 ㄴ 구매자: 김구매 ㄴ 주소: 서울시 ㄴ 구매 내용 ㄴ 목록: 펜 ㄴ 개수: 10개	김구매 | 서울시 | A001 | 펜 | 10개 김구매 | 서울시 | A001 | 연필 | 5개 김구매 | 서울시 | A001 | 색연필 | 3개

- 데이터 이해의 일관성을 위해 날짜, 숫자, 단위 등 동일한 속성의 데이터는 전 범위에서 동일한 포맷을 유지하고, 텍스트 형태의 값 대신 표준화된 코드(Code) 체계를 우선 적용

4 U.S. Department of Commerce. (2025). Generative Artificial Intelligence and Open Data: Guidelines and Best Practices. p.14

예시 **데이터 값 형식 국제 표준**

구분	국제 표준	형식
날짜 및 시간	ISO 8601	▸ YYYY-MM-DD(예: 2025-12-18) 또는 YYYY-MM-DDThh:mm:ssZ(UTC 시간 포함 시)
국가 코드	ISO 3166-1	▸ 알파벳 2자리(Alpha-2, 예: KR) 또는 3자리(Alpha-3, 예: KOR) 코드
통화 코드	ISO 4217	▸ 알파벳 3자리 코드 (예: 원화 KRW, 달러 USD)
언어 코드	ISO 639-1	▸ 알파벳 2자리 코드 (예: 한국어 ko, 영어 en)
지리 위치	WGS84 / ISO 6709	▸ 십진수 도(Decimal Degrees) 방식 (예: 위도 37.1234, 경도 127.5678)
측정 단위	SI	▸ m(미터), kg(킬로그램), s(초) 등 표준 기호 사용

▸ 데이터 공백 발생 시 오류가 아닌 결측치로 정확히 이해할 수 있도록 표준화되고 일관된 결측치 표기 방식을 적용 (예: NULL, N/A 등)

2.1.2 데이터 부분 조회 및 적시 접근 방식 제공

데이터 및 메타데이터의 일관된 제공과 데이터 접근성 확보를 위해 RESTful API 등 널리 활용되는 표준 기반 인터페이스를 적용한 제공 방식

✔ AI 시스템은 지속적인 업데이트를 통해 응답의 최신성, 관련성, 활용성을 유지해야 하므로 데이터가 적시에 업데이트될 수 있도록 개방형 표준 기반의 API를 사용하여 제공 필요

▸ 각 기관은 공공데이터의 유형 및 특성, 기관의 제공환경, 이용자의 활용 편의성 등을 고려하여 개방을 위한 제공방식을 결정· 관리해야 하며(관리지침 제9조), 갱신이 빈번한 데이터의 경우 최신 정보가 적시에 제공될 수 있도록 오픈 API 사용을 권고(제공관리 실무 매뉴얼 P.25)

✔ 데이터와 메타데이터를 연계·제공하여 이용자가 데이터의 구조와 의미를 직관적으로 파악할 수 있도록 API를 설계·운영 필요

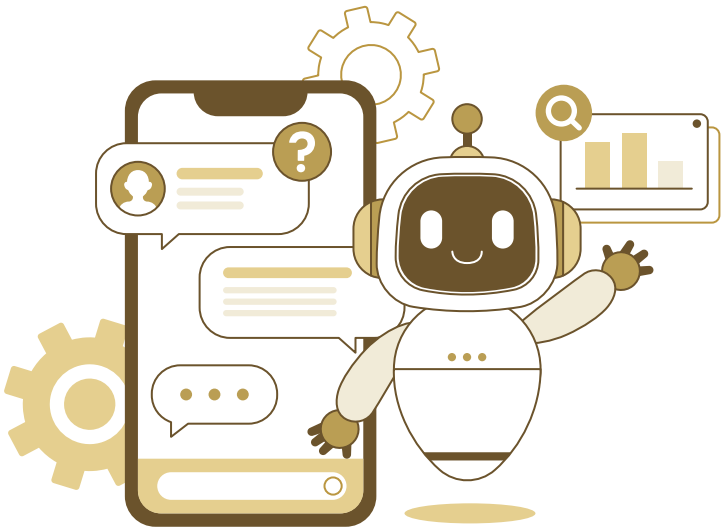
▸ RESTful 방식을 적용하여 모든 데이터를 일관된 규칙(URI, HTTP Method 등)으로 호출할 수 있도록 함으로써 상호운용성과 접근성을 확보

▸ 높은 요청 한도와 스트리밍 등 대량 데이터 전송 방식을 지원함으로써 대용량 데이터 처리 과정에서 병목 현상을 방지

▸ AI 시스템의 효율적 검색을 위해 API 호출 시 단순 값 제공에 그치지 않고, 스키마·필드 정의·의미 설명 등 메타데이터를 함께 제공하여 데이터 구조와 의미를 즉시 이해할 수 있도록 지원

참고 RESTful 적용을 통한 데이터 상호운용성 및 접근성 확보

- RESTful 방식: 모든 개체를 자원(Resource)으로 정의하여 URL로 식별하고, 조회·등록·수정·삭제 등의 행위는 HTTP 매서드로 표현하는 방식
- 예시: 특정 데이터셋의 메타데이터 조회 시 자원 경로와 메서드(GET)를 결합하여 호출
  - GET https://api.example.go.kr/datasets/traffic_001/metadata
- 표준화된 설계를 기반으로 하여 이해가 용이하고 시스템 확장이 유연함



2.2 속성정보 제공

2.2.1 메타데이터

공공데이터의 의미와 맥락을 정확히 전달하기 위한 핵심 요소로서, 데이터의 구조, 설명, 출처, 생성·수집 방식, 제약 사항 등을 구조화하여 제공함으로써 기계가 데이터를 자동으로 해석·처리할 수 있도록 하는 정보를 의미

- **고유식별자 부여:** 데이터·문서의 안정적 추적과 접근성 확보 및 이용자 피드백의 정확한 환류를 위해 고유식별자(PID)를 부여
  - ▸ 고유식별자(PID) 부여를 통해 데이터 개별 버전 및 연관 참조 정보를 영구적으로 제공함으로써 데이터 추적성을 확보하고 메타데이터 품질을 제고

  * 고유식별자(PID, Persitent Indentifier) 파일 이름, 파일 위치, 신규 버전 등이 만들어지더라도 동일한 데이터를 식별할 수 있도록 부여되는 변하지 않는 식별자

  - **고유성 및 일관성:** 단일 고유식별자(PID) 기반의 통합 관리를 통해 데이터 중복 및 혼동을 방지하고, 분산된 카탈로그 간 연계 시 데이터 식별의 일관성을 유지하여 최신 정보의 실시간 동기화를 지원
  - **연결성 및 참조 무결성:** 데이터-메타데이터-연관 문서 간의 유기적 관계를 정의하여 참조 무결성을 보장하고, 지식 그래프 탐색 및 자동 수집을 위한 지능형 참조 링크 체계 유지
  - **버전 추적 및 재현성:** 데이터 명칭이나 버전 변경 시에도 고유식별자를 통해 변경 이력을 관리함으로써 연구의 재현성을 확보하고 투명한 감사 추적 환경을 조성
- **메타데이터 제공:** 데이터 출처, 변수 정의, 변환 이력 등을 포함한 메타데이터를 제공하여 기계의 가독성과 이해도를 높임으로써 정확한 데이터 처리 및 모델 학습을 지원하고, 데이터 오해석에 따른 위험을 최소화
  - ▸ 데이터 제공자는 AI 활용 역량 강화를 위해 포괄적이고 구조화된 메타데이터를 데이터와 연계하여 제공
  - ▸ 국제 표준 기반의 메타데이터 체계를 적용하여 국가·기관·플랫폼 간 데이터 구조·의미의 동일성을 보장하고, 시스템 간 자동화된 데이터 교환·통합·검색을 지원함으로써 글로벌 상호운용성을 확보
    - 글로벌 상호운용성 및 AI 학습 도구와의 호환성 확보를 위해 DCAT, Dublin Core, Croissant 등 국제 표준 메타데이터 규격을 준용하여 적용[5)]

5 W3C, Data Catalog Vocabulary(DCAT) Ver.3; DCMI, DCMI Metadata Terms; MLCommons, Croissant Specification

[표 6] 메타데이터 체계 국제 표준

구분	설명
DCAT(Data Catalog Vocabulary)	▶ 플랫폼 간 데이터 카탈로그 공유 및 데이터 검색 용이성 확보를 위해 데이터셋 기술 방식을 규정하는 W3C 제정 표준 어휘
Dublin Core	▶ 자원 식별에 필요한 15개 핵심 요소를 정의하여 범용적 데이터 기술 및 관리를 지원하는 국제 표준(ISO) 규격
Croissant	▶ AI 모델과 학습 도구가 데이터를 즉시 처리할 수 있도록 데이터 구조와 활용법을 정의한 AI 데이터 전용 기술 규격

- 공공데이터 제공 및 이용 활성화 법률 시행규칙(이하 「공공데이터법 시행규칙」)의 '제공대상 공공데이터 등록서'와 '공공데이터 목록 등록서'의 주요 항목을 기본 메타데이터로 간주하고, 이를 기반으로 AI 활용 목적에 부합하는 상세 메타데이터를 확장 구성

[그림 3] 공공데이터 관련 메타데이터 AS-IS, TO-BE



참조 공공데이터법 시행규칙 - 공공데이터 목록 등록서 항목

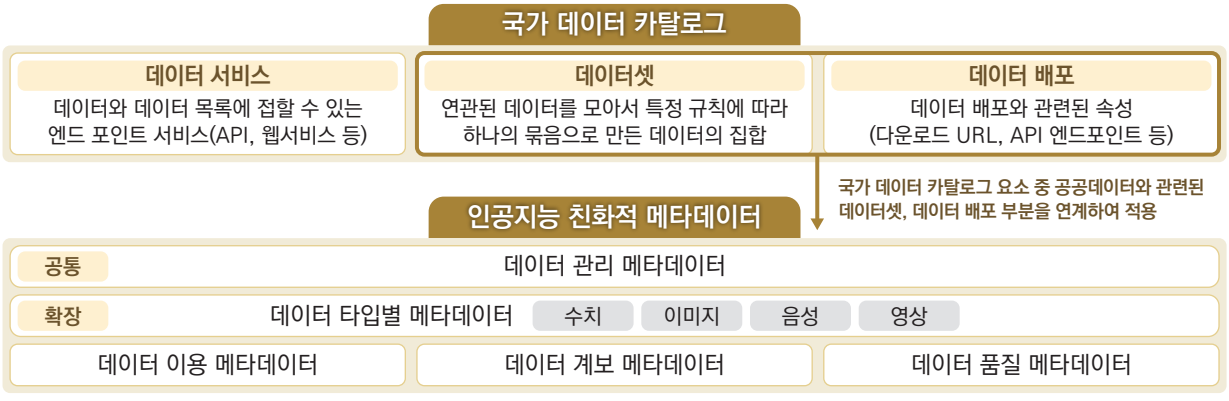
항목	내용(예시)
공공데이터 목록 명칭(국문)	▶ ○○시 연도별 인구 변화(1990~2025)
공공데이터 목록 명칭(영문)	▶ ○○ City Annual Population Data (1990-2025)
정부기능 분류체계(BRM)	▶ 일반공공행정, 교육, 보건, 환경 등
보유근거(법령)	▶ ○○○법, 지방자치단체 ○○○○조례 등
키워드	▶ 인구, ○○시, 연도별
공공데이터 설명	▶ 데이터 관련 내용 설명
제공대상 여부	▶ 제공, 부분제공, 미제공, 미제공 사유, 미제공 근거 법률 기재
향후 제공	▶ 향후 제공, 향후 제공 연도, 향후 제공 사유에 대하여 기재
제공신청에 의한 등록	▶ 예 / 아니오로 선택

참조 공공데이터법 시행규칙 – 제공대상 공공데이터 등록서 항목

항목	속성	내용
공공데이터 명칭	dct:title	▶ ○○○정보
공공데이터 설명	dct:description	▶ 데이터의 생성 목적, 생성 과정, 활용 용도 등 데이터의 내용과 가치를 인식할 수 있는 내용을 서술
제3자 권리 포함 여부 (라이선스)	dct:license	▶ 제3자 권리(저작권, 소유권 등) 포함 여부
이용허락범위(저작권)	dct:rights	▶ 저작자표시(변경금지, 변경허락 등)
업데이트 주기	dct:accrualPeriodicity	▶ 반기/분기/연/수시
수정일시	dct:modified	▶ ○○○○년 ○○월 ○○일
위치(URL)	dcat:landingPage	▶ https://data.seoul.go.kr/data.....
매체유형	dcat:mediaType	▶ 텍스트
언어	dct:language	▶ 한국어

✔ 메타데이터 체계 설계 시 국가 데이터 카탈로그와의 연계성을 고려하여 항목 구성, 속성값 형식, 필수 여부 등을 현행 국가 표준과 일치시켜 정합성을 확보

[그림 4] 인공지능 친화적 공공데이터 메타데이터 체계



[표 7] 국가 데이터 카탈로그 표준 가이드 - 네임스페이스 참조코드

No	Prefix	URI	설명
1	adms	http://www.w3.org/ns/adms#	▶ 관리 메타데이터 표현 (예: 식별자, 버전 등)
2	dcat	http://www.w3.org/ns/dcat#	▶ DCAT(Data Catalog Vocabulary)의 주요 클래스 및 속성
3	dct	http://purl.org/dc/terms/	▶ DCT(Dublin Core Terms) 메타데이터 요소 정의
4	foaf	http://xmlns.com/foaf/0.1/	▶ 사람과 조직(기관) 정보 표현 (예: foaf:Agent, foaf:name)
5	ndi	-	▶ 국가 데이터 통합플랫폼에서 정의한 확장 속성
6	odrl	http://www.w3.org/ns/odrl/2/	▶ 데이터 사용 정책 및 제약 조건 표현
7	owl	http://www.w3.org/2002/07/owl#	▶ 클래스와 속성 간의 관계를 정의 (예: 동등함, 상위/하위 간계, 제약조건 등)
8	prov	http://www.w3.org/ns/prov#	▶ 데이터의 생성 및 수정 이력 등 provenance 정보
9	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	▶ RDF 기본 구조를 위한 요소 (rdf:about, rdf:type 등)
10	spdx	http://spdx.org/rdf/terms#	▶ 소프트웨어 패키지 및 라이선스 정보 (예: 라이선스 ID, 저작권 등)
11	vcard	http://www.w3.org/2006/vcard/ns#	▶ 연락처 정보 (예: 이메일, 주소)
12	xsd	http://www.w3.org/2001/XMLS chema#	▶ 데이터 타입 명시 (예: xsd:date, xsd:string)

✔ 인공지능 친화적 공공데이터 제공의 실행력과 유연성 확보를 위해 메타데이터 구성 요소를 '필수', '권장', '선택' 3단계로 등급화

[표 8] 메타데이터 구성요소 우선순위

우선순위	속성
필수	▶ 인공지능 친화적 공공데이터 관리를 위해 반드시 작성되어야 하는 최소 요건의 핵심 메타데이터 항목
권장	▶ 데이터 활용성, 품질, 이용자 접근성 등을 높이기 위해 작성을 권장하며, 반영 시 인공지능 친화적 성숙도가 향상되는 확장 메타데이터 항목
선택	▶ 데이터별 특화 정보나 이용자 편의 향상을 위해 제공되는 항목으로, 생략 시에도 기본적 정보 제공 및 활용에는 제약이 없는 메타데이터 항목

2.2.1.1 데이터 관리 메타데이터

✔ 데이터의 식별, 검색, 접근성 확보 및 시스템 간 유기적 연계·활용을 위해 관리해야 할 핵심 속성들로 구성되었으며, 우선순위에 따라 필수, 권장, 선택으로 구분

[표 9] 데이터 관리 메타데이터

	항목명	속성	우선순위	설명	예시
기본 제공	데이터명	dct:title	필수	▶ 데이터셋의 공식 명칭	서울시 대기질 측정 데이터_2025
	데이터 설명	dct:description	필수	▶ 데이터셋의 내용, 보유 목적, 수집 대상 등 설명	서울시 내 구 대상 대기질 수집, 대기 환경 분석을 위함
	관련법령	dct:references	필수	▶ 법령 참조, 법적 근거	○○○법 제 ○○조 (URL로 연결)
	소관기관 (운영부서명)	dct:creator	필수	▶ 데이터셋을 생산·관리하는 주체	데이터 전략과
	제공기관	dct:publisher	필수	▶ 데이터셋을 대외적으로 배포·게재하는 기관	서울특별시
	제공시스템	dcat:landing Page	필수	▶ 배포 또는 추가 정보에 대한 액세스를 제공하는 웹 페이지	https://data.seoul.go.kr/data~
	갱신주기	dct:accrual Periodicity	필수	▶ 데이터셋 업데이트 빈도	http://purl.org/cld/freq/annual
	업무/ 주제분류체계	dcat:theme	필수	▶ 데이터셋의 카테고리를 참조. 데이터셋은 여러 테마와 연결 가능	환경
	키워드	dcat:keyword	필수	▶ 검색 및 분류를 위한 핵심 키워드	환경, 대기질, 미세먼지
추가 확장	고유식별자	dct:identifier	필수	▶ 중복되지 않는 영구 식별자	10.1234/seoul.air.pm.2024.v1
	담당자연락처	dcat:contact Point	필수	▶ 데이터셋 담당자에 대한 연락처 정보 ▶ 이름, 전화번호, 이메일을 하위로 제공	홍길동
	언어	dct:language	필수	▶ 데이터셋의 언어를 참조하며 여러 언어가 있는 경우 이 속성을 반복 가능	http://id.loc.gov/vocabulary/iso639-1/ko
	배포본	dcat:distribution	필수	▶ 데이터셋에 배포본이 존재함을 의미	Distribution 1
	접속 URL	dcat:accessURL	필수	▶ 데이터셋의 배포본 URL	https://data.seoul.go.kr/data~
	데이터용량	dcat:byteSize	필수	▶ 데이터의 크기(바이트)	2000
	데이터셋 유형	dcat:mediaType	필수	▶ 수치, 이미지, 음성, 영상 등 구분	image/jpeg
	시간범위	dct:temporal	권장	▶ 데이터셋에서 다루는 시간적 기간	2025.01.01.~ 2025.12.31
	공간범위	dct:spatial	권장	▶ 데이터셋에서 다루는 지리적 영역	대한민국 서울특별시 (25개 자치구)
	연계데이터셋	dct:relation	선택	▶ 연계 데이터셋이나 참조 리소스	서울시 대기질 측정 데이터_2024

예시 데이터 관리 메타데이터 작성 예시 – 2025 서울시 연령별 인구통계 정보

항목	예시
데이터명	“dct:title”: “2025 서울시 연령별 인구통계 정보”
제공범위	“dct:description”: “서울시 거주 인구를 연령대별·구별로 분류하여 인구구조 분석, 정책 수립 및 행정 서비스 개선을 위해 제공되는 통계 데이터.”
관련법령	“dct:references”: “통계법 제3조 및 국가통계작성지침 준수”
소관기관	“dct:creator”: “서울특별시 빅데이터담당관”
제공기관	“dct:publisher”: “서울특별시”
표준 데이터셋 제공시스템	“dcat:landingPage”: “https://data.seoul.go.kr/dataset/2025-pop-age”
접속 URL	“dcat:accessURL”: “https://data.seoul.go.kr/datafile/POP_2025_AGE.csv”
갱신주기	“dct:accrualPeriodicity”: “http://purl.org/cld/freq/annual”
고유식별자	“dct:identifier”: “seoul-pop-age-2025”
담당자 연락처	“dcat:contactPoint”: “fn”: “서울특별시 빅데이터담당관 인구통계 담당자”, “hasEmail”: “mailto:data@seoul.go.kr”
업무/주제분류 체계	“dcat:theme”: “인구·사회”
키워드	“dcat:keyword”: “인구”, “연령별”, “서울시”, “2025”, “인구통계”, “demographics”
언어	“dct:language”: “http://id.loc.gov/vocabulary/iso639-1/ko”
데이터 용량	“dcat:byteSize”: “2400000”
데이터셋 유형	“dcat:mediaType”: “text/csv”
관련 데이터셋	“dct:relation”: “2024 서울시 연령별 인구통계 정보”, “서울시 구별 인구 동향 데이터”
시간범위	“dct:temporal” “time:hasBeginning”: “2025-01-01”, “time:hasEnd”: “2025-12-31”
공간범위	“dct:spatial”: “서울특별시”

2.2.1.2 데이터 계보 메타데이터

✔ 데이터셋 배포 이후의 변경 이력을 투명하게 관리하고 최신성을 보장하기 위해 수정·업데이트 관련 속성을 정의하며, 데이터 신뢰도 및 재현성 제고를 위해 필수항목으로 구성

[표 10] 데이터 계보 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	버전	owl:versionInfo	필수	▶ 현재 버전 정보	2025.01.v1
	등록일시	dct:issued	필수	▶ 시스템 최초 등록 날짜 (초기 등록 후 변경 금지)	YYYY-MM-DD
	수정일시	dct:modified	필수	▶ 마지막 수정 날짜	YYYY-MM-DD
	버전 노트	adms: versionNotes	필수	▶ 이전 버전 대비 변경 사항	▶ 데이터셋 최초 공개 ▶ 결측치 처리 방식 변경 ▶ 측정지역 코드화 및 적용

예시 데이터 계보 메타데이터 작성 예시 – 딸기 생육과정 모니터링

항목	예시
버전	“owl:versionInfo”: “v1.2”
등록일시	“dct:issued”: “2025-01-15T09:00:00+09:00”
수정일시	“dct:modified”: “2025-03-02T14:30:00+09:00”
버전 노트	“adms:versionNotes”: “센서 오류값 제거 및 생육지표(엽장·착과수) 추가. 영상데이터 결측치 보정 알고리즘 개선.”

2.2.1.3 데이터 이용 메타데이터

✔ 데이터 이용 시 사용 범위와 제약 사항을 정의하여 법적 불확실성을 해소하고, 데이터의 투명한 이용 및 권리 보호를 위한 권한 속성으로 구성

[표 11] 데이터 이용 메타데이터

	항목명	속성	구분	설명	예시
기본 제공	라이선스	dct:license	필수	▶ 데이터셋의 권리 관련 공식적인 법적 문서를 기술하는 속성 - 특정 라이선스 문서(예: CC BY, GPL 등)의 URI를 직접 참조하는 용도	https://creativecommons.org/licenses/by-nc-sa/4.0/
	저작권	dct:rights	필수	▶ 데이터셋 관련 저작권 정보 전반에 관한 정보를 기술하는 속성 - 저작자 표시, 비영리 여부, 변경금지 여부, 변경허락 여부	상업적·비상업적 이용 모두 허용

예시 데이터 이용 메타데이터 작성 예시 – 대구시 업종별 사업자 분포

항목	예시
라이선스	“dct:license”: “https://...../terms/daegu-dataset-usage-policy” (이용조건 URL)
권한	“dct:rights”: “본 데이터셋의 저작권은 대구광역시에 있으며, 상업적 이용 및 제3자 제공은 내부 승인 절차에 따른다.”

2.2.1.4 데이터 품질 메타데이터

- 데이터 품질 수준 및 한계점을 사전에 인지하여 AI 모델의 신뢰성을 확보할 수 있는 속성으로 구성

[표 12] 데이터 품질 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	데이터 편향성	rai:dataBiases	권장	▶확인된 편향 유형	“도시 지역(서울·수도권)에 측정소가 집중되어 있어 농촌 지역의 대표성이 낮음”
	품질 검증 정보	dqv:has Quality Annotation	선택	▶수행된 품질 검토 및 평가 내용	“월 1회 수동 검수 완료 (환경전문가 검토)”
	데이터 한계	rai:known Limitations	선택	▶데이터셋의 구조적·내용적 제한사항 및 해석 시 유의 사항	“일부 측정소는 장비 유지보수 기간동안 측정값 누락 (측정값 누락 지역 ○○○, ○○○, ○○○)”
	결측치 정보	rai:data Collection MissingData	선택	▶결측치 비율 및 처리방식	“결측 사유: 센서 점검(2.1%), 통신 오류(1.1%)”

예시 데이터 품질 메타데이터 작성 예시 – 국내 귀화자 출신 국가 통계

항목	예시
알려진 편향 명세	“rai:dataBiases”: “일부 출신국가는 귀화 허가 대상자 규모가 매우 작아 통계적 변동성이 큼.”, “여성 비율이 높은 특정 국가 편중 현상이 존재하여 전체 귀화자 성비를 과대/과소 추정할 가능성 있음.”, “경제활동 연령대(20-40대)가 주로 귀화하는 경향으로 인해 고령층 데이터가 부족함.”
데이터셋 품질 검증 여부	“dqv:hasQualityAnnotation”: “@type”: “dqv:QualityAnnotation”, “dqv:hasQualityMeasurement”: “법무부 연례 검증 절차를 통해 원시 귀화 허가 데이터의 정확성·일관성 점검 완료(2025년 검증 기준).”
데이터 한계	“rai:knownLimitations”: “비공식 체류자의 국적 변경 이력은 포함되지 않아 귀화자 전체 규모를 완전히 대변하지 못함.”, “귀화 심사 탈락자·보류자의 통계는 포함되어 있지 않아 국가별 전환율 분석이 제한됨.”, “표본 크기가 작은 국가의 경우 단일 관측치 변화가 통계에 큰 영향을 미침.”
결측치 정보	“rai:dataCollectionMissingData” : “센서 점검(2.1%), 통신 오류(1.1%)로 결측이 발생할 수 있음.”

2.2.1.5 데이터 타입별 메타데이터

2.2.1.5.1 수치 메타데이터

✔ 데이터의 단위, 식별 값 등 숫자형 변수의 의미와 특성에 관련된 속성으로 구성

[표 13] 수치 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	레코드셋 ID	cr:RecordSet/@id	권장	▶ 해당 컬럼들이 속한 표(테이블/레코드 집합) 식별자	air_quality
	컬럼 논리 식별자	cr:Field/@id	권장	▶ 컬럼의 논리 식별자(권장: RecordSetID/컬럼명)	air_quality/station_id
	컬럼설명	cr:Field/description	권장	▶ 컬럼의 상세 설명	대기질 측정소가 위치한 지역의 고유 식별 번호
	데이터 타입	cr:Field/dataType	권장	▶ 컬럼 데이터 타입	sc:Integer

2.2.1.5.2 이미지 메타데이터

✔ 이미지 데이터의 형식, 생성 조건 등 정보를 기술한 속성으로 구성

[표 14] 이미지 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	가로 크기	schema:width	권장	▶ 이미지 가로 픽셀 수	1920
	세로 크기	schema:height	권장	▶ 이미지 세로 픽셀 수	1080
	스캔 장비 제조	exif:make	선택	▶ 이미지 입력 장비 제조사	Fujitsu
	스캔 장비 모델	exif:model	선택	▶ 이미지 입력 장비 모델명	fi-7160

2.2.1.5.3 음성 메타데이터

✔ 음성 데이터의 재생시간, 샘플레이트, 채널 수 등 음향 특성과 관련된 속성으로 구성

[표 15] 음성 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	재생시간	schema:duration	권장	▶ 오디오 파일 전체 재생 시간	PT3M25S (ISO 8601 duration)
	샘플레이트	schema:additionalProperty (PropertyValue, name=sampleRate)	권장	▶ 초당 샘플링 해 얻은 디지털 신호(Hz)	16000
	채널 수	schema:additionalProperty (PropertyValue, name=channels)	선택	▶ 채널수	1(mono)

2.2.1.5.4 영상 메타데이터

✔ 영상 데이터의 형식, 생성 조건, 영상 특성과 관련된 속성으로 구성

[표 16] 영상 메타데이터

	항목명	속성	구분	설명	예시
추가 확장	가로 크기	schema:width	권장	▶ 이미지 가로 픽셀 수	1920
	세로 크기	schema:height	권장	▶ 이미지 세로 픽셀 수	1080
	재생시간	schema: duration	권장	▶ 비디오 전체 재생 시간	PT3M25S (ISO 8601 duration)
	초당 프레임 수	schema: frameRate	선택	▶ 초당 재생 프레임 수	30
	코덱명	schema:additionalProperty (PropertyValue, name=codec)	선택	▶ 비디오 코덱 정보	MPEG-4 Part 2

참고 **AI 활용을 고려하여 고도화된 DCAT-US**

▶ DCAT-US는 W3C가 제정한 국제 데이터 카탈로그 표준(DCAT)을 기반으로, 미국 연방정부의 오픈데이터 정책과 법적 요구사항을 반영하여 관리예산처(OMB)가 도입한 메타데이터 사양

▶ 최신 버전인 DCAT-US v3.0은 생성형 AI와 데이터 자동 활용 확산에 대응해 메타데이터의 구조와 의미 표현력을 강화하였으며 글로벌 DCAT v3.0 표준과 구조적으로 호환되도록 설계

필드	설명	비고
액세스 제한	▶ 데이터에 대한 액세스 제한이 있는지 여부 표시 ▶ 민감한 과거 기록에 대한 책임 있는 액세스를 보장 ▶ 투명성을 높여 연구자와 권한 있는 이용자가 아카이브 된 자료에 대한 액세스 매개변수를 이해하고 탐색하는 데 도움을 줌	DCAT-US 공식 권장
데이터 사전	▶ 데이터 집합의 필드(변수, 차원, 측정값, 속성)를 정의하는 데이터 사전 또는 스키마를 지정 ▶ 데이터 집합을 올바르게 구문 분석하고 학습된 패턴 및 모델 출력의 품질을 개선하기 위한 컨텍스트를 제공하는 데 모두 중요	DCAT-US 공식 권장
식별자	▶ 데이터 집합의 고유 식별자 ▶ 서로 다른 데이터셋을 명확히 구분하고, 투명성과 데이터 일관성을 촉진	DCAT-US 공식 권장
키워드	▶ 데이터 집합을 설명하는 키워드 ▶ 데이터 집합의 큰 초점을 요약하는 데 유용하며, 개발자 및 기타 이용자가 새로운 관련 데이터를 찾거나 검색/크롤링하는데 도움이 될 수 있음	DCAT-US 공식 권장
라이선스	▶ 데이터 집합을 사용할 수 있는 라이선스를 나타내어 이용자와 자동화된 시스템이 데이터 라이선스를 정확하게 구문 분석 하는데 도움	DCAT-US 공식 권장
게시자	▶ 데이터 집합을 사용할 수 있게 만드는 책임이 있는 주체 ▶ 데이터 원본의 이해 및 인용, 데이터 집합에 대한 다른 컨텍스트 연결에 중요	DCAT-US 공식 권장
권한	▶ 데이터 집합과 관련된 권한을 지정하여 이용자와 자동화된 시스템이 데이터 권한을 정확하게 구문 분석하는데 도움	DCAT-US 공식 권장
시간 범위	▶ 데이터 집합이 포함하는 기간 관련 데이터로, 모델을 학습시키는 데 중요하며 모델이 시기적절하고 관련성 있는 결과를 반환할 가능성을 높임	DCAT-US 공식 권장

필드	설명	비고
업데이트/수정 날짜	▶ 데이터 집합이 변경되거나 수정된 가장 최근 날짜 ▶ 모델 개발자가 관련 정보에 액세스/업데이트하는 데 도움	DCAT-US 공식 권장
문서	▶ 데이터 집합에 대한 자세한 정보가 포함된 페이지 또는 문서 ▶ 모델 개발자가 데이터 집합에 대한 추가적인 관련 컨텍스트/메타데이터를 쉽게 식별할 수 있도록 도와줌	권장 속성
버전	▶ 현재 데이터 집합의 버전, 에디션 또는 각색인 관련 데이터 집합 ▶ 원시 데이터 집합과 파생된 데이터 집합을 식별하고 서로 다른 데이터 버전을 명확하게 구분하는 데 도움	권장 속성

2.2.2 문서화

공공데이터의 신뢰성과 활용성 확보를 위해 데이터 구조, 목적, 수집 방식, 편향성, 한계점 및 버전 이력 등을 투명하게 기술한 문서

- **목적:** AI 모델 개발자의 데이터 맥락에 대한 완전한 이해 및 수집부터 가공까지 전 과정의 투명성 확보를 통한 데이터 오용 방지 및 분석 재현성 확립
- **범위 및 내용:** 단순 데이터 파일 제공을 넘어 데이터의 활용 목적, 구조·기술적 규격, 활용 과정에서 발생할 수 있는 제약 사항(편향성 등) 등을 상세히 문서화
- **관리 원칙:** 데이터와 문서 간 정합성을 유지하기 위해 데이터 변경 시 문서 정보를 즉각적으로 현행화하고 변경 로그를 작성하여 데이터 수정·갱신 이력을 체계적으로 관리
- 인공지능 친화적 공공데이터는 다음의 문서들을 체계적으로 포함하여 제공해야 함

참고 문서화 대상 목록

구분	문서명	핵심내용
기본 명세	데이터 카드 (Data Card)	▶ 데이터 집합 개요, 활용 목적, 제약사항, 퀵스타트 등
기술 규격	스키마 정의서 (데이터 사전)	▶ 변수명(물리/논리), 데이터 타입, 단위, 필드 간 관계 등 명세
	메타데이터 명세서	▶ 관리, 업데이트, 이용, 투명성 등 메타데이터 정보
	API / Access 문서	▶ 데이터 접근 경로, 호출 규약, 인증 및 보안 설정
신뢰 보증	품질 진단 보고서	▶ 정확성, 완전성, 유효성 등 품질 지표 및 검증 결과 보고서
	개인정보/법적 리스크 분석	▶ 비식별 조치 현황, 저작권 권리관계, 라이선스 활용 조건
이력 관리	계보 및 출처	▶ 원천 데이터 소스, 수집 환경 정보
	버전 관리 및 변경 이력	▶ 유의적 기준 버전 정보, 변경 로그, 마이그레이션 노트

데이터 카드: 국내외 다수 플랫폼에서 널리 활용되는 문서로, 데이터의 적합성 판단 및 활용 여부 확인을 위한 표준화된 형태의 데이터 핵심 정보를 제공

- 데이터 카드는 기존 공공데이터 설명서나 메타데이터를 대체하는 문서가 아니라, AI 활용 관점에서 핵심 정보를 요약 및 보완하는 참고 문서
- 데이터의 주요 특성, 규격, 제약 사항 등을 한눈에 파악할 수 있도록 직관적으로 구성
- 데이터 활용 가능성을 우선 검토하게 함으로써 이용자의 데이터 탐색 비용 절감 및 접근 편의성 제고

참고 데이터 카드 상세 작성 항목[6] ※ 상세내용은 부록 4 참고

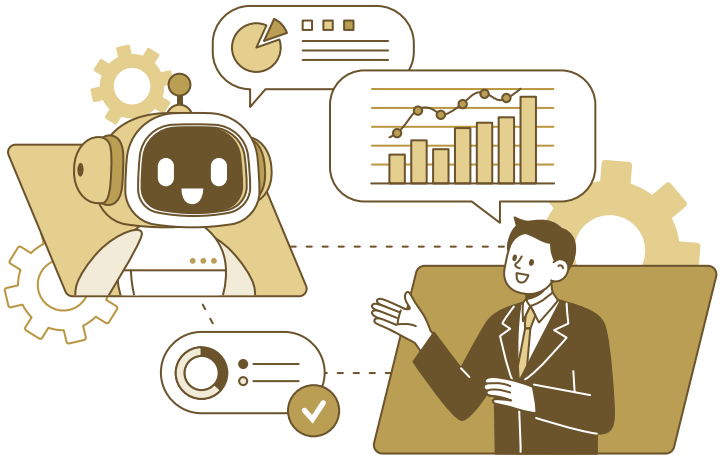
내용	
데이터 집합 개요	▸ 데이터셋 요약, 지원 태스크. 언어 정보
데이터셋 구조	▸ 데이터 인스턴스(샘플), 데이터 필드(명칭/타입/의미), 분할 정보
데이터셋 생성	▸ 구축 배경, 데이터 소스(출처), 어노테이션 과정
사용 시 고려사항	▸ 사회적 영향, 편향성 및 한계 (오용 방지 및 기술적 제약)
추가 정보 및 라이선스	▸ 라이선스, 인용, 작성자 정보
기술적 사양	▸ 재현성, 환경 설정, 매개변수 (전처리/평가 코드 및 라이브러리 버전)

예시 데이터 카드 : 도시 공원 유동인구 분석 데이터

<table>
<tr><th>항목</th><th>예시</th></tr>
<tr><td>데이터셋 개요</td><td>▸ 데이터셋 요약: 서울시 주요 공원 10개소의 실시간 인파 밀집도, 체류 시간 및 이동 방향 정보 제공
▸ 활용 예시: 시계열 유동인구 예측, 구역별 혼잡도 분류. 방문 패턴 클러스터링
▸ 언어 정보: 한국어(기본 명칭), 영어(데이터 필드 및 메타데이터)</td></tr>
<tr><td>데이터셋 구조</td><td>▸ 데이터 인스턴스: JSON 형식의 단일 레코드 샘플 제공
- { “timestamp”: “2023-10-01 14:00:00”, “park_id”: “PK_SEOUL_001”, “visitor_count”: 142, “avg_stay_min”: 15.5, “weather”: “Sunny”}
<table>
<tr><th>필드명</th><th>타입</th><th>설명</th></tr>
<tr><td>timestamp</td><td>Datetime</td><td>수집 일시 (YYYY-MM-DD HH:MM:SS)</td></tr>
<tr><td>park_id</td><td>String</td><td>공원 고유 식별 코드</td></tr>
<tr><td>visitor_count</td><td>Int</td><td>시간대별 진입 인원수 (단위: 명)</td></tr>
<tr><td>avg_stay_min</td><td>Float</td><td>구역 내 평균 체류 시간 (단위: 분)</td></tr>
</table>
▸ 분할 정보:학습용(Train) 70%, 검증용(Validation) 15%, 평가용(Test) 15% (날짜 기준 순차 분할 수행)</td></tr>
</table>

6 Hugging Face. (n.d.). Dataset Cards Guide for Hugging Face Datasets Repository. https://huggingface.co/docs

항목	예시
데이터셋 생성	▸ **구축 배경:** 도시 공원 관리 효율화 및 인파 밀집 사고 예방을 위한 조기 경보 시스템 학습 데이터 확보 ▸ **데이터 소스:** 서울시 IoT 도시데이터 센서(S-DoT) API 및 지능형 CCTV 카운팅 로그 ▸ **어노테이션:** 원천 로그 기반의 이상치 제거 및 비식별화 가공 (De-identification) 수행
사용 시 고려사항	▸ **사회적 영향:** 도시 인프라 최적화 및 시민 안전 서비스 고도화에 기여 ▸ **편향성 및 한계:** 스마트 기기 미보유자(유아, 고령층 등) 수치의 과소측정 가능성 존재 ▸ 폭우 등 기상 악화 시 센서 오차(약 5%) 발생 가능성이 있으므로 모델링 시 보정 권장
추가 정보 및 라이선스	▸ **라이선스:** 공공누리 제1유형 (출처 표기 시 상업적 이용 및 변형 가능) ▸ **인용(Citation):** Seoul Digital Foundation (2023). City Park Human Flow Pattern Dataset v1.2.0. ▸ **작성자 정보:** 데이터기획팀 (contact@sdf.seoul.kr)
기술적 사양	▸ **재현성:** 저장소 내 preprocess.py를 통해 동일한 데이터 정제 결과 도출 가능 ▸ **환경 설정:** Python 3.9+, Pandas 1.5.2, Scikit-learn 1.1.3 ▸ **주요 매개변수:** - outlier_z_score: 3.0 (이상치 제거 임계값) - smoothing_window: 15min (데이터 평활화 시간 단위) - random_seed: 42 (데이터 분할 시 무작위성 고정)



2.3 품질 지표

AI 학습·추론 과정에서 요구되는 정확성·신뢰성·의미적 완결성 등 데이터 품질 확보를 목적으로 기존 공공데이터 품질 기준을 확장·고도화한 지표

2.3.1 적용 범위

- 본 품질지표는 데이터의 형식이나 구조와 무관하게 AI 학습·추론·의사결정에 활용되는 모든 공공데이터 자산을 대상으로 적용되며, 구체적으로 다음을 포함한다.
  - 원천 데이터베이스(DB)에서 추출한 데이터를 인공지능 활용에 적합하도록 정제·가공·변환하여 구축한 데이터 집합
  - 인공지능 학습·튜닝 목적에 따라 수집·정제되어 제공되는 텍스트·이미지·영상 등 비정형 데이터셋
  - 인공지능 추론의 근거 제공을 위해 구축·관리되는 법령·지침·통계 등 지식 기반 데이터셋

2.3.2 인공지능 친화적 공공데이터 품질 지표

- **구성 원칙:** 기존 공공데이터 품질관리에서 요구하는 품질 요소를 기반으로, 완전성·일관성·정확성·적시성·유효성·유일성 등 데이터 품질 요소를 AI 학습·활용 관점에서 심화 및 확장한 지표
  - 인공지능 기술이 단순 자동화를 넘어 의사결정 지원, 대국민 서비스 제공, 국가안보 및 재난관리 등 공공행정 전반의 핵심 인프라로 활용됨에 따라 AI 모델의 성능·신뢰성·윤리적 정합성을 결정짓는 데이터 품질의 중요성이 더욱 강화되는 추세를 반영

[표 17] 인공지능 친화적 공공데이터 품질 지표

기존 공공데이터		▸ **구조적 정합성 중심:** 정형 데이터베이스(DB) 기반의 구문적 정확성, 형식적 무결성 확보에 중점을 두어 데이터 유형, 길이, 제약 조건 등 스키마 준수 여부를 진단, 데이터의 관리 효율성과 시스템 안정성을 극대화하는 것이 목표
인공지능 친화적 공공데이터		▸ **의미적 완결성 중심:** 인공지능 모델의 학습, 추론에 필수적인 정확성, 신뢰성, 다양성 등을 확보하고, 데이터 간의 맥락적 연결성과 의미적 정합성 고도화로 별도의 정제·공정 없이 AI 모델에 즉시 투입 가능한 실효적 품질 구현이 목표
품질 지표		지표 정의
완전성	기존	▸ 필수 항목의 누락여부 및 대상·기간·지역·집단 등 필요한 범위의 정보 포함 수준
	확장	▸ AI 학습·추론에 필요한 정보가 누락되지 않도록 데이터가 필요한 범위(대상·기간·지역·상황 등)를 충분히 포괄하고, 주요 활용 맥락을 반영할 수 있는 수준의 정보를 확보했는지 여부
일관성	기존	▸ 동일 개체·코드·속성값이 데이터 전반에서 논리적으로 모순 없이 적용되고 있는지 여부
	확장	▸ 데이터 요소 간 의미적 연결성이 유지되고, 동일한 개념이 다양한 맥락에서 일관되게 동일 의미로 활용되어 AI의 해석·추론 과정에서 혼동이 발생하지 않도록 의미적 일관성을 확보했는지 여부

품질 지표		지표 정의
정확성	기존	▶ 데이터 값이 실제 값 또는 합의된 기준값과 일치하는지 여부
	확장	▶ 데이터 내용이 사실과 부합하고, 신뢰할 수 있는 출처·근거에 기반하여 기록·관리됨으로써 AI가 사실과 부합하지 않는 정보를 학습하지 않도록 사실 기반의 정확성을 확보했는지 여부
적시성	기존	▶ 데이터가 정의된 주기 또는 기준 시점에 따라 최신 상태로 유지되는지 여부
	확장	▶ 데이터가 AI 학습·활용 시점에서도 활용 가능한 최신성 및 정책·제도·환경 변화 등 최신 맥락을 반영한 시의성을 확보했는지 여부
유효성	기존	▶ 정의된 스키마, 데이터 유형, 허용 범위, 제약 조건 등 사전 규칙 준수 여부
	확장	▶ 데이터 및 메타데이터(JSON, 라벨링 결과 등)가 AI 처리 파이프라인에서 오류 발생을 최소화할 수 있도록 구조·형식·스키마·제약 조건을 준수했는지 여부
유일성	기존	▶ 동일 키 또는 동일 객체에 대한 중복 레코드 존재 여부
	확장	▶ 데이터의 중복으로 인해 AI 학습 과정에서 편향, 과적합, 데이터 누수 등의 위험이 증가하지 않도록 불필요한 중복 데이터를 제거했는지 여부

2.3.2.1 완전성

정의

AI 모델이 학습과 추론을 수행하는 데 필요한 데이터·지식·시나리오의 누락 여부를 확인하고, 필요한 범위(대상·기간·지역·상황 등)

주요 속성

❶ **충분성**: 필수 데이터의 누락 방지 및 시나리오 커버리지 확보
❷ **대표성·균형성**: 데이터 분포의 균형 확보 및 편향 완화

[표 18] 완전성 ❶ 충분성 품질 지표

충분성 품질 지표		
필수 속성 충족도	정의	▶ AI 모델의 학습·추론에 필요한 핵심 속성 및 라벨값의 누락 여부를 측정
	평가 목적	▶ 데이터 결측으로 인한 학습 오류 및 성능 저하를 예방하여 학습용 데이터로서의 활용 가능성을 확보
	진단기준	▶ AI 학습에 필수적인 컬럼(라벨, 핵심 변수 등)의 값에 결측(Null) 또는 공백 존재 여부를 파악
	진단방법	▶ 필수 항목을 대상으로 전체 레코드 건수 대비 값이 Null 또는 공백이 아닌 레코드 건수의 비율을 측정
	오류예시	▶ **시계열 단절**: 교통량 예측 데이터에서 핵심인 측정 일시나 교통량 수치가 비어 있는 경우 ▶ **학습 곤란**: 지역별 민원 분석 데이터에서 행정동 코드가 누락되어 지역 패턴 학습이 곤란한 경우 ▶ **라벨 부재**: 학습용 이미지 데이터에 파일은 존재하나, 정답지인 라벨이 매칭되지 않은 경우

충분성 품질 지표		
시나리오 변수 충분성	정의	▸ 정책 목표 달성 및 AI 서비스 구현에 필요한 데이터의 핵심 속성 및 변수를 스키마에 포함하고 있는지 평가
	평가 목적	▸ 잠재적 AI 서비스 모델을 기준으로 데이터의 활용 가치를 사전 검증
	진단기준	▸ AI 개발을 위한 사전 정의된 분석 시나리오에서 요구하는 입력/출력 변수가 데이터 스키마에 확보되었는지 여부
	진단방법	▸ **시나리오 기반 매핑**: 참조 AI 서비스를 선정·기획하고, 필수 데이터 요건을 정의하여 미확보 항목을 식별 ※ 연계 데이터셋 분석: 외부 데이터와 결합하여 분석하는 경우, 목적 적합성을 위해 [외부 데이터와 원활하게 연결될 수 있는가?(Linkability)] 여부를 판단하며, 이 경우 공통 결합 키 존재 여부로 진단
	오류예시	▸ **교통단속 AI**: 불법주정차 단속을 위해 CCTV 영상과 시간은 확보했으나, 과태료 부과 및 위반유형(소화전, 횡단보도 등) 코드가 설계에서 누락 된 경우 ▸ **AI 기반 상권 분석 서비스**: 매출 예측에 필요한 업종 코드나 시간대별 유동 인구 정보가 스키마 설계 단계에서 누락된 경우

[표 19] 완전성 ❷ 대표성·균형성 품질 지표

대표성·균형성 품질 지표		
데이터 다양성	정의	▸ 데이터가 특정 조건이나 환경에 편중되지 않고 다양한 속성(지역, 날씨, 인구 등)과 예외 상황을 포괄하며, 학습(Train), 검증(Validation), 평가(Test) 데이터셋 간의 통계적 분포가 유사한지 평가하는 지표
	평가 목적	▸ AI 모델의 학습하지 않은 새로운 환경이나 변수에 대한 대응 역량(일반화 성능)을 강화하기 위함
	진단기준	▸ **속성 분포**: 주요 환경 변수별 데이터가 특정 조건에 과도하게 쏠리지 않고 균형 있게 분포하는지 진단 ▸ **분할 일치성**: 학습(Train), 검증(Validation), 평가(Test) 데이터셋 간 주요 속성의 통계적 분포의 유사성 진단
	진단방법	▸ **시각화 분석**: 속성별 빈도수 히스토그램을 통해 편향 여부 시각적 점검 ▸ **통계적 검증**: 학습셋과 검증셋 간 분포 차이를 검증하기 위해 카이제곱 검증(범주형) 또는 KS 테스트(연속형) 수행 ※ 시각화 분석, 통계 검증은 1차 정량 진단으로 희귀 케이스, 시나리오 의도된 데이터 등의 다양성 진단을 위해서는 추가 방법이 필요 (체크리스트, 추가 지표 확인 등)
	오류예시	▸ **지역 편향**: 전국 대상 서비스임에도 불구하고 수도권 데이터가 90% 이상을 차지하여 지방 환경에 대한 인식 성능 저하 우려 ▸ **분할 오류**: 학습셋(Train)은 모두 맑은 날 데이터로 구성하고, 검증셋(Validation)에만 비 오는 날을 배치하여 모델 성능 평가가 왜곡되는 경우
라벨 다양성	정의	▸ AI 모델이 예측해야 할 정답 클래스(Label/Class)가 정의된 범위 내에 포함되어 있는지, 또한 클래스 분포의 불균형 수준이 과도하지 않은지 평가하는 지표
	평가 목적	▸ 학습 데이터가 다수 범주에 편중되어 소수 범주에 대한 성능 저하가 발생할 수 있는 클래스 편향 위험 완화
	진단기준	▸ **최소 수량**: 정의된 모든 클래스가 사전에 설정한 최소 유효 수량 이상 존재해야 함 ▸ **균형 비율**: 최다 빈도 클래스와 최소 빈도 클래스 간의 비율이 사전에 설정한 임계치 이내여야 함

대표성·균형성 품질 지표		
라벨 다양성	진단방법	▸ **불균형 비율 산출:** (최다 클래스 데이터 수 / 최소 클래스 데이터 수) 값을 측정하여 기준치 초과 여부를 확인 ※ 최소 클래스 데이터 수가 0인 경우 별도 판정 ▸ **지니 계수 활용:** 클래스 분포의 불평등 정도를 0~1 사이 값으로 정량화
	오류예시	▸ **객체 인식:** 도로 장애물 10종을 정의했으나, 데이터의 95%가 라바콘이고 주요 객체인 보행자는 0.1%에 불과한 경우(보행자 인식 성능 저하 우려) ▸ **텍스트 분류:** 민원 자동 분류 서비스를 위한 데이터셋에서 기타 유형이 전체의 80%를 차지하여 변별력 저하 우려가 있는 경우

2.3.2.2 일관성

정의

동일 개체·코드·속성값이 데이터셋 전반에 걸쳐 논리적 모순 없이 적용되고, 데이터 간 참조 관계가 사전에 정의된 규칙에 부합하는지 평가하는 지표

주요 속성

❶ **의미적 일관성:** 데이터 간 내용상의 충돌 방지 및 기준 통일

[표 20] 일관성 ❶ 의미적 일관성 품질 지표

의미적 일관성 품질 지표		
관계 일관성	정의	▸ 단일 레코드(행)내에 존재하는 속성 간 관계가 시간적 선후 관계 또는 사전 정의된 업무 규칙(Business Rule)에 위배되지 않고 논리적 정합성을 유지하는지 평가하는 지표
	평가 목적	▸ 데이터 내 논리적 관계 및 선후 관계 오류를 제거하여, AI 모델이 잘못된 논리 또는 현실과 부합하지 않는 규칙을 학습할 위험 최소화
	진단기준	▸ **시간 규칙:** 종료일은 시작일보다 빠를 수 없음 (종료일 ≥ 시작일) ▸ **상태 규칙:** 처리 상태 값에 따라 연관된 날짜나 사유 필드가 필수적으로 존재해야 함 ▸ **수치 규칙:** 총합은 구성 요소의 합계와 일치해야 함
	진단방법	▸ 속성 간 논리 관계를 정의한 검증 로직을 수행하여 오류 레코드 건수 및 비율 측정
	오류예시	▸ **시간 역전:** 사업 시작일(2025-01-01)보다 종료일(2024-12-31)이 이전으로 기록됨 ▸ **상태 모순:** 민원 처리 상태는 '반려'이나, 필수항목인 '반려 사유'가 공란이거나 '처리 완료일'이 기재되어 있음
기준 정보 일관성	정의	▸ 동일한 개체(사람·기관·사업체 등)를 지칭하는 데이터가 여러 파일이나 테이블에 산재할 경우, 해당 개체의 주요 속성(이름, 주소, 식별정보 등)표기가 불일치 없이 일관되게 관리되는지 평가하는 지표
	평가 목적	▸ 동일 대상(예: A기업)에 대한 표기 불일치를 해소하여, AI가 이를 서로 다른 대상으로 오인하여 학습 효과가 분산되거나 개체 식별 성능이 저하될 수 있는 위험을 완화
	진단기준	▸ 동일 식별자를 가진 개체의 속성값이 모든 데이터셋(테이블)에서 일치하는지 여부를 확인
	진단방법	▸ 서로 다른 테이블/파일 간 동일 식별자(KEY) 기준 조인(JOIN) 비교 또는 중복 그룹화를 통해 속성값의 분산 여부 측정

의미적 일관성 품질 지표		
기준 정보 일관성	오류예시	▸ **상호명 혼재**: 사업자번호는 동일하나, 테이블 A는 '㈜한국AI', 테이블 B는 '한국에이아이 주식회사'로 상이하게 표기 ▸ **속성 충돌**: 동일 고객 ID에 대해 회원정보 테이블은 '30세', 구매이력 테이블은 '40세'로 나이 정보가 상충
참조 무결성	정의	▸ 데이터셋 내의 분류 코드나 외래키(FK)값이 기준이 되는 마스터 코드 테이블에 실재하는 유효한 값인지를 평가하는 지표
	평가 목적	▸ 존재하지 않거나 폐기된 코드를 참조하여 발생하는 데이터 단절을 방지하고, AI가 잘못된 관계 정보를 학습하거나 특성 결합 시 데이터가 유실되는 위험 최소화
	진단기준	▸ 상세 데이터의 모든 코드/참조 값이 표준 코드 테이블 내에 존재하는지 확인
	진단방법	▸ 상세 데이터 테이블과 기준 코드 테이블을 연결하여 확인하거나 Not Exists 조건을 사용하여 매핑되지 않는 데이터 검출
	오류예시	▸ **미등록 코드**: 입력된 행정동 코드가 기준 행정표준코드 목록에 존재하지 않는 값인 경우 ▸ **폐기 코드 참조**: 2024년 데이터에서 이미 폐기(또는 변경)된 과거 조직 코드를 사용하는 경우

2.3.2.3 정확성

정의

AI 모델이 학습·추론에 사용하는 데이터 값과 라벨이 현실의 사실 및 의미와 부합하는지, 또한 제공된 정보가 신뢰할 수 있는 출처·근거에 기반하고 있는지를 평가하는 지표

주요 속성

❶ **내용 정확성**: 라벨(Label) 및 데이터 값의 사실·의미 부합 여부
❷ **데이터 신뢰성**: 명확한 출처·근거 제시 및 허위·오류 정보 유입 최소화

고려사항

▸ **샘플링 검수**: 대용량 데이터 특성상 전수검사 한계를 보완하기 위해 통계적 샘플링과 정답(Ground Truth) 데이터셋을 활용한 검증이 일반적
▸ **정답 데이터셋**: 데이터 구축 시 정답 데이터셋을 함께 구축하여 모델의 정확성을 지속적으로 평가할 수 있는 기반 제공 필요

※ 정답(Ground Truth) 데이터셋 : AI 모델이 학습하고 성능을 평가할 때 '정답으로 간주되는 값·라벨을 사전에 검증하여 구축한 기준 데이터셋'을 말함

[표 21] 정확성 ❶ 내용 정확성 품질 지표

내용 정확성 품질 지표		
기준 정합성	정의	▸ 데이터의 값이 법령, 고시, 지침, 표준 코드표 등 공신력 있는 외부 기준과 의미적·수치적으로 일치하는지 평가하는 지표
	평가 목적	▸ AI가 기준과 상충되는 데이터(예: 철 지난 지원금액, 폐기된 법정동 코드)를 학습하여 이용자에게 사실과 다른 정보 또는 오류를 제공할 위험을 최소화
	진단기준	▸ 해당 데이터의 기준 시점(Base Date)에 유효한 법령·고시의 기준값 및 표준 코드 목록과 대조하여 일치 여부 판정

내용 정확성 품질 지표		
기준 정합성	진단방법	▸ **기준 테이블 비교**: 기준 테이블 구축 또는 실시간 검증 방식을 통해 공신력 있는 기준 정보가 진단 대상 데이터와 일치하는지 검증
	오류예시	▸ **정보 안내 오류**: 2025년 고시된 수당 인상 금액 미반영으로 이전 연도 수당 금액 안내
규칙 정확성	정의	▸ 정책 요건이나 비즈니스 로직 기준에 따라 여러 속성을 종합하여 판단한 결과 값(분류, 판정, 등급 등)의 경우 사전에 정의된 업무 규칙(Business Logic)과 산출 로직이 정확히 일치하는지 평가하는 지표
	평가 목적	▸ AI 모델은 입력 변수(원인)와 결과 변수(결과) 사이의 관계(패턴)를 학습하므로, 논리적으로 잘못된 판단 데이터를 제거하여 오작동 및 편향 위험을 최소화
	진단기준	▸ '소득+재산 → 대상자 선정'과 같은 업무 시나리오별 결정 규칙(Business Rule)을 검증 쿼리로 구현하여 위배 여부 확인 ※ 데이터셋 내에 입력 변수와 결과 변수 모두가 존재하는 경우로 시나리오 로직 자체가 아니라 데이터의 논리적 모순을 검증하는 것
	진단방법	▸ 정책 지침서상의 판단 로직을 쿼리문이나 스크립트로 구현하여, 실제 저장된 값과 로직상 도출된 값이 다른 경우를 검출
	오류예시	▸ **판정 오류**: 소득 및 재산 기준은 모두 충족 했으나, 결과 컬럼에 불충족으로 기록된 경우
라벨 정확성	정의	▸ 지도학습의 정답지 역할을 하는 라벨이 실제 정답(Ground Truth)과 일치하는지 여부를 의미하며, 정답(Ground Truth) 수준의 품질을 확보했는지 평가하는 지표
	평가 목적	▸ 부정확한 라벨은 모델 성능 저하의 주요 원인이 될 수 있으므로, 전문 검수 인력 또는 검증된 기준 셋을 통해 학습 데이터의 무결성을 확보
	진단기준	▸ 샘플링 데이터에 대해 전문가 검수 결과(정답)와 데이터셋의 라벨 값이 일치하는 비율이 사전에 설정한 허용 오차 범위 이내인지 진단
	진단방법	▸ **교차 검증**: 전체 데이터 중 표본(예: 5%~10%)을 무작위 추출하여 숙련된 검수자가 검수를 수행하며, 필요시 다수 검수자가 교차 검증을 수행함 ▸ **기준 데이터셋**: 사전에 구축된 골드 스탠다드(Gold Standard) 데이터셋과 현재 라벨을 대조하여 정밀도(Precision) 및 재현율(Recall) 등 혼동 행렬 기반의 정량적 정확도를 산출
	오류예시	▸ **불법주정차 오분류**: 실제로는 '정상 주차' 구역에 있는 차량 사진에 '불법 주정차' 라벨링

[표 22] 정확성 ❷ 데이터 신뢰성 품질 지표

데이터 신뢰성 품질 지표		
출처 신뢰성	정의	▸ 데이터의 수집 또는 생성 기원이 투명하게 관리되고, 해당 데이터의 출처가 정책 영역·업무 기능·법적 소관에 부합하는 공공기관인지를 평가하는 지표
	평가 목적	▸ 출처 불명의 데이터 학습으로 인한 지식 재산권 분쟁 및 사실과 다른 정보생성·제공 위험을 최소화하고, 향후 AI 판단 결과에 대한 설명 책임과 투명성을 확보
	진단기준	▸ 데이터셋의 메타데이터에 발행 기관, 원천 시스템, 참조 링크(URL)정보가 표준 규격에 맞춰 누락 없이 기재되었는지 확인 ▸ 기재된 발행 기관이 해당 도메인 데이터 생산에 대한 공식적 권한을 보유하는지 확인

데이터 신뢰성 품질 지표		
출처 신뢰성	진단방법	▸ **메타데이터 표준 검증**: 메타데이터가 데이터 카탈로그 표준 필드에 따라 작성되었는지 확인하고, 발행 기관, 원천 시스템, 참조 링크 등 출처 관리에 필요한 필수 항목이 누락 없이 기재되어 있는지 점검 ▸ **소관 기관 적합성 검증**: 메타데이터에 기재된 발행 기관이 해당 데이터의 정책 영역·업무 기능·법적 소관에 부합하는 공공기관인지 여부를 확인하여, 데이터 내용과 발행 주체 간의 책임 연계성이 확보되어 있는지 점검
	오류예시	▸ **비공식 출처**: 의료 데이터를 보건복지부나 의학회의 공식 자료가 아닌 검증되지 않은 개인 블로그 또는 위키피디아 게시글을 활용해서 수집된 경우 ▸ **검증 불가**: 출처가 명시되어 있으나 제공된 URL이 삭제된 페이지(404 Error) 이거나 보안으로 접속이 불가한 등 데이터의 원본을 대조·확인할 수 없는 경우

2.3.2.4 적시성

정의

데이터가 현재 시점의 현실 세계 정보를 시의적절하게 반영하며 시간 경과에 따른 데이터 분포 및 개념의 변화에 대한 체계적인 모니터링과 관리를 이행하는지 평가하는 지표

주요 속성

❶ **최신성**: 데이터 발생 후 제공까지의 지연을 최소화하고, 사전에 정의된 갱신주기를 준수하여 데이터의 최신성과 적시성을 보장
❷ **개념·분포 변화 관리**: 데이터·개념 분포 변화 탐지 및 관리

[표 23] 적시성 ❶ 최신성 품질 지표

최신성 품질 지표		
최신성	정의	▸ 데이터가 현재 시점의 현실 세계 정보를 시의적절하게 반영하고 있으며, 사전에 정의된 갱신 주기에 맞춰 최신 상태로 유지되고 있는지를 평가
	평가 목적	▸ 시간이 경과함에 따라 데이터의 가치가 하락하거나 사실관계가 변하는 것을 관리
	진단기준	▸ 데이터의 갱신 주기(예: 실시간, 일 단위, 월 단위, 연 단위 등)가 사전에 정의되어 있으며 데이터의 최종 갱신 시점이 정의된 갱신 주기에 부합하는지 확인
	진단방법	▸ **갱신 주기 준수 점검**: 메타데이터에 명시된 갱신 주기와 데이터의 최종 수정일을 비교하여, 현재 시점 기준으로 갱신 주기 초과 여부 및 초과 기간을 산출 ▸ **유효 정보 대조**: 데이터와 현재 유효한 법령, 행정규칙, 표준코드 등 기준 정보를 대조하여 만료·폐기된 정보 포함 여부 및 비율을 확인

[표 24] 적시성 ❷ 개념·분포 관리 품질 지표

개념·분포 관리 품질 지표		
데이터 연속성	정의	▸ 데이터가 시간 흐름에 따라 지속적으로 수집·관리되고 있으며, 기준·분류·코드 체계가 변경되어도 과거 데이터와 현재 데이터를 함께 비교·분석할 수 있게 데이터 의미와 구조가 일관되게 유지되고 있는지를 평가
	평가 목적	▸ 시계열적 특성의 단절로 인한 AI 모델의 추론 강건성 저하를 방지하고, 정책·제도·행정 환경 변화 전후의 데이터를 일관된 맥락에서 결합·분석할 수 있도록 장기간 시계열 분석 가능성을 확보하기 위함
	진단기준	▸ **물리적 연속성:** 데이터가 시간 축 상에서 정의된 주기(일·월·연 등)에 따라 지속적으로 축적되어 있으며, 결측 구간, 장기 공백, 비정상적 수집 중단 등 시계열 단절 구간이 존재하는지 여부 ▸ **논리적 연속성:** 정책 변경, 행정구역 개편, 분류체계 개편 등으로 코드 변경 시 연계를 위한 매핑 정보가 제공되는지 여부
	진단방법	▸ **시계열 연속성 점검:** 데이터의 시간 필드(기준일, 발생일 등)를 기준으로 연속된 시계열 구간을 분석하여 결측 구간, 비정상적 공백을 확인 ▸ **코드 변경 관리 점검:** 코드 정의서 내 변경 이력을 확인하여, 이전 코드와 신규 코드 간 연계 정보가 문서화되어 있는지 확인
	오류예시	▸ **시계열 단절:** 시스템 장애로 저장되지 못한 데이터가 있는 경우. 예를 들어, 설 연휴 기간 데이터 누락으로, 명절 트렌드를 학습할 수 없는 경우 ▸ **맥락 단절 - 행정구역:** 'A동'이 'B동'으로 통합되었으나 데이터상 연결 고리가 없어, AI는 'A동'이 소멸하고 'B동'이 신설된 것으로 잘못 인식하여 인구 추이 분석 시 오류가 발생할 우려가 있는 경우

2.3.2.5 유효성

정의

AI가 오류 발생을 최소화하며 읽고 처리할 수 있도록 형식과 정의된 스키마, 데이터 유형, 허용 범위, 제약 조건 등을 준수하는지 평가하는 지표

주요 속성

❶ **구조·형식 유효성:** 스키마 준수 및 기계 이해성(Machine Understanding) 확보
❷ **값·범위 유효성:** 데이터 입력 규칙 및 허용 범위 준수

[표 25] 유효성 ❶ 구조·형식 유효성 품질 지표

구조·형식 유효성 품질 지표		
형식 유효성	정의	▸ 데이터가 사전에 정의된 형식(데이터 타입, 길이, 날짜·숫자 형식 등)을 준수하여 기계가 오류 없이 읽고 처리할 수 있는 상태인지를 평가하는 지표
	평가 목적	▸ AI 학습·전처리 과정에서 형식 오류로 인한 파싱 실패, 처리 중단, 데이터 손실을 최소화
	진단기준	▸ 데이터가 원천 데이터베이스(DB) 또는 데이터 생성 단계에 정의된 데이터 타입·형식 규칙을 일관되게 따르고, 날짜, 숫자 코드 등 주요 필드가 기계 처리 과정에 오류 없이 해석가능한 형식으로 구성되어 있는지 여부

구조·형식 유효성 품질 지표		
형식 유효성	진단방법	▸ **형식 규칙 기반 검증:** 원천 DBMS에 정의된 데이터 타입, 길이 제한, 형식 규칙 또는 기존 검증 스크립트를 기준으로 데이터를 점검하여 형식 불일치 데이터 비율을 산출 ▸ **파싱 오류 점검:** CSV, JSON 등 파일을 표준 처리 도구(데이터 분석 도구, 스크립트, ETL 도구 등)를 통해 날짜·숫자 필드로 읽어들이는 파싱 과정에서 오류가 발생하는 레코드를 검출
	오류예시	▸ **날짜 오류:** YYYY-MM-DD 필드에 2024-13-40(월, 일 범위 초과) 또는 24.01.01(형식 불일치) 입력된 경우 ▸ **타입 오류:** 나이(Integer) 필드에 숫자 대신 `Unknown`, `-` 등의 문자열이 포함된 경우
기술적 유효성	정의	▸ 텍스트, 이미지, 음성, 영상 등 비정형 데이터 파일이 손상되지 않았으며, AI 학습에 필요한 최소한의 기술적 사양(해상도, 길이, 용량 등)을 충족하고 있는지 평가하는 지표
	평가 목적	▸ 파일 헤더 손상으로 열리지 않거나, 해상도·음질이 현저히 낮아 AI 인식·분석 성능이 저하되는 문제를 사전에 식별·조치
	진단기준	▸ 파일 포맷별로 정상적인 열기(Open/Read)가 가능한지 여부 및 해상도, 재생 길이, 용량 등 사전에 정의된 최소 기술 요건 충족 여부 확인
	진단방법	▸ **자동화 점검:** 자동화 스크립트를 통해 파일 열기(Open/Read) 시도 시 오류 발생률과 기술 사양 미달 비율을 측정
	오류예시	▸ **파일 손상:** 확장자는 .jpg이나, 파일 헤더가 손상되어 열리지 않는 이미지 파일 ▸ **규격 미달:** 자율주행 학습용 이미지의 해상도가 50×50 픽셀로 낮아 사물 식별이 어려운 경우

[표 26] 유효성 ❷ 값·범위 유효성 품질 지표

값·범위 유효성 품질 지표		
수치 범위 유효성	정의	▸ 연속형 수치 데이터가 사전에 정의된 허용 구간이나 제약 조건을 준수하는지 평가하는 지표
	평가 목적	▸ 물리적으로 불가능하거나 업무 규칙상 허용되지 않는 값(예: 음수 나이)이 학습되어 AI 모델의 예측 결과가 왜곡되는 위험을 최소화
	진단기준	▸ 컬럼별로 정의된 최솟값·최댓값, 음수 허용 여부 등 기본 제약 조건 위배 데이터 비율
	진단방법	▸ 컬럼 단위로 허용 범위를 벗어난 값의 건수를 산출
	오류예시	▸ **구간 오류:** 나이 필드에 -5가 입력되거나, 월(Month) 필드에 13이 입력된 경우 ▸ **제약 오류:** 지원 금액에 음수(-10000) 입력, 습도 센서값이 200% 기록
통계적 타당성	정의	▸ 데이터 분포상 통계적으로 발생하기 어려운 극단적인 값이나, 의미 없는 기본값·임의값이 포함되어 있는지를 평가하는 지표
	평가 목적	▸ 999999, -999와 같은 더미값이나 극단적 이상치로 인해 AI 모델의 평균, 가중치, 학습 결과가 왜곡되는 위험을 최소화
	진단기준	▸ 데이터 분포 기준으로 비정상적으로 튀는 값(이상치) 또는 반복적으로 입력된 특정 값(예: 999999, 0000 등)의 과도한 빈도 여부
	진단방법	▸ 데이터 분포 분석(Histogram, Boxplot)을 통해 일반적인 범위를 크게 벗어난 값이 존재하는지 점검하거나 특정 패턴의 반복 빈도를 분석
	오류예시	▸ **더미 값:** 입력자가 임의로 입력한 999999, 0000 등이 존재하는 경우 ▸ **극단치:** 성인 남성 키 데이터에 10cm나 300cm 등 비현실적 값 포함

2.3.2.6 유일성

정의

데이터셋 내 불필요한 중복이 존재하지 않도록 관리하여 특정 패턴에 대한 과적합(Overfitting) 위험을 최소화하고, 학습(Train)·검증(Validation)·평가(Test) 데이터셋 간 중복을 제거하여 데이터 누수 발생을 방지했는지 평가하는 지표

주요 속성

❶ **유일성:** 데이터 중복에 따른 과적합 및 평가 왜곡 위험 최소화

[표 27] 유일성 ❶ 유일성 품질 지표

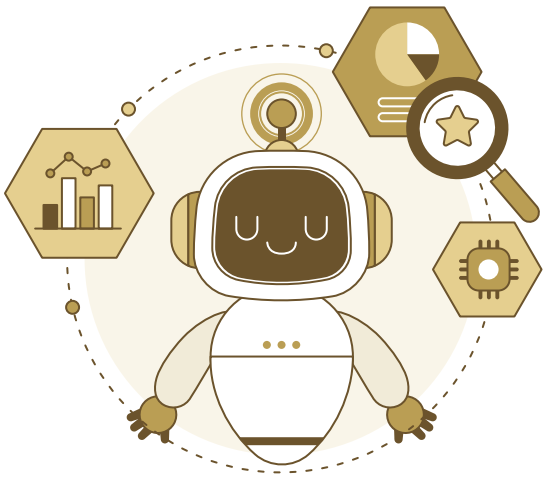
유일성 품질 지표		
데이터 유일성	정의	▸ 동일한 개체(사람·기관·사업체 등)나 사건을 식별하는 고유 식별자(Key)를 기준으로, 시스템 오류 등으로 인해 동일한 내용의 레코드가 중복저장 되어 있지 않은지 평가
	평가 목적	▸ 동일 데이터가 반복 학습되어 AI 모델의 판단이 특정 데이터에 과도하게 치우치는 현상(가중치 왜곡, 과적합)을 방지하고, 학습 데이터와 성능 평가 데이터 간에 동일한 데이터가 동시에 포함되는 데이터 누수(Data Leakage) 위험을 예방
	진단기준	▸ 개체 고유성(Entity Uniqueness)원칙에 따라, 실제 세계의 단일 개체가 데이터셋 내에서 단 하나의 유효한 식별자(Unique ID) 및 레코드로 존재하며 물리적 중복성 제거된 상태 평가
	진단 방법	▸ 사업자번호, 시설ID, 사건 번호 등 개체 식별용 키 값을 기준으로 중복 레코드 존재 여부를 확인하고, 건수 및 중복 비율을 산출
	오류예시	▸ **중복 데이터:** 사업자번호 '123-45-67890'를 가진 'A식당' 정보가 시스템 오류로 동일 레코드 3건으로 중복 저장된 경우
유사 중복 제어	정의	▸ 텍스트·이미지의 겉보기 표현은 일부 다르지만, 내용과 의미가 사실상 동일하거나 매우 유사한 데이터가 과도하게 반복 포함되어 있는지 평가
	평가 목적	▸ 표현만 다른 유사 데이터가 대량 포함되어 AI가 특정 표현이나 상황에 편향하여 학습할 위험을 완화
	진단기준	▸ 데이터 간 의미 또는 시각적 유사도를 측정하여, 사전에 설정한 임계치를 초과하는 유사 데이터 집합이 과도하게 존재하는지를 판단
	진단 방법	▸ **텍스트:** 문장 내 중복된 단어 비율을 비교하거나 문장의 의미를 수치로 변환(임베딩)하여 유사도 분석 ▸ **이미지:** 이미지의 전체적인 형태와 특징을 요약한 값(퍼셉추얼 해시)을 비교하여 유사 이미지 그룹 식별
	오류예시	▸ **민원 중복:** "길가 차 정차"와 "길가에 차가 계속 서 있어요"는 표현은 다르지만, '불법주정차 민원' 의미를 가지는 유사 중복 데이터로 판단될 수 있음 ▸ **뉴스 중복:** 동일한 사건에 대해 언론사별로 조사(은/는/이/가)만 다른 기사가 수천 건 수집될 수 있음

2.3.3 AI 모델 기반 데이터 품질 평가

AI 모델의 추론 결과 및 성능 지표를 분석하여, 데이터셋 내 정보 누락, 라벨의 부정확성, 특정 상황·사례에 대한 편향 여부 등을 동적으로 진단하는 데이터 품질 평가 방식

[표 28] 모델 기반 데이터 품질 평가

완전성	▶ **지식·문맥 완전성** AI가 답변을 생성하는 데 필요한 정답(Ground Truth) 근거 정보가 데이터셋 또는 지식베이스에 충분히 포함되어 있는지 평가 • **예시** 검색 기반 생성(RAG) 시스템에서 사용자의 질문에 대해 관련 법령·지침·매뉴얼 등 정답 근거 문서가 검색되지 않는 경우, 데이터셋 또는 지식베이스에 정보 공백(Knowledge Gap) 이 존재
정확성	▶ **답변 신뢰성 및 근거 부합성** AI 모델이 생성한 결과 값 또는 질의응답 결과가 해당 결과가 참조한 근거 문서(법령, 고시, 지침, 통계 자료 등)와 의미적으로 일치하는지를 평가 및 LLM·RAG 응답에서 사실과 다른 내용이 생성되거나 근거 없이 내용을 추론·창작하는 현상(비사실적 생성, Hallucination)이 발생하는지 여부를 통해 데이터 또는 지식베이스의 신뢰성 부족 가능성을 간접적으로 점검 • **예시** 청년 전세자금대출 이자율에 대해 시스템이 산출하거나 답변한 값이 실제 공식 고시문 또는 관계 부처 발표 자료에 기재된 이자율과 정확히 일치하는지 비교
유효성	▶ **데이터 처리 적합성 및 기술적 유효성** 인코딩, 토큰화(Tokenization), 파싱 등 전처리 및 처리 과정에서 발생하는 오류 레코드 또는 입력 불가 데이터의 비율을 측정하여 데이터의 기술적 처리 가능성을 진단 • **예시** LLM 입력 과정에서 문자 인코딩 깨짐 또는 제어문자 포함으로 인해 토큰화에 실패한 텍스트 비율, 이미지 인식 모델에서 규격 미달 또는 파일 손상으로 인해 입력 단계에서 제외된 이미지 샘플 비율



2.4 표준 코드

공공데이터 연계·결합 시 데이터 상호운용성 및 정확성 향상과 데이터 내용의 일관성 유지를 위해 적용되는 표준화된 코드 체계

- 표준코드 적용 시 기관·시스템 간 데이터 연계가 가능하여 표기 다양성에 따른 모호성과 오류가 감소하여 데이터의 품질 향상
  - 표준코드는 동일 의미인데 다양한 표기가 혼재된 내용(값) 문제를 제거하여 모델 결과의 정확성을 향상하는 중요한 요소
    - 행정구역 등 내용(값)의 명칭이 변경되더라도 코드는 유지되므로, 과거-현재 데이터 간 시간적 일관성을 확보할 수 있어 장기 운영과 버전 관리에 용이
    - 국제적으로 통용되는 코드 사용 시, 추가 전처리 없이 국가 간 데이터 연계 가능
  - 다른 기관의 데이터 연계·활용 시 추가 전처리 등 비효율 발생 방지를 위해 공통 표준, 기관 표준을 적용하여 공공데이터 구축 필요

※ 공공데이터베이스 표준화 관리 매뉴얼 P.15 공공데이터베이스 표준 관리

참조 공공데이터베이스 표준화 관리 매뉴얼 – 공공데이터베이스 표준 관리

- 공공기관의 장은 신규로 공공데이터베이스를 구축할 때에는 반드시 공통 표준, 기관 표준, 행정표준코드 등을 적용해야 한다.
- 기존에 운영 중이며 표준이 적용되지 않은 공공데이터베이스는 해당 정보시스템을 전면 재구축하기 전까지 비표준 데이터를 표준 데이터와 매핑하여 표준화된 데이터로 연계·활용할 수 있도록 관리해야 하며, 재구축 등 데이터베이스 개선 업무 추진 시 공통 표준을 적용해야 한다.
- 공공기관의 장은 기관 표준을 새롭게 제정하거나 개정된 경우, 또는 기관 표준의 공동 활용이 필요하다고 판단되는 경우 표준 관리 통합시스템을 통해 행정안전부장관에게 신청할 수 있으며 등록된 정보는 최신성을 유지해야 한다.

- 공공데이터는 행정표준코드관리시스템(code.go.kr)의 표준코드를 우선 적용하고, 국제 표준코드가 있는 경우 국제 기준 사용 권장
- 공개 범위 고유식별 정보(가명처리 또는 비식별화 조치된 주민등록번호, 사업자등록번호 등)는 관련 법령에서 정한 표준 체계를 준수하여 적용
- 표준코드의 기준과 출처는 메타데이터와 문서에 명확히 표기하여, 데이터에 적용된 코드에 대한 이용자의 명확한 이해와 활용 지원

참고 행정표준코드관리시스템 내 표준코드 – 기관, 법정동, 직급[7]

구분	예시	
	코드	명칭
기관	0000001	▶ 대통령
	0000003	▶ 입법부
	1741000	▶ 행정안전부
	B554393	▶ 한국지능정보사회진흥원
법정동	1100000000	▶ 서울특별시
	4100000000	▶ 경기도
	3611000000	▶ 세종특별자치시
	2700000000	▶ 대구광역시
직급	00092	▶ 행정사무관(행정부국가공무원, 일반)
	00911	▶ 지방행정주사보(행정부지방공무원, 일반)
	01777	▶ 기능10급 철도원(행정부국가공무원, 기능)
	04772	▶ 시·도지사(행정부지방공무원, 정무직)

참조 공공데이터베이스 표준화 관리 매뉴얼 – 공통표준 구성요소

▶ 공통표준용어, 공통표준단어, 공통표준도메인 구성하며 해당 요소를 대상으로 정의 기준, 방법, 절차 및 관리 항목을 정의하여 우선적으로 적용

구성요소	설명
공통표준용어	▶ “공통표준단어”와 “공통표준도메인”의 조합으로 생성 ▶ 공통표준단어의 조합으로 공통표준용어명이 생성되고, 공통표준도메인에서 정의한 데이터 형식 반영
공통표준단어	▶ 공통표준단어명, 공통표준단어영문명, 공통표준단어영문약어명, 공통표준도메인분류명 등으로 구성 ▶ 업무용어에 기반한 최소단위의 명사형 단어
공통표준도메인	▶ 공통표준도메인분류명, 공통표준도메인명, 데이터 타입 및 길이, 저장·표현 형식 등으로 구성

7 행정표준코드관리시스템. (n.d.). 행정표준코드관리시스템. https://www.code.go.kr

CHAPTER 03

인공지능 친화적 공공데이터 관리원칙

3.1 데이터 책임 주체 및 관리체계 수립

공공데이터의 품질·보안·윤리·업데이트 등 관리 전반에 대한 책임 주체의 명확화와 부문별 체계적 관리 절차 수립

- 데이터의 품질, 보안, 윤리, 업데이트 관련 최종 책임 주체를 명확히 정의하고 민원, 오류 등 데이터와 관련된 문제 발생 시 빠르게 조치할 수 있도록 체계 마련 필요
- 각 책임 주체(품질, 보안, 윤리, 업데이트)는 데이터를 체계적으로 관리하고 공개하기 위한 절차 및 기준에 대해 각 주체가 협력할 수 있는 방식으로 고려하여 수립 필요
  - ▸공공기관이 AI 시스템을 개발, 도입, 운영하는 경우, 본 가이드라인의 인공지능 친화적 기준을 충족한 데이터셋을 우선적으로 활용하도록 권고

※ 공공데이터베이스 표준화 관리 매뉴얼 P.77 예방적 품질관리 검토 항목별 이행점검 기준

참조 공공데이터베이스 표준화 관리 매뉴얼 –
예방적 품질관리 검토항목 별 이행점검 기준 중 데이터 관리체계

진단사항		결과 이행점검 산출물
데이터 품질관리 체계	▸데이터 품질관리를 위한 정책, 절차, 조직(담당자) 관련 사항 반영 여부 ▸데이터 품질관리를 위한 필수산출물 등 관련 사항 반영 여부	❶ 데이터 품질관리 체계를 확인할 수 있는 산출물(매뉴얼 등) ※ 데이터 품질관리 정책 및 절차, 조직 등의 운영 계획 (또는 운영관리)에 관한 문서 ❷ 데이터 품질관리 체계 수립을 수행하였는지 확인할 수 있는 사업 감리보고서 또는 사업 완료 보고서 등 ➜ ❶, ❷ 중 택1 제출(범정부 EA포털에 자료 등록)
데이터 개방 및 메타데이터 관리체계 수립	▸개방 예정 데이터 목록 등 기관의 공공데이터 개방계획과 본 사업의 연관성 관련사항 반영 여부	❶ 데이터 개방을 수행한 내역을 확인할 수 있는 산출물 (아래 목록 중 택1) ▸구축 시스템의 데이터 개방관리 체계를 확인할 수 있는 문서 (개방을 위한 별도 DB 구축, API 개발 등) ▸공공데이터 포털, 기관 개별데이터 포털을 통해 데이터 개방을 수행한 내역(캡쳐화면 등) 등 공공데이터 개방을 수행한 결과 ❷ 데이터 개방을 수행한 내역을 확인할 수 있는 사업 감리 보고서 또는 사업 완료 보고서 등 ➜ ❶, ❷ 중 택1 제출(범정부 EA포털에 자료 등록)
	▸공공데이터베이스의 메타데이터 등록, 관리 및 현행화 관련사항 반영 여부	❶ 메타데이터 등록 및 현행화를 확인할 수 있는 산출물 (아래 목록 중 택1) ▸메타데이터관리시스템에 구축 시스템의 메타데이터 등록 내역(화면 캡처 인정) ▸(메타데이터관리시스템이 없는 경우) 구축 시스템의 메타 데이터 항목을 작성한 문서

진단사항		결과 이행점검 산출물
데이터 개방 및 메타데이터 관리체계 수립	▸ 공공데이터베이스의 메타데이터 등록, 관리 및 현행화 관련사항 반영 여부	❷ 메타데이터 등록 또는 현행화를 수행한 내역을 확인할 수 있는 사업 감리보고서 또는 사업 완료 보고서 등 ➜ ❶ , ❷ 중 택1 제출(범정부 EA포털에 자료 등록)

3.1.1 데이터 검증 절차 및 공유 품질 기준 수립

3.1.1.1 데이터 품질 검증 절차

✔ 데이터 품질 검사 기준과 도출된 측정값을 공개하여 이용자에게 데이터에 대한 신뢰성 제공

※ **참조:** 공공데이터 품질관리 매뉴얼 Ⅱ. 데이터 품질관리

▸ 데이터 자체의 품질뿐 아니라 메타데이터에 대한 품질도 함께 평가, 데이터 이용자와 인공지능 모델에게 데이터에 대한 정확한 정보를 제공하여 적용될 수 있도록 절차 및 기준 마련 필요

참조 공공데이터 품질관리 매뉴얼 (II. 데이터 품질관리 내용 참조)

- ▸ **데이터 표준 수립, 적용 및 점검:** 데이터 품질을 보장하기 위한 활동으로 데이터 정보 요소의 명칭, 정의, 형식, 규칙에 대한 원칙을 수립하여 적용하고 점검하는 활동
- ▸ **데이터 구조의 일관성 확보 및 오류 데이터 입력 방지:** 데이터 중복 최소화 하도록 데이터 모델을 설계, 구축하고 불가피한 중복이 발생하는 경우 중복 데이터의 일관성이 확보되도록 관리하는 활동
- ▸ **연계 데이터 정합성 관리:** 데이터 보유기관과 이를 활용하는 기관 간의 연계 데이터에 대한 표준화된 관리 및 최신성 확보를 통해 연계 데이터의 정합성 확보
- ▸ **데이터 산출물 관리:** 데이터 품질과 관련된 산출물 생성, 검증, 진행사항을 반영하는 활동

✔ 품질 지표를 정량적으로 관리하여 실시간 모니터링을 통해 데이터 품질의 투명성이 확보될 수 있도록 '핵심 품질 KPI'를 정의하고 '공개 대시보드 구축'을 통해 운영 필요

▸ **핵심 품질 KPI 정의:** 데이터 품질 속성을 기반으로 핵심 품질 지표(KPI)를 정의

▸ **모니터링 체계 구축:** 주요 품질 KPI 현황을 실시간으로 모니터링하고 시각화하는 공개 대시보드를 구축하여 품질 상태의 투명성을 확보

✔ 데이터 지속가능성 확보를 위해, 다중 백업과 복원 테스트를 포함한 백업·보존 절차를 마련하여 품질 검증 체계에 반영 필요

[표 29] 다중 백업 및 복원 테스트

구분	내용
다중 백업	▸ 데이터 손실 방지를 위해 최소 2개 이상의 물리적으로 분리된 위치 (예: 다른 클라우드 리전(Region))에 백업을 보관
복원 테스트	▸ 백업된 파일이 온전한지 해시 값으로 검증하는 것뿐만 아니라, 실제 상황을 가정한 전체 복원 테스트를 수행하여 백업 정책의 실효성 검증

3.1.1.2 공공데이터 공유 시 품질 요건

- 공유 플랫폼의 데이터 신뢰성을 유지하고 AI 모델의 학습 적합성을 보장하기 위해, 모든 공공데이터는 등록 시점의 최소 요건을 충족해야 하며 등록 후에도 주기적인 품질 관리가 필요
- 공유 데이터는 다음의 4가지 요건을 충족하는 것을 원칙으로 함 (*기준 예시)
  - **내용 완전성:** 데이터셋 내 결측치 비율이 20% 미만이어야 하며, 핵심 정보의 누락이 없어야 함
  - **메타데이터 충실성:** 아래 명시된 표준 메타데이터 필수 항목은 모두 기재하여 데이터의 탐색 가능성을 확보
  - **최신성 유지:** 데이터 발생 후 제공까지의 지연을 방지하기 위해, 최근 갱신일로부터 갱신 주기가 2회 이상 경과하지 않아야 함
  - **기술 표준 준수:** AI 이해 효율성을 위해 플랫폼 표준 포맷(CSV, JSON, XML, Parquet 등) 중 하나를 준수함
- 데이터 검색·활용·AI 학습 적합성 판단을 위해 다음의 항목들을 메타데이터에 포함할 것을 권장

[표 30] 표준 메타데이터 필수 항목

메타데이터 항목	설명
관리 기관	▸ 원천 데이터 보유 기관
생산 주기	▸ 일/주/월/분기/연/비정기
최근 갱신일	▸ 실제 데이터 기준(등록일 아님)
데이터 기간	▸ 수록 데이터의 시작~종료 기간
레코드 수	▸ 등록 시점 기준 건수
컬럼 수 및 설명	▸ 각 컬럼의 명칭, 데이터 타입, 설명
데이터 등급	▸ C/S/O 내지 1/2/3 등급
비식별 처리 여부	▸ 처리 방법 포함
활용 허용 범위	▸ 열람/분석/결합/재배포 허용 여부
관련 법령	▸ 근거 법률 및 제공 제한 법률
데이터 관리 담당자	▸ 담당 부서 및 담당자 연락처

- 공유 플랫폼에 등록된 데이터의 품질을 지속 관리하기 위해 연1회 품질 점검을 실시
  - 우수 기관에 대해 품질 인증을 부여하고, 미흡 기관에 대해 데이터 공유 제한, 품질 개선 조치 요청 등 가능

3.1.2 데이터 유지관리 절차

- 최신성 및 정확성이 확보된 데이터 관리를 위해 데이터 유지관리 담당 주체별 권한과 역할을 명확히 하고 문서에 명시하며, 이를 기반으로 한 유지관리 절차와 기준을 마련[8)]
  - 데이터 유지관리 과정에서 데이터 담당 기관, 시스템 운영기관, 인공지능 모델 운영기관이 각자의 역할과 권한을 나누어 책임 있게 관리하며 기관별 유지관리 기준에 대한 명확한 설명 필요
  - 데이터는 변경·갱신이 반복되므로, 데이터 구조와 기준을 일관되게 업데이트하여 기존 이용자의 데이터 파이프라인에 영향을 주지 않도록 관리 필요
  - 변경 사항 발생 시 변경 사항을 명확히 메타데이터와 문서로 안내하여 이용자 재사용성을 높이고, 필요한 경우 변경 내용을 시각적으로 제공
  - 데이터 갱신 시 스키마 준수 여부, 이상치 비율 등 핵심 품질 지표를 자동으로 점검할 수 있도록 절차에 포함
  - 품질 기준에 미달 시 자동으로 공개를 보류하는 '품질 게이트(Quality Gate)'를 데이터 파이프라인에 적용하여 안정적인 운영을 지원

3.1.3 데이터 생애주기 및 이력 관리

- AI에 사용되는 데이터는 분석의 재현성을 보장하기 위한 이력관리 체계와 데이터의 최신성 유지가 중요.
- 데이터의 활용 목적과 성격에 따라 보존 기간을 차등 설정하며, 만료 시 정해진 절차에 따라 보존 연장 여부를 결정

예시 데이터 접근 유형별 보존기간

데이터 유형	보존 기간	연장 조건
정책 분석용 데이터	5년	▸기관 신청 시 5년 단위로 기간 연장 가능
행정 업무용 데이터	3년	▸행정 업무 종료 시점부터 기산하여 보존
시범·실험성 데이터	1년	▸정식 데이터로 등록 전환 시 보존 기간 재산정
법령 근거 데이터	법령에서 정한 기간	▸개별 법령에서 정한 보존 기간을 최우선으로 적용

8 U.S. Department of Commerce. (2025). Generative Artificial Intelligence and Open Data: Guidelines and Best Practices. p.23-24 ; Open Data Institute. (2025). A Framework for AI-ready Data. p.8, 14

- 보존 기간이 만료된 데이터는 해당 데이터 관리주체의 의사결정을 거쳐 안전하게 처리함
  - **의사결정:** 데이터 관리 주체는 보존 기간 만료 전에 해당 데이터의 연장, 이관, 또는 삭제 중 하나를 선택
  - **자동 조치:** 기한 내 응답이 없을 경우 공유 플랫폼은 자동으로 열람을 중단하며, 30일의 추가 유예 기간을 거친 후 최종 삭제 처리
  - **이력 관리:** 삭제 이력은 메타데이터 로그를 5년간 보관하여 관리 책임성을 유지하나 데이터 자체는 삭제
- 데이터 활용 기관이 과거 데이터 시점을 재현하고, 환각(할루시네이션) 없는 최신 정보를 제공하기 위해 변경 사항을 체계적으로 관리함
  - **버전 부여:** 데이터 갱신 시 기존 데이터를 유지하는 방식으로 부여. 고유한 버전 번호(V1.0, V2.0 등)를 부과하여 별도로 보존
  - **분석 재현성:** 활용 기관은 분석 목적에 따라 특정 버전을 고정하여 사용할 수 있어, 데이터 변화에 따른 결과값의 왜곡을 방지할 수 있어야 함
  - **변경 통보:** 컬럼의 추가·삭제·명칭 변경 등 주요 스키마 변경 시 메이저 변경으로 처리하고 활용 기관에 사전 통보
  - **자동 기록:** 버전별 변경 내용, 사유, 담당자 정보는 메타데이터의 버전 필드 등에 자동으로 기록되어 투명하게 공개

3.1.4 데이터 보안 및 윤리 절차와 기준 마련

- 데이터 접근 시 보안 강화를 위해 디지털 서명·암호화 등을 적용, 데이터의 원본성 검증 및 변조 방지 필요
- 비식별화, 암호화 등 보안 조치를 기준으로 각 이용자 권한에 따른 적정한 데이터 접근이 이루어지도록 절차 및 기준 마련 필요
- 데이터 윤리, 데이터 관련 법적 권리 등과 관련되어 요구되는 규범이나 조건을 준수할 수 있도록 절차와 기준 설정 시 적용

참조 공공부문 인공지능 윤리기준, 행정안전부(2025.11)

- **공공성:** 공공 AI 서비스를 공공의 이익과 국민의 복지를 위해 제공한다.
- **형평성:** 공공 AI 서비스를 모든 국민에게 공정하고 차별 없이 제공한다.
- **투명성:** 공공 AI의 도입과 활용에 대한 과정을 투명하게 공개한다.
- **책임성:** 공공 AI에 대한 책임 주체를 명확히 정하고, 제도적 기준에 따라 활용한다.
- **안전성:** 공공 AI 시스템이 국민에게 피해를 주지 않도록 안전한 방식으로 운영한다.
- **프라이버시 보호:** 공공 AI 시스템이 개인정보와 사생활 등을 침해하지 않도록 보호장치를 마련한다.

3.2 이용자 접근 편의성 인프라 구축

공공데이터 이용자의 데이터 탐색 편의성 제고를 위한 이용자 중심 인프라 환경 구축

- 이용자가 공공데이터에 쉽게 접근할 수 있도록 데이터 포털 등 인프라 환경이 구축되어야 함[9)]
  - 데이터에 대한 접근뿐만 아니라 데이터 집계를 위한 도구 세트, 데이터 품질을 탐색, 평가, 검증할 수 있는 도구도 함께 제공
  - 인프라 구성 시에는 이용자가 적용하고자 하는 AI 모델에 적합한 데이터를 찾기 유용하도록 구성
    - AI 개발자뿐만 아니라 모든 이용자가 데이터에 효과적으로 접근할 수 있도록 데이터 검색을 위한 다양한 방식 마련
    - 데이터 접근 방식은 최소 데이터 관련 전문가가 RESTful API를 통해 프로그래밍 방식으로 접근, 또는 누구나 직접 다운로드를 통해 데이터에 접근할 수 있도록 하는 방식 권장

    * REST(Representational State Transfer) 원칙을 따르는 웹 기반 데이터 접근 인터페이스로 인터넷 주소를 이용하여 데이터를 조회, 등록, 수정, 삭제하는 방식
- 데이터와 관련된 문서, 메타데이터를 기본적으로 한 번에 제공받을 수 있도록 설계하여 데이터 이용과 관련된 정보를 얻기 위해 여러 번 접근하지 않도록 편의성 제공
- 데이터 파일은 ZIP 등 압축 포맷을 적용하여 다운로드 속도를 개선하고, 대규모 AI 작업용 데이터는 Parquet 등 컬럼 지향 저장 포맷을 활용하여 효율적으로 제공

3.3 이용자 관점 데이터셋 평가 및 피드백 체계 구축

데이터에 대한 이용자 평가 및 요구사항 접수 창구 운영을 통해 데이터 문제점의 발굴과 개선 활동을 수행하고, 이용자에게 적합한 데이터 제공을 가능하게 하는 피드백 체계 수립

- 데이터 제공자가 모든 이용자의 요구를 사전에 반영하기 어려우므로 이용자 평가 기능을 제공, 데이터 이용 예정자가 데이터의 적합성을 신속히 판단할 수 있도록 평가구조 마련

  ※공공데이터 제공 관리 실무 매뉴얼 P.64 공공데이터 품질관리 평가 및 조치, 공공데이터 활용성 제고를 위한 환류 체계 운영 참고
  - 이용자 평가 기능은 별점, 품질 코멘트, 사용 사례 등으로 마련 가능

9 U.S. Department of Commerce. (2025). Generative Artificial Intelligence and Open Data: Guidelines and Best Practices. p.11 ; Open Data Institute. (2025). A Framework for AI-ready Data. p.8

참고 공공데이터 활용성 제고를 위한 환류 체계 운영

- 각 기관이 데이터를 활용하면서 발생하는 품질 오류 신고 및 처리 기준을 마련하고 이용자들의 요구사항을 심층 분석하여 기관 품질관리 계획(정책) 등에 환류
- **수요자 요구사항 접수 창구 운영:** 내부와 외부 이용자 대상으로 품질 오류 신고, 요구사항 등을 신청 접수할 수 있는 창구를 마련하고, 처리 절차 및 기준을 정의 및 관리
  ※ 일자별 신고 접수 내역과 처리결과를 기록하여 관리하고 처리결과 신청자에게 통지 의무화
- **수요자 요구사항 분석·활용:** 수요자 요구사항을 정기적으로 분석하여 요구사항 파악 및 문제점을 발굴하고, 개선 활동 수행
- 수요자 데이터 이용 실태(만족도 조사, 활용성이 높은 데이터 수요 조사)를 조사· 분석하여 정책에 반영 및 개선 수행

✔ 단순 평가뿐 아니라 데이터 제공자와 이용자가 피드백 할 수 있는 커뮤니티 등을 구축하여 지속적으로 데이터 문제점 발굴하여 개선할 수 있는 체계 마련

- 제공자-이용자 간의 데이터에 대한 피드백은 제공자가 발견하지 못한 품질 문제 등의 발견으로 이용자 친화적인 데이터 제공을 가능하게 하여 데이터셋의 활용성을 증대시키는 효과 존재

참고 Hugging face, Kaggle Datasets 피드백을 통한 데이터 평가 방식 예시[10]

- Hugging face의 경우 해당 데이터셋을 제공한 제공자에 대한 팔로워 수, 데이터셋에 표시된 '좋아요' 수를 기반으로 해당 데이터셋이 이용자에게 유용했는지를 판단할 수 있도록 간단한 평가 방식 제공
- Kaggle Datasets는 사용성 평가(10점 만점), 다운로드 수, 토론, 제안, 질문과 답변 등을 통해 데이터셋에 대해 평가하고 피드백할 수 있는 기능 제공

참고 아일랜드 오픈 데이터 게시 지침 – 이용자 커뮤니티와 협업 권장

- 데이터 공유의 모든 단계에서 이용자와의 소통은 데이터 재사용을 장려하고 궁극적으로 경제적, 사회적, 환경적, 정치적 영향을 촉진하는 데 필수적
- 가장 효과적인 접근 방식은 문제 또는 과제 기반 접근 방식으로, 데이터 이용자 및 영향을 받는 사람들과 협력하여 데이터를 통해 해결할 수 있는 문제를 파악하고, 이를 해결하는 데 필요한 데이터 접근성을 높이는 것. 이는 단순히 데이터 목록을 작성하는 게 아닌 과제에 집중하는 것으로, 데이터의 재사용 가능성을 높이는 데에 도움
- 이용자 커뮤니티와 협력하면 데이터를 중심으로 한 생태계의 유기적 성장이 촉진되고 대규모 영향력 전파 가능

권장 사항

- 게시된 각 데이터셋 대한 연락처 정보를 제공하며, 연락처 정보에는 개인이 아닌 역할과 관련된 일반 이메일 주소 포함
- 데이터 요청에 대한 토론에 참여
- 데이터셋에 대한 의견 공유
- 데이터셋에 대한 피드백을 듣고, 적절한 경우 이 피드백에 따라 조치
- 소셜 미디어, 뉴스레터, 이벤트 등 모든 커뮤니케이션 채널을 통해 오픈 데이터를 홍보
- 이용자 및 기타 게시자와 함께 문제/과제 기반 오픈 데이터 이벤트를 공동으로 조직[11]

10 Hugging face 홈페이지 ; Kaggle 홈페이지
11 Data.gov.ie. (n.d.). Ireland Open Data Publication Guidelines. https://data.gov.ie/guidelines

3.4 데이터 접근 권한 및 활용 체계

공공데이터의 책임성 확보를 위해 데이터 접근 권한을 차등 관리하고, AI 학습 및 결합 시 명시적 동의와 샌드박스 활용을 통해 데이터에 대한 보안을 강화

- 공공데이터의 안전한 활용을 위해 데이터 생산 및 관리 주체, 이용 목적에 따라 접근 권한을 체계적으로 관리해야 함. 기술 실무자와 데이터 제공 주체는 법·제도 및 규제 준수를 전제로 데이터의 적법성과 이용 권한을 확보

예시 데이터 접근 유형별 허용 데이터 등급

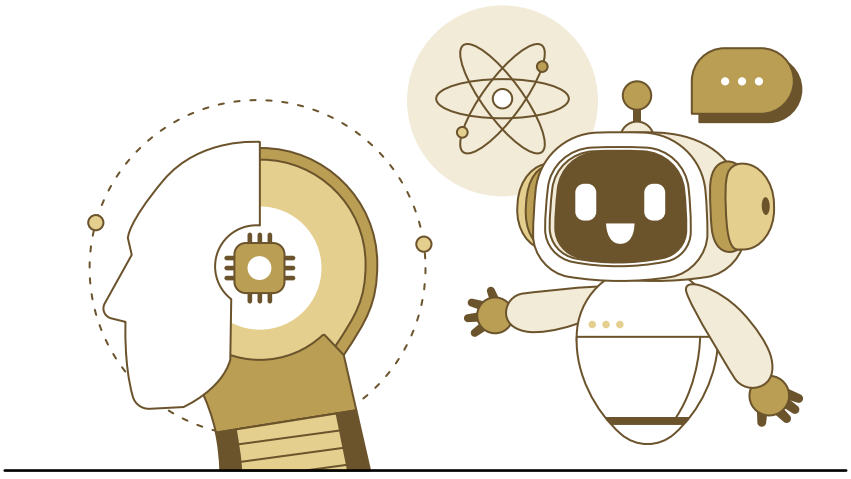
접근 유형	대상	허용 데이터 등급
전체 접근	▸국가공유데이터 플랫폼 관리 기관	▸S·O 내지 1·2등급 전체
일반 접근	▸중앙행정기관, 광역지자체	▸1(O)등급 전체, 2(S)등급 신청 후
제한 접근	▸기초지자체, 소규모 공공기관	▸1(O)등급 한정, 2(S)등급 별도 심사
민간 접근	▸공공데이터포털 이용자	▸1(O)등급 중 개방 승인 데이터에 한정

- 데이터 이용 효율을 위해 데이터별로 아래 4가지 활용 유형의 허용 여부를 메타데이터에 명시
  - **열람:** 플랫폼 내 조회 및 시각화
  - **분석:** 다운로드 후 통계·정책 분석 목적 활용
  - **결합:** 타 기관 데이터와 연계·결합
  - **AI 학습:** 기계학습 모델 훈련 데이터로 활용할 경우, 데이터 오너 기관의 명시적 동의 및 행정안전부의 승인을 조건으로 허용
- 2개 이상의 기관이 보유한 데이터를 결합하여 새로운 가치를 창출하고자 할 때는 데이터 유출 방지를 위해 엄격한 승인 절차를 거쳐야 함
  - **결합 신청:** 활용 기관은 결합 목적과 대상 데이터를 명시한 결합 신청서를 제출
  - **원천 기관 동의:** 신청된 데이터의 소유권을 가진 각 원천 기관으로부터 활용 동의을 얻음
  - **샌드박스 처리:** 모든 결합 작업은 외부 유출이 차단된 플랫폼 내 샌드박스에서 수행
  - **반출 재심사:** 결합이 완료된 결과물을 외부로 반출하기 전, 개인 식별 가능성 등을 검토하는 최종 재심사 절차를 거침
- 모든 관계자는 법적·제도적 규제를 준수하며 데이터의 적법성과 이용 권한을 확보하고, 개인정보 보호 및 공정성 등의 윤리적 원칙을 철저히 이행함
  - 기술 실무자와 규제기관은 데이터의 출처와 적법한 이용 권한이 확보되었는지 사전에 철저히 검증함
  - 데이터 처리 전 과정에서 개인정보 보호는 물론, 특정 집단에 대한 차별이나 편향이 발생하지 않도록 비차별·공정성 원칙을 준수함

3.5 데이터 거버넌스와 책임 구조

데이터를 통한 행정의 투명성을 높이고 기관 간 원활한 데이터 공유와 협력을 위해 데이터의 생산부터 활용까지 전 주기에 걸친 책임 소재를 명확히 함

- **데이터 거버넌스 체계:** 공공데이터에 대한 제공 여부는 공공기관의 장이 최종결정권을 가지며, 공공데이터의 관리 체계를 책임 있게 운영함
  - **공공데이터법 제12조(공공데이터제공책임관 등):** 각 공공기관은 공공데이터 업무를 총괄하는 책임관(고위직)과 실무를 담당하는 실무담당자를 지정함
  - **데이터기반행정법 제19조(데이터활용역량 강화):** 정부는 공공기관 종사자의 데이터 분석 및 활용 역량을 높이기 위한 교육과 훈련을 지원함
- **품질 문제 발생 시 책임:** 공공데이터 오류 발생 시 공공기관은 신속한 조치를 통해 데이터의 신뢰성을 유지하도록 노력함
  - **이슈 신고:** 플랫폼 내 품질에 대한 오류나 불만사항 발생 시, 데이터 제공기관은 10일 이내 데이터를 수정하거나, 관련 사항을 소명함
- **활용 결과 환류:** 데이터의 품질을 지속적으로 개선하는 선순환 구조를 위해 데이터 활용 결과를 공유하도록 노력함
  - **실적공유:** 데이터 활용 기관은 분기별로 데이터 활용사례, 결과분석, 품질점검 결과 등을 데이터 플랫폼에 공개할 수 있음
  - **품질개선:** 데이터 관리기관은 제공된 피드백을 참고하여 데이터 품질을 보완하고 갱신 주기를 최적화
  - **사례확산:** 우수 활용사례를 플랫폼에 공개하여 타 기관의 벤치마킹과 데이터 활용 활성화를 유도



3.6 법적 충돌 해소 및 데이터 보호 체계

공공데이터 관련 법령 간의 관계를 정립하고 부처 간 이견을 조정하는 체계적인 프로세스 운영을 통해 공공데이터의 공유와 활용 과정에서 발생할 수 있는 법적 불확실성을 해소

3.6.1 법률 간 충돌 조정 체계

- 「데이터기반행정법」과 개별 법령 간 상충할 경우 명확한 우선순위 원칙을 수립하고, 전문적인 조정 기구를 통해 데이터 공유의 법적 근거를 확보
- 현장에서의 혼선을 방지하기 위해 「데이터기반행정법」과 상충할 가능성이 있는 주요 법률 및 충돌 유형을 사전에 검토함
  - **개인정보 보호법:** 제3자 제공 제한 관련 사항
  - **신용정보의 이용 및 보호에 관한 법률:** 금융정보 보호 관련 사항
  - **통계법:** 통계 목적 수집 데이터의 목적 외 이용 제한 관련 사항
  - **국가정보원법 및 보안업무규정:** 국가 보안 및 비밀 유지 관련 사항
  - **기타:** 보건, 의료 관련 법률 중 개인정보 관련 사항
- 공공기관은 데이터 제공을 위해 적극적으로 노력함
  - 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제36조(면책)에 따라 데이터 품질이나 제공 중단 등으로 발생한 손해에 대해 고의·중과실이 없으면 데이터 제공주체가 책임을 지지 않음
  - 공공데이터 제공 담당자는 면책을 위한 기준과 사례에 대해 「공공데이터 적극 제공을 위한 면책 안내서」(행정안전부, 2026.2)를 참고할 수 있음
- 공공데이터 요청에 대한 제공여부에 이용자와 공공기관간 이해관계가 충돌하는 경우 '공공데이터분쟁조정위원회'에 조정을 요청할 수 있음
  - **「공공데이터의 제공 및 이용 활성화에 관한 법률」 제29조:** 공공기관의 공공데이터 제공거부 및 제공중단에 관한 분쟁조정을 위하여 공공데이터분쟁조정위원회를 설치하고 구성하도록 규정
  - **「데이터기반행정 활성화에 관한 법률」 제13조:** 공공기관 간 데이터 제공 거부에 대한 조정 업무도 공공데이터분쟁조정위원회에서 담당하도록 명시

3.6.2 데이터 보호 및 권리 준수

- 기술 실무자, 데이터 제공 주체, 규제기관은 법·제도 및 규제 준수를 전제로 데이터의 적법성·출처·이용 권한을 확보하고, 개인정보보호와 다양성·비차별·공정성 등 윤리적 원칙을 함께 이행해야 함
- 특정 집단이 과대 표집(Oversampling)되거나 과소 표집(Undersampling)되어 차별적 모델이 나오거나 민감정보가 추론될 수 없도록 데이터 생성 시부터 윤리적인 요소를 고려해야 함
  - 다양성·대표성 부족에 따른 잠재적 편향 및 문제적 사용 사례 발생 가능성에 대한 테스트, 노인·장애인 등 공정성에 영향을 받는 집단에 대한 고려와 협의 등 수행

참고 EU, 신뢰할 수 있는 인공지능 평가 목록 – 개인정보 보호 및 다양성, 차별 금지 및 공정성[12)]

❶ 개인정보보호
- AI 시스템이 사생활 보호 권리, 신체적·정신적·도덕적 무결성에 대한 권리, 데이터 보호 권리에 미치는 영향을 고려했습니까?
- 사용 사례에 따라 AI 시스템과 관련된 개인정보 보호 문제를 신고할 수 있는 메커니즘을 마련했습니까?

❷ 다양성, 차별 금지 및 공정성
- AI 시스템에서 입력 데이터 사용 및 알고리즘 설계 측면 모두에서 불공정한 편향을 생성하거나 강화하지 않도록 방지하기 위한 전략이나 절차 집합을 수립했습니까?
- 데이터에서 최종 이용자 및/또는 대상의 다양성과 대표성을 고려했습니까?
- AI 설계자와 개발자가 AI 시스템 설계 및 개발 과정에서 발생할 수 있는 편향을 더 잘 인식할 수 있도록 교육 및 인식 제고 활동을 마련했습니까?
- AI 시스템의 편향성, 차별성 또는 성능 저하와 관련된 문제를 신고할 수 있는 메커니즘을 마련했습니까?
- AI 시스템 구축 과정의 어느 단계에서든 귀하의 공정성 정의(노인 또는 장애인 대표)가 일반적으로 사용되고 구현되고 있습니까?

- 데이터 축약 및 범주화 처리방식, 부분 삭제 처리, 식별 요소 숨김 처리 등을 통해 데이터에 포함된 개인정보를 비식별화 처리
- 개인정보, 비공개 대상 정보 포함·가공 여부를 파악하여 「개인정보 보호법」, 「공공기관의 정보공개에 관한 법률」(「정보공개법」), 등 관련 법령을 준수하여 처리

※ 공공데이터 제공 관리 실무 매뉴얼(2023) P.14 참조

[표 31] 비식별화 방법

구분	내용
데이터 축약 및 범주화	▶ 상세 주소 → 시/군 단위, 상세 나이 → 연령대로 변경
부분 삭제 처리	▶ 개인정보 관련 정보 중 일부를 부분 삭제하나, 데이터 분석 목적을 해치지 않는 수준에서 변환 ▶ 차량번호 12가3456 → ○○가0001
식별 요소 숨김 처리	▶ 일부 문자를 * 또는 다른 기호로 가려 식별이 불가능하도록 처리 ▶ 전화번호 010-1234-5678 → 010-****-5678

12 EU High-Level Expert Group on AI. (2019). The Assessment List for Trustworthy AI (ALTAI)

참고 개인정보 비식별화 관련 실행 방안 방법론

❶ 개인정보 등 비공개 대상정보의 포함 여부 확인 (「정보공개법」 제9조에 따른 비공개 대상 정보의 포함 여부 확인)
❷ 개인정보 등 비공개 대상 정보의 기술적 분리 가능 여부 확인
- 개인정보 등 비공개 대상 정보가 포함된 경우, 해당 정보를 기술적으로 분리할 수 있는 경우 관련 사항을 별도 기록·관리하고, 부분 제공 대상으로 1차 확인

❸ 개인정보 등 비공개 대상 정보의 가공을 결정한 경우 부분제공대상으로 1차 확정하고, 불가능한 경우에는 제공 제외 대상으로 최종 확정하여 관련 사항을 별도 기록·관리
❹ 개인정보에 해당하는 부분을 가명 정보, 익명 정보화 하여 처리 후 공개
※ 개인정보의 익명처리 여부는 기관 판단사항으로 추후 재식별 위험 등을 충분히 고려해야 함

참고문헌 개인정보 관련 사항은 「개인정보 보호법령」 준수 및 「가명정보 처리 가이드라인(개인정보보호위원회)」, 「공공분야 가명정보 제공 실무 안내서(행정안전부)」

▸ 이용자 및 기타 게시자와 함께 문제/과제 기반 오픈 데이터 이벤트를 공동으로 조직[13)]

- 데이터 공개 목적상 개인 식별이 완전히 불가능하면 안되는 경우, 가명정보 처리 절차에서 가명화한 데이터를 별도 보관, 키 관리 등을 통해 보호조치를 취하고 있는지 지속적으로 점검 및 확인
- 개인정보보호과 관련하여 데이터 수집 당시 적법한 동의를 받았는지에 대한 명시적 동의 획득 기록(서면 또는 전자동의서) 보관
- 데이터 소유권 및 사용 권한, 데이터의 재사용 또는 재배포에 대한 제한사항에 대한 명확한 정보를 제공
  - ▸ 공공데이터의 저작권이 해당 기관에 있지 않고 일부라도 외부인 및 타 기관 등에 있는 경우, 권리의 소재를 파악하고 이용 허락 여부와 범위를 반드시 확인

  ※ 공공데이터 관리지침 제8조 제1항 참조

  - ▸ 「저작권법」 제24조의2 및 「공공저작물 저작권 관리 및 이용 지침」 제11조에 근거하여 공공기관이 저작물이 포함된 공공데이터를 제공하는 경우 공공저작물 자유이용 허락표시 기준(이하 '공공누리')를 반드시 적용

  ※ 공공데이터 관리지침 제8조 제2항, 공공데이터 제공 관리 실무 매뉴얼(2023) P.19 참조

  * "공공저작물"이란 공공기관 등이 그 저작재산권의 전부 또는 일부를 가지고 있는 저작물을 말하므로 (공공저작물 저작권 관리 및 이용지침 제3조), 공공데이터에 저작물이 포함된 경우는 공공저작물에 해당함

13 Data.gov.ie. (n.d.). Ireland Open Data Publication Guidelines. https://data.gov.ie/guidelines

[표 32] 공공저작물 자유이용 허락표시 기준(이하 '공공누리')

이용허락유형	표시마크	이용허락범위
[제0유형] 자유이용		▶출처표시 조건 없음 ▶상업적, 비상업적 이용가능 ▶변형 등 2차적 저작물 작성 가능
[제1유형] 출처표시		▶출처표시 ▶상업적, 비상업적 이용가능 ▶변형 등 2차적 저작물 작성 가능
[제2유형] 출처표시+상업적 이용금지		▶출처표시 ▶비상업적 이용만 가능 ▶변형 등 2차적 저작물 작성 가능
[제3유형] 출처표시+변경금지		▶출처표시 ▶상업적, 비상업적 이용가능 ▶변형 등 2차적 저작물 작성 금지
[제4유형] 출처표시+상업적 이용금지+변경금지		▶출처표시 ▶비상업적 이용만 가능 ▶변형 등 2차적 저작물 작성 금지
[AI유형] 인공지능학습용		▶출처표시 조건 없음 ▶상업적, 비상업적 이용가능 ▶변형 등 2차적 저작물 작성 가능

- 라이선스 표기를 통한 데이터 보호에 대한 고지는 AI 실무자가 완전한 확신을 가질 수 있도록 명확하게 해야 하고, 메타데이터와 데이터 설명 문서 내 포함하여 정보를 제공해야 함
  - ▶라이선스 이름에는 가급적 국제적으로 통용되는 Creative Commons CC 계열, Open Data Commons와 공공누리 라이선스를 표기하여 이용자 혼란 축소
  - ▶라이선스 전문은 데이터 파일과 함께 배포 또는 메타데이터의 'license' 필드에 포함하며, 메타데이터 내 스키마를 활용하여 기계가 읽을 수 있도록 라이선스 정보 제공
  - ▶데이터셋 설명 문서에도 데이터 이용자가 라이선스에 관하여 올바르게 인지하고 사용할 수 있도록 라이선스 관련 내용을 모두 포함하여 제공

참고 데이터셋 라이선스 모니터링

- ▶향후 법적 분쟁 가능성을 낮추기 위하여 데이터셋 관리자는 정기적으로 사용 현황에 대한 점검 필요
- ▶사용 현황 리스트에서 라이선스 조건 위반 사례에 대한 꾸준한 모니터링 필요
- ▶데이터 업데이트(새 데이터 추가 등)가 있을 경우 라이선스 범위 내에 존재하는지 확인, 개정사항이 있을 경우 이를 데이터셋 설명서(이용가이드), 메타데이터, 링크 내 반영하고 변경된 사항에 대하여 해당 데이터를 사용 중인 이용자들이 알 수 있도록 정보를 제공해야 함

부 록 1

인공지능 친화적 공공데이터 관리 원칙

원칙 01 데이터는 개방형 방식으로 제공되어야 하며, 기계 이해가 가능한 형식으로 구조화 되어야 한다.

원칙 02 데이터는 부분 조회가 가능하여야 하며, 실시간 또는 준실시간 접근 및 활용을 지원하여야 한다.

원칙 03 데이터의 구조와 의미를 설명하는 포괄적이고 기계가 이해 가능한 메타데이터를 제공하여야 한다.

원칙 04 데이터 구조, 활용 목적, 편향성 및 한계점, 버전 및 변경 이력을 투명하게 문서화하고 공개하여야 한다.

원칙 05 인공지능 활용에 적합한 공공데이터 품질을 지속적으로 점검하고 개선하여야 한다.

원칙 06 기관 간 데이터 연계성과 상호운용성을 확보하기 위하여 표준 코드를 일관되게 적용하여야 한다.

원칙 07 데이터 품질, 보안, 윤리, 업데이트에 대한 책임 주체와 관리 체계를 명확히 지정하여야 한다.

원칙 08 사용자 중심의 데이터 인프라를 구축·운영하여 데이터에 대한 접근성을 강화하여야 한다.

원칙 09 사용자 평가 및 피드백 체계를 운영하고, 그 결과를 데이터 품질 개선 과정에 반영하여야 한다.

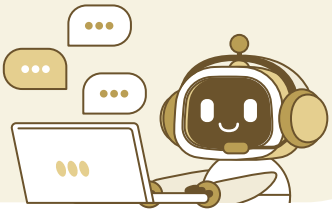
원칙 10 데이터 접근 권한을 차등 관리하고, 데이터에 대한 보안을 강화하여야 한다.

원칙 11 데이터의 생산부터 활용까지 전 주기에 대한 책임 소재를 명확히 하여야 한다.

원칙 12 데이터 생애주기 전반에 윤리적 법적 기준을 준수하고, 데이터 무결성과 신뢰성을 확보하여 공공의 이익을 보호하여야 한다.

부 록 2

주요 용어 정의



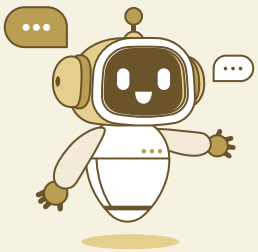
	용어	설명
1	AI (Artificial Intelligence)	▸데이터 기반 학습을 통해 판단·예측·추론 등의 기능을 수행하는 지능형 기술 ▸자동화·지능화된 서비스를 제공하기 위한 알고리즘 기반 시스템
2	인공지능 친화적 (Artificial Intelligence Ready)	▸AI 학습·추론에 적합하도록 품질·구조·메타데이터가 정비된 데이터 상태 ▸지속적 활용을 위해 표준 포맷·계보·문서화 등이 갖추어진 데이터셋 구성 요건
3	API (Application Programming Interface)	▸시스템 간 데이터 교환을 위한 표준화된 호출 방식 ▸응용프로그램이 외부 서비스 기능을 접근·활용하기 위한 인터페이스 규격
4	CSV (Comma-Separated Values)	▸쉼표로 구분된 텍스트 기반의 표 형식 데이터 구조 ▸다양한 환경에서 호환성이 높은 데이터 교환 포맷
5	DCAT (Data Catalog Vocabulary)	▸데이터 카탈로그 상호운용성을 확보하기 위한 W3C RDF 기반 메타데이터 표준 ▸데이터셋·배포·서비스 등의 기술 요소를 구조화한 기술 규격
6	DOI (Digital Object Identifier)	▸디지털 객체에 영구 식별자를 부여하여 위치와 무관한 지속적 참조를 가능하게 하는 체계 ▸학술자료·데이터셋 등의 장기 식별·추적을 위한 고유 식별 시스템
7	DVC (Data Version Control)	▸데이터·모델 파일을 Git과 연동하여 버전별로 관리하기 위한 분산형 버전관리 도구 ▸대규모 데이터 변경 이력의 재현성과 관리 효율성을 위한 관리 체계
8	EDA (Exploratory Data Analysis)	▸데이터의 분포·특성·이상치를 시각화와 통계로 파악하는 사전 분석 절차 ▸모델 구축 전 품질 점검 및 가설 검토를 위한 탐색적 분석 방식
9	GeoJSON (Geographic JavaScript Object Notation)	▸지리공간 정보를 JSON 구조로 표현하는 국제 표준 포맷 ▸좌표·지오메트리·속성 정보를 구조화한 공간 데이터 표현 방식
10	Git (Global Information Tracker)	▸코드·문서·데이터 변경 이력을 분산 방식으로 관리하는 버전관리 시스템 ▸협업·롤백·추적을 위한 변경 이력 관리 도구
11	Hallucination	▸모델이 사실과 다른 정보를 생성하거나 오류를 포함한 결과를 산출하는 현상 ▸데이터 품질·프롬프트 조건 등에 의해 발생하는 그럴듯해 보이지만 잘못된 응답 출력 상황
12	JSON (JavaScript Object Notation)	▸키-값 구조 기반의 경량 데이터 교환 포맷 ▸웹서비스·API 응답에서 널리 사용되는 구조화 데이터 형식
13	LLM (Large Language Model)	▸대규모 텍스트 데이터를 학습하여 자연어 이해·생성을 수행하는 모델 구조 ▸다양한 도메인의 지식을 종합하여 응답을 산출하는 초거대 언어모델 체계
14	LOD (Linked Open Data)	▸URI 기반으로 데이터 간 관계를 연결하여 활용하는 공개형 의미론 데이터 구조 ▸공공데이터의 연계·통합을 위한 링크드 데이터(Linked Data) 구현 방식
15	Parquet	▸컬럼 기반 저장 방식으로 압축 효율과 조회 성능을 높인 데이터 포맷 ▸대규모 분석·AI 학습에서 활용되는 고성능 저장 구조
16	PGP (Pretty Good Privacy)	▸데이터 암호화·전자서명·무결성 보장을 위한 공개키 기반 보안 기술 ▸민감정보 전송 및 공유 시 안전성을 확보하기 위한 보호 체계
17	PID (Persistent Identifier)	▸디지털 객체를 영구적으로 식별하는 고유 식별자 ▸데이터셋·문서 등의 안정적 참조·추적을 위한 식별 체계

	용어	설명
18	RAG (Retrieval-Augmented Generation)	▶ 외부 지식을 검색·결합하여 근거 기반 응답을 생성하는 AI 아키텍처 ▶ 모델 환각 감소와 최신성 확보를 위한 검색-생성 결합 방식
19	RDF (Resource Description Framework)	▶ 데이터 간 관계를 삼중 구조로 표현하는 의미론적 데이터 모델 ▶ DCAT·PROV-O 등 공공데이터 메타데이터 표현을 위한 기반 기술 체계
20	RESFful 방식 (Representational State Transfer)	▶ HTTP 규칙을 준수하여 자원 중심으로 API를 설계하는 방식 ▶ 단순·일관된 인터페이스 제공을 위한 웹 아키텍처 원칙
21	Top-K	▶ 모델이 산출한 후보 중 상위 K개의 결과를 선택하는 랭킹 기반 방식 ▶ 검색·추천·생성 모델에서 우선순위 결과를 도출하기 위한 선택 기준
22	URI (Uniform Resource Identifier)	▶ 웹 자원을 고유하게 식별하기 위한 표준 식별자 ▶ LOD·메타데이터·API 등에서 참조 일관성을 확보하는 주소 체계
23	YAML(YAML Ain't Markup Language)	▶ 계층형 구조를 직관적으로 표현하는 데이터 직렬화 포맷 ▶ 설정 파일 및 데이터 파이프라인 정의에 활용되는 경량 구성 언어
24	가명정보	▶ 추가 정보 없이는 특정 개인을 식별할 수 없도록 처리한 개인정보 ▶ 통계·연구·AI 활용에 적합하도록 익명성 수준을 강화한 정보 형태
25	과대표집	▶ 특정 그룹이 실제 비중보다 과도하게 포함된 데이터 구성 ▶ 분석·모델링 과정에서 편향을 유발할 수 있는 표본 불균형 상태
26	과소대표집	▶ 특정 그룹이 실제 비중보다 부족하게 포함된 데이터 구성 ▶ 성능 저하 및 편향 발생 가능성이 높은 표본 불균형 상태
27	데이터셋(Dataset)	▶ 특정 목적이나 분석을 위해 수집·정제된 데이터의 집합체 ▶ 구조화된 형태로 저장·관리되며 메타데이터와 함께 공유·재사용 가능한 정보 단위
28	데이터 인벤토리	▶ 기관 보유 데이터의 유형·구조·형식·위치를 종합 파악하는 목록 ▶ 데이터 관리·현황 파악·발견성을 높이기 위한 자산 관리 체계
29	데이터 재사용	▶ 기존 데이터를 새로운 분석·정책·모델에 반복 활용하는 과정 ▶ 비용 절감과 가치 창출을 위한 데이터 순환 이용 방식
30	데이터 중복 (Leakage)	▶ 학습·검증·테스트 간 동일 데이터가 포함되어 성능이 과대추정되는 현상 ▶ 모델 신뢰성을 저해하는 검증 데이터 누출 상황
31	데이터 집합	▶ 공통된 특성이나 주제를 가진 개별 데이터 항목들의 모임 ▶ 통계 분석·머신러닝·의사결정 등 다양한 목적에 활용되는 기초 정보 묶음
32	데이터 카드 (Dataset Card)	▶ 데이터셋의 목적·구성·품질·한계를 요약하여 제공하는 표준 문서 ▶ 이용자가 주요 특성을 신속히 이해하도록 지원하는 정보 요약 도구
33	데이터 카탈로그	▶ 기관 내·외부 데이터 자산을 메타데이터 기반으로 검색·발견·공유하는 체계 ▶ DCAT 등 표준을 활용하여 상호운용성과 접근성을 높이는 관리 구조
34	데이터 파이프라인	▶ 데이터 수집·저장·처리·배포 등의 과정을 자동화한 처리 경로 ▶ AI 학습·분석 환경의 안정성을 확보하기 위한 데이터 처리 구조
35	라이선스 (License)	▶ 데이터의 사용 조건·저작권·배포 범위를 명시한 규정 ▶ 이용자의 권리·의무를 정의하는 법적 준거 문서
36	멀티모달 (Multimodal)	▶ 텍스트·이미지·음성 등 복수 형태의 데이터를 통합 처리하는 방식 ▶ 복합 정보 입력 기반의 고도화된 AI 분석·생성을 위한 모델 구조
37	메타데이터 (Metadata)	▶ 데이터의 구조·내용·맥락·품질을 설명하는 부가 정보 ▶ 검색·연계·재사용을 지원하는 필수 관리 요소

	용어	설명
38	문서화 (Documentation)	▶ 데이터 생성·처리·활용 과정 및 근거를 체계적으로 기록한 문서 ▶ 재현성·검증·관리 강화를 위한 정보 기록 체계
39	민감정보	▶ 개인 신상·건강·재산 등 공개 시 위험이 큰 정보 ▶ 보호법령에 따라 강화된 안전조치가 요구되는 정보 범주
40	백업 (Backup)	▶ 데이터 손실에 대비해 복제본을 별도로 저장하는 보호 조치 ▶ 장애·재해 발생 시 복구를 위한 필수 데이터 보존 절차
41	비식별화 (De-identification)	▶ 개인 식별요소를 삭제·대체하여 재식별 가능성을 최소화하는 처리 방식 ▶ 개인정보 보호 및 공개 데이터 활용을 위한 안전 조치 기법
42	스키마 (Schema)	▶ 데이터의 구조·필드·제약조건 등을 정의한 설계 규칙 ▶ 데이터셋의 일관성과 검증 가능성을 보장하는 구조 정의 체계
43	오픈 포맷 (Open Format)	▶ 누구나 제한 없이 사용할 수 있는 공개 파일 형식 ▶ 상호운용성·접근성·지속가능성을 확보하는 데이터 제공 방식
44	원시 데이터(Raw Data)	▶ 정제·가공·전처리되지 않은 상태의 데이터 ▶ AI 학습 전 단계에서 정제 필요
45	원천 데이터 (Raw Data)	▶ 가공·정제·전처리 이전의 최초 수집 데이터 ▶ 분석·AI 학습의 기초가 되는 원본 자료
46	응답 SLA(Service Level Agreement)	▶ 시스템 응답 시간·가용성 등 서비스 품질 기준을 정의한 합의 규격 ▶ 서비스 제공자와 이용자 간 책임 범위를 명시한 운영 기준
47	인프라 (Infrastructure)	▶ 데이터 처리·저장·전송을 위한 네트워크·서버·플랫폼 기반 환경 ▶ 안정적 서비스 운영을 위한 기술적 기반 구조
48	임베딩 유사도 (Embedding Similarity)	▶ 텍스트·이미지 등을 벡터 공간에 표현한 후 거리·유사성을 계산하는 방식 ▶ 검색·추천·RAG 등에서 의미 기반 비교를 수행하기 위한 지표
49	전처리 (Preprocessing)	▶ 분석·학습 전 데이터를 정제·정규화·변환하는 준비 과정 ▶ 품질 향상 및 모델 성능 제고를 위한 사전 처리 절차
50	토큰화 (Tokenization)	▶ 텍스트를 단어·문장·하위 단위로 분할하는 처리 방식 ▶ 자연어 처리 모델 입력을 위한 기본 변환 과정
51	파싱 (Parsing)	▶ 데이터 구조를 분석하여 의미 단위로 해석·변환하는 처리 과정 ▶ 문법·규칙에 따라 데이터를 기계가 읽기 적합한 구조로 변환하는 절차
52	편향 (Bias)	▶ 데이터 구성 또는 모델 학습 과정에서 특정 집단·값이 과도하게 영향받는 현상 ▶ 공정성·신뢰성 확보를 위해 관리가 필요한 왜곡 요소
53	프롬프트 (Prompt)	▶ AI 모델에 작업 지시 또는 맥락을 전달하기 위한 입력 문구 ▶ 모델 출력 품질을 결정하는 핵심 제어 요소

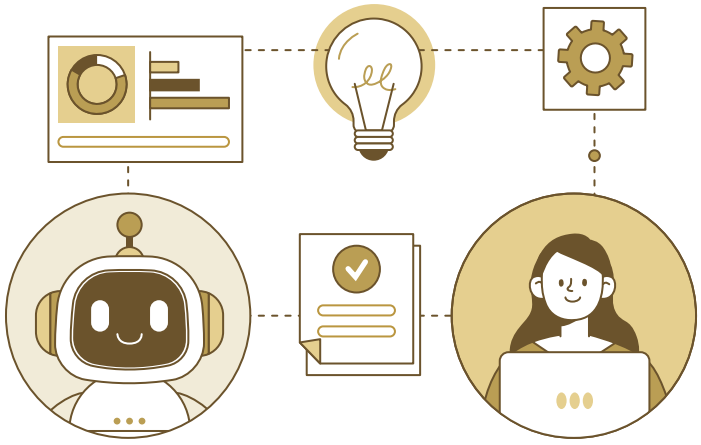
부 록 3

메타데이터 예시



구분	항목	예시
데이터 관리 메타데이터	데이터명	“dct:title”: “2025 서울시 연령별 인구통계 멀티모달 데이터셋”
	제공범위	“dct:description”: “2025년 기준 서울특별시 거주 인구를 연령대별·구별로 집계한 수치 데이터와 함께, 인구 분포 시각화 이미지, 정책 설명 음성 자료, 인구 구조 분석 영상 콘텐츠를 통합 제공하는 멀티모달 통계 데이터셋”
	관련법령	“dct:references”: “통계법 제3조, 국가통계작성지침, 공공데이터 제공 및 이용 활성화에 관한 법률”
	소관기관	“dct:creator”: “서울특별시 빅데이터담당관”
	제공기관	“dct:publisher”: “서울특별시”
	표준데이터셋 제공시스템	“dcat:landingPage”: “https://data.seoul.go.kr/dataset/2025-population-multimodal”
	접속 URL	“dcat:accessURL”: “https://data.seoul.go.kr/datafile/POP_2025_MULTIMODAL”
	갱신주기	“dct:accrualPeriodicity”: “http://purl.org/cld/freq/annual”
	고유식별자	“dct:identifier”: “seoul-population-multimodal-2025”
	담당자명	“dcat:contactPoint”: “fn”: “서울특별시 빅데이터담당관 인구통계 담당자”, “hasEmail”: “mailto:data@seoul.go.kr”
	업무/주제분류체계	“dcat:theme”: “인구·사회”
	키워드	“dcat:keyword”: “인구”, “연령별”, “서울시”, “2025”, “인구통계”, “demographics”, “multimodal”
	언어	“dct:language”: “http://id.loc.gov/vocabulary/iso639-1/ko”
	데이터 용량	“dcat:byteSize”: “1850000000”
	데이터셋 유형	“dcat:mediaType”: “text/csv”
	관련 데이터셋	“dct:relation”: “2024 서울시 연령별 인구통계 정보”, “서울시 구별 인구 동향 데이터”
	시간범위	“dct:temporal”: “time:hasBeginning”: “2025-01-01”, “time:hasEnd”: “2025-12-31”
	공간범위	“dct:spatial”: “서울특별시”
데이터 계보 메타데이터	버전	“owl:versionInfo”: “v1.2”
	배포일시	“dct:issued”: “2025-01-15T09:00:00+09:00”
	최종 업데이트/수정일시	“dct:modified”: “2025-12-15T14:30:00+09:00”
	버전 노트	“adms:versionNotes”: “일부 구 단위 인구 누락값 보정, 연령 구간 세분화 (5세 단위), 시각화 이미지 및 정책 설명 영상 신규 추가.”
데이터 이용 메타데이터	라이선스	“dct:license”: “https://...../terms/daegu-dataset-usage-policy”
	권한	“dct:rights”: “본 데이터셋은 서울특별시가 저작권을 보유하며, 행정·정책 수립 및 비상업적 연구 목적에 한하여 이용 가능합니다.”

구분	항목	예시
데이터 품질 메타데이터	알려진 편향 명세	“rai:dataBiases”: “외국인 단기 체류자는 통계에 포함되지 않아 실제 생활 인구와 차이가 있을 수 있음.”, “고령층의 주소 미변경으로 일부 자치구 인구가 과소 추정될 가능성 있음.”
	데이터셋 품질검증 여부	“dqv:hasQualityAnnotation”: “@type”: “dqv:QualityAnnotation”, “dqv:hasQualityMeasurement”: “법무부 연례 검증 절차를 통해 원시 귀화 허가 데이터의 정확성·일관성 점검 완료(2025년 검증 기준).”
	데이터 한계	“rai:knownLimitations”: “주민등록 기준 통계로 실제 상주 인구를 완전히 반영하지 못함.”, “연령대별 이동 인구 분석에는 부적합함.”
	결측치 정보	“rai:dataCollectionMissingData”: “센서 점검(2.1%), 통신 오류(1.1%)로 결측이 발생할 수 있음.”



부 록 4

인공지능 친화적 공공데이터 관리 체크리스트

- 본 체크리스트는 데이터의 단순 개방을 넘어, AI 모델이 즉각적으로 학습(Training) 및 실시간 추론에 활용할 수 있는 '인공지능 친화적' 상태를 점검하기 위한 것임
- **공통** : 모든 공공데이터가 준수해야 하는 항목
- **유형별** : 데이터의 형태나 도메인 특성에 따라 추가적으로 요구되는 항목
- **필수** : 공공데이터 제공 시 필수로 준수해야 하는 항목
- **권장** : 공공데이터 제공 시 권장(선택) 항목
- **□ 적정 □ 미흡 □ 해당없음** : 세 개의 체크 사항으로 표기하여 관리

공통 공공데이터셋 점검 항목

분류	구분	점검 내용	체크
형식 및 구조	필수	▶ 데이터를 특정 소프트웨어에 종속되지 않는 개방형 표준 포맷으로 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음
	필수	▶ 데이터를 기계 이해 및 자동 처리가 가능하도록 구조화된 형태로 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음
	필수	▶ 데이터 구조(스키마)는 필드(키)·자료형·관계·중첩 규칙등을 포함하여 명확히 정의되어 있는가?	□ 적정 □ 미흡 □ 해당없음
	필수	▶ 메타데이터를 표준 스키마(예: DCAT)또는 기관·국가 적용 프로파일에 따라 기계 이해 가능한 형태로 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음
	필수	▶ 메타데이터 작성 기준에서 정의한 필수 항목을 누락 없이 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음
	권장	▶ 데이터 검증 기준(예: 허용값, 필수값, 형식, 관계 제약 등)을 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음
	권장	▶ 대용량 데이터는 압축 제공 또는 대용량 처리에 적합한 포맷(예: Parquet 등)으로 제공하여 이용자가 용이하게 다운로드·활용할 수 있는가?	□ 적정 □ 미흡 □ 해당없음
의미 및 표준	필수	▶ 변수(항목)별 정의·단위·산정기준·코드값 의미를 포함하는 데이터 사전(또는 이에 준하는 문서)을 제공하고 있는가?	□ 적정 □ 미흡 □ 해당없음

분류	구분	점검 내용	체크
의미 및 표준	필수	▶ 날짜·시간·코드 등 주요 값의 표기 형식은 표준 포맷에 따라 일관되게 적용되고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	필수	▶ 동일한 개념은 동일한 라벨·용어로 일관되게 사용하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 적용한 표준 포맷 및 적용 기준(예: 날짜·시간 형식, 코드 체계, 누락 값 표기 등)을 문서화하여 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 도메인 표준 용어체계 또는 온톨로지가 존재하는 경우, 해당 표준과의 정합성 점검 기준(매핑·검증 절차 포함)을 마련·적용하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
품질 및 이력	필수	▶ 결측값 및 특수값 표기 규칙을 통일하여 일관되게 적용하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	필수	▶ 데이터 품질평가의 수행 여부 및 결과 내용을 메타데이터에 명시하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	필수	▶ 인공적으로 생성된 합성 데이터나 기계가 자동으로 단 주석이 포함된 경우 이를 명확히 구분하여 표기했는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 결측값 및 특수값의 정의, 발생 사유, 처리 원칙(예: 제외, 대체, 별도 코드 처리 등) 등의 정보를 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터의 원천·수집경로·가공/전처리(정제·비식별) 이력을 기록·관리하고 공개 가능한 범위에서 정보를 제공하는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터 정제 및 변환에 사용된 소스 코드는 라이선스를 명시하여, 공개 가능한 범위 내 오픈소스 형태로 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터, 메타데이터 및 핵심 문서(스키마/데이터사전·코드북/접근 안내)는 이용자가 동일한 배포 경로에서 함께 확인할 수 있도록 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터 또는 문서가 변경되는 경우, 관련 메타데이터 및 핵심 문서를 동시에 개정·현행화하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터 특성(용도·품질·리스크 등)을 요약한 데이터 카드(또는 이에 준하는 설명 문서)를 함께 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
관리 및 보안	필수	▶ 개인정보(또는 민감정보)가 포함된 경우, 관련 법령·지침에 따라 비식별/가명처리 등 보호조치를 적용하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	필수	▶ 이용허락 범위(재사용, 재배포, 상업적 이용, 2차 저작물 작성 등)를 라이선스로 명확히 표시하고, 해당 라이선스 정보를 메타데이터로 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음

분류	구분	점검 내용	체크
관리 및 보안	권장	▶ 재식별 위험에 대한 평가·검토를 수행하고, 그 결과 및 조치 사항을 문서화하여 관리하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터에 영구 식별자(PID 또는 이에 준하는 지속 식별자)를 부여하여 관리하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
데이터 제공	권장	▶ 데이터는 이용자의 자동 수집 및 연계를 지원할 수 있도록 웹 기반 표준 API(예: RESTful API) 형태로 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터가 API로 제공되는 경우, 이용 편의 및 장애 대응을 위해 파일 다운로드 등 최소 1개 이상의 추가 제공 경로를 함께 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ API 제공 시 데이터 응답에는 필요한 메타데이터를 함께 포함하거나, 메타데이터를 확인할 수 있는 표준화된 참조 링크를 함께 제공하도록 설계되어 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 대용량 데이터 처리 및 자동 수집을 위해 필터링, 조건 검색, 분할 조회(페이지네이션) 등 효율적인 조회 기능을 제공하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
관리 유지	필수	▶ 품질·보안·윤리·업데이트에 대한 책임 주체(역할 기준)를 명확히 하고 이용자 문의/오류 신고 연락 창구를 명시하고 운영하는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	필수	▶ 이용자 오류 신고 및 개선 요청을 접수·처리할 수 있는 절차를 마련하고, 피드백 체계를 운영하고 있는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터 변경 사항은 버전별로 추적 가능하도록 변경 이력 정보를 제공하는가?	☐ 적정 ☐ 미흡 ☐ 해당없음
	권장	▶ 데이터의 변조 방지 및 신뢰성 확인을 위한 전자서명이나 해시(Hash)값 제공 등 기술적 검증 방식 운영하는가?	☐ 적정 ☐ 미흡 ☐ 해당없음

유형별 데이터 특성에 따른 추가 점검 항목

데이터 유형	구분	점검 내용
수치	권장	▶ 산식/집계 단위/기준시점 등 산출 규칙을 제공하는가?
이미지 / 영상	필수	▶ 수집 환경(장비/센서·조건 등)메타정보를 제공하는가?
	필수	▶ 라벨링 가이드라인·검증 절차 정보를 제공하는가?
개인정보 보호	필수	▶ 얼굴, 차량번호, 목소리 등 민감 정보의 마스킹/익명화 처리를 완료하고, 재식별 위험성 평가를 거쳤는가?
	권장	▶ 비식별·가공으로 인해 제공 정보의 공백/왜곡 가능성이 발생하는 항목에 대해, 가공 범위·원칙·주의사항을 이용자에게 명확히 안내하는가?

부 록 5

인공지능 친화적 공공데이터 데이터 카드 상세

분류	항목	작성 내용
데이터셋 개요	데이터셋명/ID	▶ (공식 명칭) / (고유 식별자)
	발행기관/담당부서/연락처	▶ (기관명, 부서명, 이메일/전화)
	데이터셋 요약	▶ 데이터 내용, 어떤 행정·정책 목적에 활용되는지 3~5문장으로 기술
	지원 태스크(활용 목적)	▶ 예: 분류 / 예측 / 추천 / 검색 / 요약 / 시계열 예측 / 이상탐지 등(해당 시)
	언어 정보	▶ ko / en / 다국어 여부, 텍스트 데이터의 언어 비율(가능 시)
	대상 범위	▶ 기간(시작~종료), 공간(행정구역 단위), 대상(기관/업무/서비스)
	갱신 정책	▶ 갱신주기(일/주/월), 갱신 기준일(전일/실시간), 갱신 지연 허용 범위
	제공 형태	▶ 파일 다운로드 / API 제공 여부(엔드포인트 유무)
데이터셋 구조	데이터 인스턴스(샘플)	▶ 1~3개 샘플 레코드를 예시로 제시(비식별/마스킹 적용)
	데이터 필드 정의	▶ 필드 정의 표로 제공
	키/식별자 규칙	▶ 기본키/복합키, 중복 허용 여부, 참조키 정의
	결측·특수값 규칙	▶ 결측 표기 방법(NULL/빈값/NA), 0과 결측 구분, 마스킹 값 규칙
	분할 정보(Splits)	▶ 학습/검증/평가 분할 제공 여부(해당 시), 분할 기준(시간순/무작위)
	포맷·인코딩·표준	▶ 데이터포맷, UTF-8, 날짜·시간 ISO 8601, 표준 코드체계(행정표준코드 등)
	기계 이해 스키마 제공	▶ 스키마 파일 또는 메타데이터 파일 제공 위치
데이터셋 생성	구축 배경	▶ 정책/행정 목적, 구축 필요성, 제공 범위 결정 근거
	데이터 소스(출처)	▶ 원천 시스템 / 기관, 수집 방식, 수집 기간
	정제·전처리 과정	▶ 중복 제거, 오류 정정, 표준화(코드/단위/형식), 이상치 처리 원칙
	어노테이션 과정(해당 시)	▶ 라벨 정의, 작업자(내부/외부), 가이드라인, 품질검수 등
	비식별·보호 조치	▶ 개인정보 포함 여부, 비식별 처리 방식, 재식별 위험 완화(소량 셀 마스킹 등)
	품질관리 절차	▶ 생성 단계별 검증 항목(스키마 검증/코드값 검증/샘플링 대조 등)과 책임 주체
사용 시 고려사항	권장 사용 범위	▶ 가능한 사용 목적(정책분석/서비스 개선/AI 학습·평가 등) 명시
	비권장/금지 사용	▶ 감시/차별/재식별 시도 등 오용 방지 목적의 제한 명시
	사회적 영향	▶ 데이터 활용이 시민 / 취약계층 / 정책결정에 미칠 수 있는 영향 요약
	편향성 및 한계	▶ 대표성 한계(지역/기간/집단/업무유형 편중), 측정오차, 누락 가능성
	기술적 제약	▶ 해상도(시간/공간), 지연시간, 표본 크기, 변경 이력 영향(버전 간 비교 주의)
	리스크 완화 권고	▶ 모델링 시 주의사항(시간순 분할, 편향 점검, 품질 플래그 활용 등)
추가 정보 및 라이선스	라이선스	▶ 적용 라이선스(공공누리 등), 허용 / 제한 조건(출처표시/변경금지 등)
	인용	▶ 데이터셋 인용 문구(기관명, 데이터명, 버전, 발행일, 접속일)
	작성자/관리자 정보	▶ 작성 / 검토 / 승인 책임부서, 문의 채널
	버전/변경 이력	▶ 버전 규칙, 주요 변경사항, 하위호환성 안내
	관련 문서	▶ 데이터사전, 코드북, 품질보고서, 활용사례, API 문서 링크/문서번호
기술적 사양	재현성	▶ 동일 조건에서 동일 결과를 얻기 위한 전처리·생성 파이프라인 설명
	환경 설정	▶ 전처리 / 검증 / 평가 실행 환경(OS, 컨테이너/가상환경 사용 여부)
	매개변수/규칙	▶ 집계 기준, 필터링 기준, 마스킹 임계값, 표준화 규칙 등 핵심 파라미터
	코드 및 라이브러리 버전	▶ 전처리 / 검증 / 평가 코드 제공 위치, 의존성 목록 및 버전
	무결성 검증 정보	▶ 체크섬 제공 여부, 파일 크기 / 레코드 수 검증 기준
	접근 사양(해당 시)	▶ API 버전, 인증 방식, 호출 제한, 응답 스키마, 오류코드 정의

- 본 공공데이터의 인공지능 친화적 관리 가이드라인 내용의 무단전재 및 재배포를 금하며, 가공 인용시에는 반드시 행정안전부, 한국지능정보사회진흥원의 출처를 밝혀주시기 바랍니다.
- 본 공공데이터의 인공지능 친화적 관리 가이드라인 내용은 행정안전부와 한국지능정보사회진흥원의 공식 견해와 다를 수 있고, 문의 및 제안은 아래 연락처로 문의해주시기 바랍니다.
- **문의 :** 한국지능정보사회진흥원 공공지능데이터본부 공공지능데이터인프라팀 류 도 수석 (doryu@nia.or.kr)

공공데이터의 인공지능 친화적 관리 가이드라인

2026. 03

행정안전부 NIA 한국지능정보사회진흥원