데이터 구축 공정 순환 기반 메타데이터: 품질·운영관리 적용 방안

박정하 TTA AI반도체검증팀 책임연구원
박경은 TTA AI데이터품질팀 선임연구원

1. 머리말

AI 학습용 데이터는 거의 모든 AI 기반 어플리케이션의 바탕을 이루는 동시에, 다양한 산업 분야에 AI 기술을 접목·도입해 혁신을 일으키는 핵심 동력이라 할 수 있다. AI가 초거대 AI로 발전하는 과정에서 데이터는 단순한 학습 재료 이상의 가치를 가지며, 컴퓨터 비전, 자연어 처리, 추천 시스템, 산업 자동화 분야 등 그 활용도가 다양해지고 있다. 이러한 다양한 분야의 AI는 데이터에 의해 그 성능이 좌우되며, 고도화된 AI 모델을 서비스화해 운영하는 과정에서 새로운 데이터가 다시 생성·수집되면서 지속적인 순환 과정을 이룬다. 또한 데이터를 배포한 이후에도 재사용, 재가공, 재학습하는 방식으로 생태계 전체를 순환한다. 이 순환 과정에서 누가, 어떤 목적과 의도를 가지고, 어떤 기준과 절차로 만들었는지에 대한 설명은 곧 AI 신뢰성의 최소 조건이 된다. 이와 같이 데이터 구축 기준, 절차, 변경 이력을 구조화해 남긴 메타데이터는 구축 공정의 기록으로 기능하게 된다.

이번 원고에선 데이터 구축 계획 수립, 데이터 획득 및 수집, 데이터 정제 과정, 데이터 가공 과정, 데이터 학습 과정, 데이터 활용 과정에 이르는 데이터 구축 공정에서 요구되는 메타데이터의 주요 항목과 구성요소를 살펴본다. 더불어 품질관리 및 운영관리 관점에서의 시사점과 적용 방안을 알아보고자 한다.

2. 공정 기반 메타데이터의 핵심

2.1 AI 학습용 데이터 구축 공정 중심 메타데이터의 필요성

데이터 수집 및 구축부터 AI 모델 훈련, 새로운 데이터 생성 및 활용에 이르는 데이터 순환 과정은 메타데이터에 의해 효율적으로 관리·최적화된다. 메타데이터는 검색 및 구성을 용이하게 하고, 이를 활용해 데이터 오류나 불일치를 식별·수정함으로써 전반적인 데이터 품질을 향상시킬 수 있다. 특히 AI 학습용 데이터는 AI를 위한 목적과 의도가 명확하므로, 그러한 내용을 절차, 산출물 형식의 메타데이터로 남겨 이용자 관점의 투명성을 확보할 수 있어야 한다. AI 학습용 데이터의 순환 과정은 '구축 계획 수립→획득·수집→정제→가공→학습 후 AI 활용에 따른 데이터 재생성 및 수집' 단계를 다시 거친다. 그러므로 구축계획 수립부터 학습까지 데이터 구축 공정에 초점을 맞춘 메타데이터의 필수 세부 정보를 구체화하고, 구조를 체계화할 필요가 있다.

2.2 구축 공정 기반 메타데이터의 핵심 기능

구축 공정 기반 메타데이터의 핵심 기능은 <표 1>과 같다. 우선, 데이터가 어떤 공정(수집·가공·검수 등)을 거쳐 생성됐는지 체계적으로 기록함으로써 데이터 출처와 변경 이력을 투명하게 관리하고, 데이터 이력 및 추적성을 확보한다. 또한 공정 단계별 검수 기준과 검수 결과(데이터 등)를 함께 관리해 최종 산출물이 요구되는 품질 수준을 충족하는지 확인할 수 있는 객관적 근거를 제공하고, 이를 통해 품질 관리와 신뢰도 향상에 기여한다. 더불어 표준화된 공정 정보와 구축 환경·처리 절차 기록을 통해 유사 프로젝트 수행 시 시행착오를 줄이고, 담당자 변경이나 시스템 업그레이드 상황에서도 신속한 대응을 가능토록 해 유지보수성과 재사용성을 높인다. 데이터 자체에 대한 메타데이터 관점에선 데이터 형식, 스키마, 속성 정보를 정의·구조화해 사용자가 필요한 데이터의 구조와 특성을 쉽게 파악할 수 있도록 지원한다. 마지막으로 AI 학습 관점에선 학습 데이터 생성·검수 이력과 관련 정보를 근거로 모델 성능과 편향성을 분석하고 개선·보완하는 데 필요한 기초 자료를 제공한다.

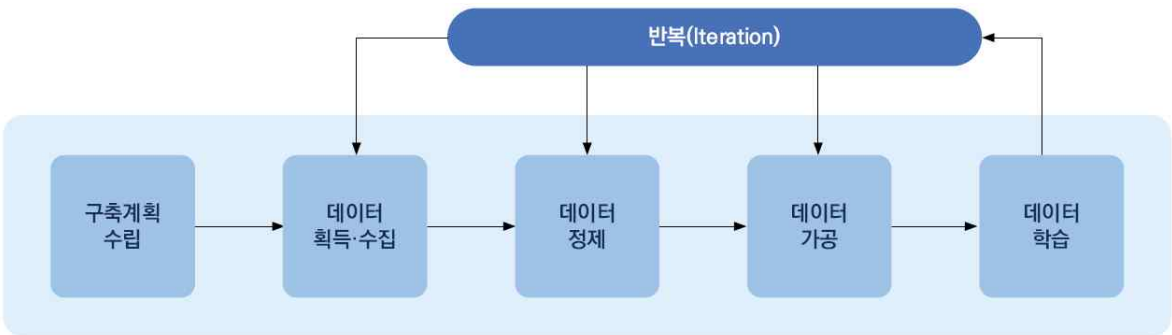
<표 1> 구축 공정 기반 메타데이터 핵심 기능

핵심 기능	설명
데이터 추적성 확보	특정 데이터가 어떤 원천에서 어떠한 가공 과정을 거쳐 생성됐는지 추적 가능
품질관리 및 신뢰도 향상	공정별 검수 이력을 관리함으로써 데이터의 정확성과 완전성을 보증
유지보수 및 재사용성 증대	구축 환경과 로직이 기록돼 있어, 담당자 변경이나 시스템 업그레이드 시에도 신속한 대응 가능
데이터의 구조적 이해	데이터 형식, 스키마, 속성 정보를 정의해 사용자가 필요한 데이터의 구조를 쉽게 파악
AI 모델 학습의 근거 제공	AI 모델의 성능 편향성을 분석·개선하는 데 필수적인 기초 자료로 사용 가능

3. 구축 공정 기반 메타데이터 항목체계와 구성요소

3.1 구축 공정 기반 메타데이터 항목체계

메타데이터 항목체계는 AI 학습용 데이터 구축 공정([그림 1])을 '계획 수립→획득·수집→정제→가공→학습' 5단계 공정으로 보고, 각 단계마다 이뤄지는 공정작업의 목적·의도·절차와 그에 따른 생성 정보·산출물에 기반해 9가지 메타데이터 항목(<표 2>)으로 구분한다. 그리고 데이터 셋 배포와 활용에 필요한 세부 내용을 각 항목별 구성요소(<표 3>)로 정의한다[1].



[그림 1] AI 학습용 데이터 구축 공정

<표 2> 구축 공정 기반 메타데이터 항목

연번	항목명(국문)	항목명(영문)	설명
1	데이터 개요 정보	Data Overview	구축하고자 하는 AI 학습용 데이터에 대한 전반적인 정보를 의미하며, 데이터 분야, 유형, 구축 목적 등을 작성한다.
2	유형별 메타데이터 정보	Metadata by type	데이터 유형별 메타데이터 정보로, 데이터 유형은 음성, 텍스트, 이미지, 3D, 시계열 유형이다.
3	데이터 통계 정보	Data statistics	데이터 구축 규모 및 수량 정보와 카테고리(주제, 매체 등)별 분포 정보를 의미한다.
4	데이터 저장구조 및 포맷 정보	Data storage structure and format	데이터를 저장·관리하는 디렉토리 폴더 구조에 대한 요약 정보, 파일 명명규칙 및 포맷 정보를 의미한다.
5	데이터 구축 절차 정보	Data construction procedures	데이터 구축계획 수립, 데이터 획득 및 수집, 데이터 정제, 데이터 가공, 데이터 학습에 이르는 과정을 의미한다.
6	AI 모델 유효성 검증 정보	AI model validation	AI 모델을 통해 구축된 데이터의 유효성을 검증하기 위한 정보를 의미한다.
7	데이터 활용 정보	Data utilization	데이터 운영 및 활용 방안과 데이터 하자 및 유지보수에 관한 정보를 의미한다.
8	데이터 사용 시 고려사항	Dataset Construction Considerations	데이터 구축공정 기준에 포함할 수 없으나 구축 시 주의해야 하는 사항 및 해결방법 기술
9	기타	Others	데이터 변경 이력, 데이터 사용 제약 사항 등에 대한 정보를 의미한다.

특히 메타데이터를 '데이터 개요'에만 묶지 않고, 유형별 정보(음성·텍스트·이미지·3D·시계열), 통계 분포, 저장구조·포맷, 구축 절차, AI 모델 유효성 검증, 활용 정보, 사용 시 고려 사항, 버전·변경 이력까지 9개 항목으로 그 구성을 확장했다. 이는 데이터 구축 공정에 따른 내용과 함께 데이터 사용 시 고려 사항 등과 같은 기타 내용을 보완해 AI 투명성 확보를 위한 정보를 명확하게 제공한다.

3.2 구축 공정 기반 메타데이터 항목별 구성요소

학습용 데이터 구축은 '구축계획 수립–획득·수집–정제–가공–학습' 단계로 구분되고, 각 단계별 산출물과 의사결정 근거는 메타데이터로 체계화된다. 이후 데이터 개요, 구축 절차, 구성·통계 정보, 활용 정보, 유효성 검증, 사용 시 고려사항, 버전·이력 등의 범주가 세워지고, 이는 세부 구성요소로 세분화된다. 각 세부 구성요소는 약 40개 수준의 구성요소로 제시되고, 각 구성요소에 대해 ID, 요소명, 정의·설명, 필수 여부 같은 속성체계가 만들어지며, 각 구성요소의 필수 여부에 따라 필수(M)와 선택(O), 권고(R)로 구분된다. 또한 데이터 셋이 확장될수록 검색과 비교 및 책임 소재가 분명하게 나타낼 수 있도록 했다.

구체적으로는 메타데이터 항목별 세부 구성요소를 통해 데이터 셋이 어떤 맥락에서 만들어졌고(데이터 개요 정보), 어떤 내용과 구조로 이뤄진 데이터 셋이며(유형별 메타데이터 정보, 데이터

1) AI 데이터 품질관리 가이드라인 v3.5, https://www.aihub.or.kr/aihubnews/qlityguidance/view.do?currMenu=135&topMenu=103&nttSn=10404

<표 3> 메타데이터 항목별 세부 구성요소

메타데이터 항목	ID	메타데이터 구성요소	설명	필수 여부
데이터 개요 정보	1-1	데이터 셋 명	데이터 셋의 국문명 및 영문명	M
	1-2	데이터 구분	데이터의 특성	M
	1-3	데이터 분야	데이터가 속하는 산업 또는 서비스 분야	M
	1-4	데이터 구축 목적	데이터의 구축 목적	M
유형별 메타데이터 정보	2-1	데이터 영역	데이터가 활용되는 기술적 분야	M
	2-2	데이터 유형	데이터가 구축된 형태	M
	2-3	데이터 파일 형식	데이터의 파일 형식	M
	2-4	데이터 출처	데이터의 수집 원천	M
	2-5	라벨링 유형	라벨링 수행 방법	M
	2-6	라벨링 파일 형식	라벨 정보가 저장된 파일 형식	M
데이터 통계 정보	3-1	데이터 구축 규모	원시·원천·가공 데이터 구축 수량	M
	3-2	데이터 분포	원천·가공 데이터의 대·중·소 카테고리 분류 현황	M
	3-3	공정별 데이터 구성	획득·수집·정제·가공 공정별 데이터 형태·포맷	M
	3-4	기타 데이터 구성	획득·수집·정제·가공 공정 외 기타 데이터 형태·포맷	O
	3-5	원천가공 데이터 파일 관계	원천·가공 데이터의 파일 매칭 관계	M
	3-6	어노테이션 포맷	가공 데이터의 상세 포맷	M
	3-7	단위	가공 데이터의 수량 단위	M
데이터 저장구조 및 포맷 정보	4-1	폴더(디렉토리) 구조	데이터 셋 저장·관리 폴더 구조 요약	M
	4-2	파일 명명 규칙	데이터 셋 저장·관리를 위한 파일 명명 규칙	M
데이터 구축 절차 정보	5-1	획득·수집 절차	획득·수집 공정 상세 절차	M
	5-2	정제 절차	정제 공정 상세 절차	M
	5-3	가공 절차	가공 공정 상세 절차	M
	5-4	획득·수집 기준	획득·수집 공정 구축 기준 및 제외 기준	O
	5-5	정제 기준	정제 공정 구축 기준 및 제외 기준	O
	5-6	가공 기준	가공 공정 구축 기준 및 제외 기준	O
AI 모델 유효성 검증 정보	6-1	모델 설치 환경	모델 구동을 위한 환경 설치 방법	M
	6-2	모델의 설명 및 기능	모델 설명 및 학습 임무	M
	6-3	모델에 사용된 데이터 및 통계	모델에 사용된 학습·검증·평가 데이터별 비율	M
	6-4	학습 및 평가 세부 사항	모델 학습·평가에 사용된 세부 조건 및 요소	M
	6-5	평가 성능지표 및 목표	모델 성능 평가를 위해 활용한 지표 및 목표 수치	M
데이터 활용 정보	7-1	활용 서비스	데이터 셋이 활용되는 서비스 분야의 설명 및 예시	M
	7-2	데이터 셋 구축년도	데이터 셋이 구축된 연도	M
	7-3	데이터 셋 구축 수량	데이터 셋 전체 수량	M
	7-4	획득·수집 공정 결과	획득·수집 공정 결과	R
	7-5	정제 공정 결과	정제 공정 결과	R
	7-6	가공 공정 결과	가공 공정 결과	R
데이터 사용 시 고려 사항	8-1	데이터 구축 유의사항	데이터 구축 공정 기준에 포함할 수 없으나 구축 시 주의해야 하는 사항 및 해결 방법	M
기타	9-1	데이터 셋 버전	데이터 셋의 버전 정보	M
	9-2	데이터 셋 변경 이력	데이터 셋 수정 또는 업데이트 이력	M

통계 정보, 데이터 저장구조및 포맷 정보), 데이터 셋이 어떤 정제 가공 기준으로 만들어졌는지(데이터 구축 절차 정보), 이 데이터 셋으로 무엇을 할 수 있으며(AI 모델 유효성 검증 정보, 데이터 활용 정보), 무엇을 하지 못하는지(데이터 사용 시 고려 사항, 기타)까지 설명할 수 있도록 한다. 세부 요소를 살펴보면 출처·권리(라이선스), 수집·제외 기준, 정제 규칙(중복·노이즈 제거, 비식별화), 가공·라벨링 규격과 검수 방식, 학습·검증·시험 분할 원칙, 데이터 분포, 파일 포맷과 저장구조처럼 재현성에 필요한 정보가 중심축을 이룬다.

예를 들어[2] 법률 문서 데이터는 수집처·선별 기준, 중복 제거·문장 분할·부적합 데이터 정제, 개인정보·민감정보 비식별화, 교차 검수, 학습·검증·시험 분할 등 실제 작업 흔적을 메타데이터로 남겨 '어떻게 만들어졌는가'를 설명한다. 또 법·제도 준수와 비윤리·혐오·편향 데이터 지양을 기준으로 제시해, 메타데이터가 책임 있는 개발을 위한 체크리스트가 될 수 있음을 보여준다.

4. 운영·리스크 관리의 관점

구축 공정 기반 메타데이터는 단순히 필드 목록이 아니라 목적·범위, 권장·금지 사용, 위험(편향·한계·오용 가능성)까지 이용자 관점에서 제공해야 할 설명서로 활용할 수 있다. 아울러 배포 형태(원천·가공본), 라이선스 조건, 접근 경로, 삭제·정정 요청 처리 절차까지 명시하면, 이용자는 준수해야 할 운영 규칙을 사전에 이해할 수 있다. 이는 데이터 제공자 입장에서도 분쟁 가능성을 줄이고, 변경이 발생했을 때 영향 범위를 추적하는 데 도움이 된다. 실무적으로도 배포 이후의 관리(버전, 변경이력, 배포 형태, 접근·보안 정책)까지 포함해 운영 가능한 데이터 셋으로 활용할 수 있다.

특히 실질적인 데이터 활용을 위해선 '사용 시 고려사항'의 비중을 높일 필요가 있다. 분포 편향, 수집 환경의 제약, 라벨링 주관성, 시간 경과에 따른 데이터 노후화는 모델의 일반화 성능과 공정성에 직접 영향을 준다. 또한 개인정보·민감정보 처리, 저작권 준수, 접근 권한과 로그 관리 같은 거버넌스 요소도 메타데이터 범주로 끌어올림으로써 책임 있는 활용의 전제 조건을 분명하게 규정할 수 있다.

5. 품질관리 관점에서의 접점

AI 모델의 지도학습을 위한 데이터 품질 관리 요소는 설계·획득·정제·라벨링 단계에 매핑하며 다양성, 출처신뢰성, 규격적합성, 통계적 충분성, 라벨 정확성, 유효성 등을 요구한다[3]. 이 프레임은 공정 기반 메타데이터와 맞닿아 있다. 품질은 검사 결과만이 아니라 어떤 기준으로 관리했는가의 증거로 설명돼야 하며, 그 증거를 담는 그릇이 메타데이터다. 실무적으로 활용할 때는 단계별 필수 메타데이터가 누락될 경우 공정을 중단하고, 표본 검수 결과와 오류 유형 및 데이터 분포 변화 감지에 따른 버전 관리를 메타데이터와 연계해 일관된 품질관리를 수행할 수 있다. 특히 유효성 검증 정보에 대한 메타데이터 항목을 통해 어떤 모델·임무(Task)에서 어떤 지표(정확도, F1-score, WER(Word Error Rate) 등)로 검증했는지, 학습·평가 환경과 데이터 분할이 무엇인지가 남아 있어야 성능 주장을 해석할 수 있다. 검증 결과를 메타데이터에 포함하면, 데이터 셋은 단순

2) AI 학습용 데이터 구축 공정 기반의 메타데이터, 부록 1 사용예시, TTAK.KO-10.1595, 2025
3) 지도학습을 위한 데이터 품질관리 요구사항 동향, TTA저널 vol199, 2022

참고 자료가 아니라 품질 보증서 역할을 하게 된다.
따라서 구축 공정 기반 메타데이터 구성요소를 통해 공정 과정의 소홀함이 있는지 여부를 파악하고, 품질검증의 기준과 관점을 부여할 수 있다.

6. 맺음말

메타데이터는 단순히 데이터를 배포하고 활용하기 위한 포장 작업이 아니라, 공정 설계와 동시에 고려·기획해야 하는 것이다. 이번 원고에서 제시한 구축 공정 기반 메타데이터 항목과 구성요소는 데이터 구축 시작 시점뿐만 아니라 데이터 확장 구축 시점에서 단계별 의사결정(수집·정제·가공 기준), 품질 지표와 샘플링 검수 기록, 버전 변경 사유를 자동으로 수집·연결하는 템플릿으로 활용될 수 있다. 향후에는 메타데이터 구성요소의 속성체계(ID, 요소명, 설명, 필수여부)를 참고해 템플릿을 만들고 단계별 로그를 자동 수집해 검수 결과와 연결한 뒤, 배포 시 '데이터 카드' 형태로 제공하는 운영 흐름까지 구축해 볼 계획이다.
메타데이터가 충실할수록 설명 가능성은 높아지고 재사용 비용은 낮아진다. 결국 좋은 데이터는 좋은 기록에서 시작함을 인지하고, 구축 공정 전 단계에서 메타데이터를 표준화·구조화해 지속적으로 축적하는 체계를 마련해야 한다. 이를 통해 데이터 출처와 품질을 명확히 증빙하고, 모델 학습의 근거를 일관되게 확보함으로써, 데이터 운영의 안정성과 AI 활용 신뢰도를 함께 높일 수 있다.

[본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성사업(2025-0360, 2025년 초거대AI 확산 생태계 조성사업)에 의해서 수행됐음]

[참고문헌]
[1] [제1권] AI 데이터 품질관리 가이드 v3.5.pdf, AIHub, 2025, https://www.aihub.or.kr/
[2] [제2권] AI 데이터 구축 가이드 v3.5.pdf, AIHub, 2025, https://www.aihub.or.kr/
[3] AI 학습용 데이터 구축 공정 기반의 메타데이터, TTAK.KO-10.1595, 2025
[4] 지도학습을 위한 데이터 품질관리 요구사항, TTAK.KO-10.1339, 2021

※ 출처: TTA 저널 제223호