

AI Ready Data의 개념 정립과 표준화 방향 연구

김 학 래*

중앙대학교 문헌정보학과 교수

Conceptual Framework and Standardization Direction for AI-Ready Data

Haklae Kim*

Professor, Department of Library and Information Science, Chung-Ang University, Seoul 06974, Korea

[요 약]

AI 시스템에서 데이터를 활용하려면 데이터 품질을 넘어 특정 AI 목적에 실제로 사용할 수 있는 상태인지를 판단할 수 있어야 한다. 그러나 기존 데이터 품질과 AI 준비 데이터(AI Ready Data) 논의는 필요한 요소를 제시하는 데 머물러 데이터의 준비 상태와 활용 가능성을 일관된 기준으로 판정하기 어렵다. 본 연구는 AI Ready Data를 데이터의 고정된 속성이 아니라 진단과 변환에 따라 달라지는 상태로 보고, 이를 판정하는 형식 모델을 제안한다. 데이터셋의 공통 준비 상태와 활용 목적별 준비 상태를 구분하여 Raw, Discoverable, Quality-Ready, Purpose-Ready, Stale로 판정한다. 결정 게이트로 상태 전이를 판정하고, 공통 품질 등급(Q-Tier)과 AI 활용 목적별 준비 등급(P-Tier)을 구분하여 판정한다. 판정 결과와 근거는 기계 판독 가능한 AIRD Pack으로 구성한다. 공공데이터 적용 결과, 동일한 데이터도 품질 상태와 활용 목적에 따라 준비도 판정이 달라짐을 확인하였다. 이를 바탕으로 AI Ready Data의 표준화 기능 계층과 공공데이터 정책 연계 방향을 제시한다.

[Abstract]

For data to be used in AI systems, their quality and ability to support a specific AI purpose must be determined. Existing data quality approaches and discussions regarding AI-Ready Data (AIRD) largely focus on identifying required elements, making it difficult to assess data readiness and usability according to consistent criteria. This study views AIRD not as a fixed property but as a state that changes through diagnosis and transformation and proposes a formal model for assessing such states. The model assigns a dataset one common readiness state and one readiness state per AI purpose, determines state transitions at decision gates, and distinguishes a common quality tier (Q-Tier) from purpose-specific readiness tiers (P-Tiers). Assessment results and supporting evidence are organized into a machine-readable AIRD Pack. Application to public data shows that readiness assessments of the same data can vary depending on their quality state and intended use. Furthermore, this study proposes functional layers for AIRD standardization and directions for linking the framework to public data policy.

색인어 : AI Ready Data, 데이터 준비 상태, 상태 전이 모델, 목적 준비도 벡터, 데이터 거버넌스

Keyword : AI-Ready Data, Data Readiness State, State Transition Model, Purpose Readiness Vector, Data Governance

<http://dx.doi.org/10.9728/dcs.2026.27.9.2695>



This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 02 August 2026; **Revised** 21 August 2026

Accepted 31 August 2026

***Corresponding Author; Haklae Kim**

Tel: +82-2-820-5561

E-mail: haklaekim@cau.ac.kr

1. 서 론

인공지능 기술은 의료, 제조, 금융, 행정 등 여러 분야로 확산되고 있으며, 각국 정부도 인공지능(Artificial Intelligence, AI)을 국가 전략에 포함하고 있다. AI 모델의 성능이 향상되어도 데이터 문제로 기대한 운영 가치를 얻지 못하는 시스템은 여전히 존재한다. Sambasivan et al.은 AI 실무자 인터뷰를 바탕으로 데이터 문제의 누적과 과급을 ‘데이터 종속(data cascade)’으로 설명하였다[1]. 이러한 데이터 문제는 AI 파이프라인에서 구조 모호성(Schema Ambiguity), 의미 공백(Semantic Gap), 운용 불가(Non-operability)의 형태로 나타날 수 있다. 구조 모호성은 스키마가 불명확하여 데이터를 파싱하지 못하는 문제이고, 의미 공백은 데이터의 의미를 해석할 맥락 정보가 부족한 문제다. 운용 불가는 품질이 일정 수준을 충족하더라도 파이프라인이 요구하는 형식과 구조를 갖추지 못한 상태를 뜻한다. 데이터의 정제, 출처, 처리 규칙이 함께 관리되지 않으면 자동화된 처리 과정에서 오류가 누적되고 운용 단계의 문제로 이어질 수 있다[1].

기존 접근은 데이터 품질과 AI Ready Data 논의로 구분할 수 있다. ISO/IEC 25012에 기반한 데이터 품질(Data Quality) 접근은 데이터의 품질 특성을 정의하지만, AI 파이프라인이 데이터를 자동으로 처리하고 활용할 수 있는지 직접 판정하지 않는다[2]. 품질 점수가 높아도 파이프라인에 투입하기 어려울 수 있고, 낮은 점수가 특정 목적의 활용 가능성을 모두 부정하는 것도 아니다. AI Ready Data에 관한 논의는 오픈데이터연구소(Open Data Institute, ODI), 미국 상무부(Department of Commerce, DoC), 경제협력개발기구(Organisation for Economic Co-operation and Development, OECD), Majithia et al., Verhulst et al.의 FAIR-R 등으로 확산되고 있다[3]~[7]. 그러나 이들 논의는 준비 상태의 전이, 활용 목적에 따른 준비도의 차이, AI 파이프라인의 자동 처리 가능성을 하나의 판정 체계로 제시하지 않는다. AI Ready의 구성 요소를 정의하더라도 실제 데이터가 해당 상태에 도달했는지를 판정하는 기준은 별도로 필요하다.

정책 분야에서도 같은 문제가 확인된다. 행정안전부는 2026년 ‘공공데이터의 인공지능 친화적 관리 가이드라인’을 발표하였고[8], 이정윤·김학래는 AI Ready Data의 개념과 한국 정책의 현황을 정리하였다[9]. 영국 정부디지털청(Government Digital Service, GDS)과 과학혁신기술부(Department for Science, Innovation and Technology, DSIT)는 정부 데이터셋의 AI 준비도를 4개 요소로 정의하고 자가진단 체크리스트를 제시하였다[10].

정책·개념 프레임워크는 필요한 요소를 구체화하였지만, 자가진단 체크리스트나 정성 척도를 넘어 데이터가 AI 파이프라인에 실제로 투입 가능한지를 판정하는 형식 기준까지 제시하지는 않는다. 이정윤·김학래는 조작적 정의와 정량 지표의 개발을 향후 과제로 제시하였다[9]. 본 연구는 이러한

공백을 바탕으로 준비 상태의 전이와 활용 목적별 준비도를 판정하는 개념적·이론적 프레임워크를 구성하고, 다음 연구 질문을 설정하였다.

- RQ1. 데이터 준비 상태를 정적 품질 속성이 아닌 동적 상태(dynamic state)로 어떻게 형식화할 것인가.
- RQ2. 이질적인 AI 활용 목적에 따라 준비도를 어떻게 운용 가능하게 정의할 것인가.
- RQ3. 이러한 모델을 실제 데이터 거버넌스·정책과 어떻게 통합할 것인가.

각 연구 질문은 다음과 같이 모델 설계에 반영하였다. RQ1은 데이터셋의 준비 상태와 운영 활동을 구분하고, 5개 준비 상태의 전이 조건을 결정 게이트로 규정한다. RQ2는 활용 목적 유형마다 준비도를 독립적으로 판정하고, 품질 등급과 목적 등급을 구분하여 기록하는 체계를 제안한다. RQ3은 제안한 모델을 도메인 중립적인 표준화 기능 계층으로 구성하고, 공공데이터를 대표 적용 분야로 설정하여 정책 연계 방향을 제시한다.

연구는 문헌·표준 분석, 공백 도출, 모델 설계, 사례 적용, 표준화 방향 도출의 5단계로 수행하였다. 데이터 준비성 연구, AI Ready Data 프레임워크, 데이터 품질·메타데이터 국제 표준을 검토하고, 판정 가능성을 기준으로 상태 전이 모델, 목적 의존성, 운용 기준의 공백을 도출하였다. 준비 상태와 운영 활동을 분리한 게이트 기반 전이 모델과 목적별 준비도 판정 체계를 설계한 뒤, 전국등록공장현황 공공데이터(<https://www.data.go.kr/data/15105482/fileData.do>)의 두 가지 형태에 적용하여 판정 결과를 대조하였다. 마지막으로 제안한 모델을 표준화 기능 계층으로 정리하고 정책 연계 방향을 제시하였다. 사례 적용은 판정 절차와 산출 형태를 설명하는 예시(worked example)이며, 실측에 근거한 성능 검증은 아니다. 제안한 모델의 일부는 제정 중인 정보통신단체 표준 ‘AI 레디 데이터’ 시리즈의 설계에 반영되었다. 논문에서 대응한 표준안은 2026년 8월 제출예정본이다.

본 논문의 구성은 다음과 같다. 2장은 AI Ready Data의 최신 동향, 데이터셋 문서화 표준, 정책 프레임워크를 검토한다. 3장은 데이터 품질과 AI 준비 상태의 개념적 차이, AI Ready Data의 정의, 그리고 이를 구현하는 프레임워크의 구성을 제시하고, 그 종합으로서 필요조건 명제(명제 1)와 공백 명제(명제 2)를 정의한다. 4장은 준비 상태와 운영 활동을 분리하는 라이프사이클을 도입하고 상태 전이를 결정 게이트 조건으로 형식화한다. 5장은 목적 준비도 벡터를 정의하고, 6장은 공장현황 공공데이터를 사례로 모델의 작동을 검토하고, 7장은 개념·측정·평가 체계의 표준화와 정책 연계의 제안 로드맵을 제시한다. 8장은 연구의 결론과 향후 연구방향을 기술한다.

II. 데이터 준비성과 AI Ready Data 프레임워크의 한계

2-1 데이터 준비성 프레임워크

데이터 준비성(data readiness) 개념의 학술적 출발점은 Lawrence(2017)의 데이터 준비 수준(Data Readiness Levels, DRL)이다[11]. DRL은 항공우주 분야의 기술성숙도(Technology Readiness Level)를 모방하여 데이터를 접근성 수준에서 과제 맥락 적합성 수준까지 분류하고, 데이터를 정적 속성이 아니라 목적 지향적 준비 상태로 개념화하였다. ABC 프레임워크는 이를 밴드 분류와 정제 절차로 구체화하였고[12], 준비도 보고서는 평가 결과와 처리 이력을 데이터셋에 첨부하는 방식을 제안하였다[13]. 이후 의학 데이터셋의 기계학습 준비성을 계산하는 DREAMER[14], 140여 편의 문헌에서 준비도 지표를 분류한 연구[15], 과학 데이터 파이프라인을 2차원 성숙도 매트릭스로 분석한 연구[16]가 제시되었다. 기존 데이터 준비성 연구는 데이터가 어느 수준에 있는지를 기술하지만, 본 연구의 검토 범위에서 다음 수준으로 전이하는 조건과 결정 게이트는 형식화되지 않았고 활용 목적에 따른 준비도의 차이도 체계화되지 않았다[15].

2-2 AI Ready Data 특화 프레임워크: 최신 동향과 한계

AI Ready Data는 AI의 학습·추론·의사결정에 적합하도록 필요한 조건을 갖춘 데이터의 상태를 지칭하며, 아직 국제적으로 합의된 정의나 측정 기준이 없는 초기 개념이다[9]. 강조점은 이해관계자마다 다르다. 산업계는 목적 적합성과 자동화된 파이프라인을, 학계·연구기관은 재현성과 책임성, 국제기구·표준화 기구는 상호운용성과 규범 준수를 중시한다. 이정윤·김학래는 65편의 문헌을 고찰하여 이러한 접근을 품질·신뢰성, 표준화·상호운용성, 거버넌스·책임성, 다양성·포용성, 지속적 운영 가능성의 다섯 요소로 종합한다[9].

개념 정립과 병행하여 2025~2026년에는 AI Ready Data

를 직접 대상으로 한 프레임워크가 발표되었다. 접근성, 품질·대표성, 거버넌스, 맥락·문서화의 4개 범주로 구성된 체크리스트[6], 유럽 데이터 생태계를 위한 AI Ready Data Products 청사진[17], 생의학 도메인의 준비성 프레임워크 FAIRSCAPE[18]가 그 예다.

AI Ready Data에 관한 논의는 데이터 관리의 국제 원칙인 FAIR(Findable, Accessible, Interoperable, Reusable)[19]와 밀접하게 연결된다. Verhulst et al.의 FAIR-R은 FAIR에 AI 준비성을 뜻하는 'R'을 추가하여 발견·재사용 기능성과 AI 활용 준비성을 함께 다루었고[7], AI Magazine의 특집도 FAIR 원칙과 기계학습·AI 준비성의 관계를 다루었다[20]. 이들 접근은 FAIR의 발견·접근·상호운용·재사용 원칙을 바탕으로 AI 활용에 필요한 조건을 추가한다. 다만 FAIR 자체는 특정 AI 파이프라인의 자동 활용 가능성(machine-actionability)이나 활용 목적에 따른 준비도의 차이를 판정하지 않는다[19]. FAIR-R도 AI 준비성의 구성 요소를 제시하지만, 실제 데이터의 준비 상태와 목적별 준비도를 판정하는 형식 기준까지 규정하지 않는다[7].

2-3 데이터셋 문서화와 메타데이터 표준

데이터셋 문서화와 메타데이터 표준은 데이터셋의 목적·구성·수집 과정·편향을 문서화하는 체계[21], 기계학습 프레임워크가 데이터셋을 직접 적재하는 메타데이터 형식[22], 데이터 카탈로그 어휘[23], 품질을 메타데이터로 표현하는 어휘[24], 분석·기계학습용 데이터 품질의 모델·측정·관리 체계[25]로 발전하였다.

이들 연구와 표준은 데이터의 기술(description)과 메타데이터 표준화를 지원하지만, 데이터가 AI 파이프라인에 투입 가능한 상태인지 판정하는 상태 모델과 활용 목적별 메타데이터 요건의 차이를 함께 규정하지는 않는다. Croissant 1.0은 ML에 특화된 Layer 3 메타데이터를 정의하지만 RAG, KG, Stats 등 다른 목적 유형은 포괄하지 않는다.

한편 데이터 엔지니어링 영역에서 Great Expectations,

표 1. 주요 프레임워크와 본 연구의 차별성 비교

Table 1. Comparison of major frameworks and the differentiation of this study

Aspect	MOIS Guidelines (2026)[8]	Majithia et al. (2026)[6]	FAIR-R (2025)	Brewer et al. (2026)[16]	This study
State concept	Openness and quality levels	Checklist	Fifth principle "R"	Five-level operational scale	Gate-based readiness state model (states + per-purpose predicate)
Purpose dependency	Single quality criterion	Single framework	Not defined	Domain-specific	PRV: per-purpose readiness vector
Decision criteria	Quality score	Item-by-item check	Conceptual level	Maturity matrix	QI_{min} , mandatory gates, per-tier pass lines
State transition condition	Not defined	Not defined	Not defined	Not defined	Formalized as gate conditions
Threshold treatment	Not specified	Not specified	Not specified	Not specified	Externalized as parameters

TFDV 같은 런타임 검증 도구, 스키마 레지스트리, 그리고 스키마·품질·SLA를 생산자·소비자 계약으로 결합하는 데이터 계약(data contracts) 관행이 발전해 왔으며, 이 흐름은 본 연구가 제안하는 AI Ready Data(AIRD) Pack과 문제의식을 공유한다. 그러나 이들은 개별 파이프라인의 운영 검증에 초점을 두어, 데이터셋 수준의 공개·교환 가능한 준비 상태 등급과 목적별 준비도를 형식화하지 않는다.

2-4 정책 프레임워크

정책 영역의 출발점은 ODI의 AI Ready 데이터 프레임워크로, 접근성·상호운용성, 품질·대표성, 거버넌스·신뢰, 맥락·문서화를 핵심 요소로 제시하였다[3]. 영국 GDS·DSIT는 이 ODI 4요소를 토대로 정부 데이터셋의 AI 준비도를 재구성하고, 데이터를 Bronze(원천)·Silver(정제)·Gold(큐레이션) 계층으로 구분하는 메달리온 구조와 자가진단 체크리스트를 함께 제시하였다[10]. 그 밖의 정책 문헌도 개방형 정부 데이터의 AI 준비도와 정부의 데이터 공급자 역할을 강조하며 [4],[5],[26],[27], 국내 연구는 공공데이터 전환계층 모델로 한국 정책을 진단하였다[9]. 이들 정책·분석 문헌은 인간 중심으로 데이터 발행자의 역할과 거버넌스를 설명하며, AI 시스템이 데이터를 직접 파싱·처리하는 데 필요한 운용성(operability) 요건과 목적별 준비 요건의 차별화를 형식 기준으로 규정하지 않는다. 행정안전부 가이드라인도 같은 공백을 공유한다[8]. 표 1은 기존 프레임워크가 AI Ready의 구성 요소를 주로 정의하고, 상태 전이 조건은 명시적으로 규정하지 않는다는 차이를 정리한다. 이 연구는 AI Ready의 준비 상태와 전이 조건을 규정하고, 활용 목적별 준비도를 독립적으로 판정하는 체계를 제안한다.

III. 개념적 프레임워크

3-1 데이터 품질과 AI 준비 상태의 구별

데이터 품질은 데이터가 의도한 용도에 적합한 정도를 평가하는 개념이다. ISO/IEC 25012는 데이터 품질을 고유(inherent) 관점과 시스템 의존(system-dependent) 관점에 걸친 15개 특성으로 정의한다[2]. 본 연구는 특정 시스템이나 AI 활용 목적에 종속되지 않고 데이터 자체에서 공통적으로 진단할 수 있는 특성을 기준으로 정확성·완전성·일관성·적시성을 선택하였다. 국내 공공데이터에 적용되는 기존 품질진단 체계와의 연속성을 고려하여 유효성·유일성을 추가하고, 이를 6개 공통 차원으로 구성하였다.

6개 차원은 데이터의 값과 규칙을 중심으로 품질을 진단한다. 그러나 AI 시스템이 데이터를 사용하려면 품질 진단만으로 확인할 수 없는 조건도 충족해야 한다. 특정 AI 처리 파이프라인이 데이터를 자동으로 적재하고 검증하고 활용할 수 있어야 한다. 데이터 품질과 파이프라인 처리 가능성을 모두 충족한 상태를 AI 준비 상태로 정의하고, 데이터 품질과 AI 준비 상태의 관계를 다음 명제로 규정한다.

명제 1 (필요조건 명제). 데이터 품질은 AI 준비 상태(AI readiness)의 필요조건이지만 충분조건이 아니다.

품질 진단을 통과하지 못한 데이터는 AI 준비 상태에 도달하지 못하며, 품질 진단을 통과했다라도 파이프라인이 그 데이터를 처리할 수 있는지 확인하지 못하면 이 상태가 성립하지 않는다. 명제 1은 4장의 목적별 판정과 5장의 등급 체계를 설계하는 전제가 된다.

본 연구는 6개 품질 차원에 AI 활용 가능성과 직접 관련되는 기계 판독성(D7; machine-readability)을 추가하여 7개 차원으로 구성하였다. 기계 판독성은 AI 도구가 데이터를 직접 적재할 수 있는지를 측정하며, 기존 6개 차원에서 확인하지 않는 영역이다. 이 차원은 데이터셋이 CSV, Parquet, JSON 등 개방형 포맷과 명시적 스키마, API 접근 체계를 갖추었는지 측정한다. 기계 판독성은 다른 차원의 점수로 대체하지 않는 조건이며, 판정 규칙은 4-2절에서 규정한다.

7개 차원은 D1 완전성, D2 일관성, D3 정확성, D4 적시성, D5 유효성, D6 유일성, D7 기계 판독성으로 표기한다. 표준화 논의는 7개 차원에 걸쳐 16개 정식 지표와 측정 공식, 전제 조건, 데이터 유형별 적용 기준을 규정하며, 이 논문은 차원 단위의 판정 구조를 정의한다.

3-2 AI Ready Data의 정의

AI Ready Data는 인공지능 모델의 학습·추론·평가·검색증강생성·지식그래프 구축 등 특정 AI 활용 목적에 맞게 품질진단, 포맷 변환, 메타데이터 기록, 목적별 최적화를 수행하고, 사전에 명시되지 않은 수작업 해석이나 구조 재설계 없이 AI 처리 파이프라인이 자동으로 적재·검증·활용할 수 있는 상태로 관리되는 데이터로 정의한다. 이 정의는 특정 활용 목적을 전제로 하지 않는 공통 기반인 Quality-Ready와 특정 활용 목적의 요건을 충족한 Purpose-Ready를 함께 포함한다. AI Ready Data는 두 가지 준비 상태를 포괄하는 개념이다. 준비 상태와 상태 전이는 4장에서 형식화하고, 활용 목적별 준비도와 등급은 5장에서 구체화한다.

이 정의는 상태성(Statefulness), 목적 의존성(Purpose Dependency), 파이프라인 운용성(Pipeline Operability)의 세 가지 성질을 함축한다. AI Ready Data는 고정된 속성이 아니라 지속적인 관리로 유지되는 동적 상태이며, 상태의 도달과 상실 조건은 4장에서 형식화한다. 같은 데이터셋도 기계 학습 목적으로 준비되었으나 지식그래프 구축에는 적합하지 않을 수 있다. 파이프라인 운용성은 사전에 명시되지 않은 수작업 전처리 없이 파이프라인이 데이터를 자동으로 적재하고

활용하는 성질을 뜻한다.

세 성질은 상태 기록, 목적별 요건, 처리 규칙을 데이터와 함께 관리하는 것을 전제한다. 따라서 AI Ready Data는 고립된 값(value)의 집합이 아니라 메타정보와 처리 규칙을 함께 갖춘 자기 설명 가능한 패키지(package) 단위로 구성된다.

3-3 프레임워크의 구성

AI Ready Data의 판정 구조는 그림 1과 같이 준비 상태와 상태별 확인 항목으로 구성한다. 데이터셋의 준비 상태는 단계와 결정 게이트에 따라 진행되며, 각 상태에서 확정하는 항목을 함께 제시한다. 마름모는 상태 전이를 판정하는 결정 게이트를 나타낸다. 이정운·김학래는 AI Ready Data에 필요한 관리 요건을 5개로 정리하였다[9]. 이 연구는 이 관리 요건을 판정 가능한 항목으로 구체화한다.

1) 메타데이터 계층

메타데이터 계층은 활용 목적과 관계없이 모든 데이터셋에 적용하는 공통 기반이며, 발견·이해·운용의 3계층으로 구분한다. 발견 계층(Layer 1; discovery)은 데이터를 찾고 접근하는 데 필요한 정보를 기술한다. 이해 계층(Layer 2; understanding)은 품질, 이력, 출처를 기술한다. 운용 계층(Layer 3; operability)은 AI 도구가 데이터를 처리하고 적재할 때 참조하는 정보를 기술하며, 활용 목적에 따라 달라진다.

상태별 메타데이터 기술 범위는 그림 1에 표현되어 있다. 메타데이터는 데이터셋의 준비 상태에 해당하는 어휘를 사용하여 기술하며, 상태가 전이되면 기술하는 정보의 범위가 누적하여 확장된다. Discoverable 상태는 발견 계층의 어휘로 기술하고, Quality-Ready 상태는 이해 계층의 어휘를 사용한 정보를 추가한다. Purpose-Ready 판정을 받은 데이터셋은 운용 계층의 정보를 추가한 메타데이터로 구성된다.

운용 계층은 활용 목적 유형별 멀티 프로파일로 구성한다.

ML·DL 형식은 Croissant 1.0, 지식그래프와 온톨로지는 VoID, 통계 형식은 SDMX 등 국제 표준 어휘를 기반으로 한다. 기존 국제 표준 어휘가 포괄하지 못하는 RAG·FineTuning 등의 요건은 aird-rag·aird-finetuning 등의 확장 프로파일로 추가할 수 있다. 다만 등록 상태의 프로파일이 확정되기 전까지 확장 프로파일은 참고용이며, 해당 활용 목적의 공식 등급 판정에는 사용하지 않는다.

2) 판정 체계

데이터셋은 품질 등급(Q-Tier; Quality-Tier), 목적 등급(P-Tier; Purpose-Tier), 진단 성숙도(DM; diagnostic maturity)로 판정한다. Q-Tier는 공통 품질 수준, P-Tier는 특정 활용 목적의 적합도를 나타낸다. DM은 품질 판정의 근거가 되는 측정 범위를 나타내며 DM-0(기본 진단), DM-1(부분 진단), DM-2(표준 진단)의 3개 수준으로 구분한다. Q-Tier, P-Tier, DM은 서로 독립적으로 판정하며 하나의 결과로 다른 결과를 추론할 수 없다.

진단 성숙도는 데이터 품질 수준을 나타내지 않는다. 따라서 진단 범위가 넓어도 품질 등급이 높아지지 않으며, 진단 범위가 좁으면 산출한 등급을 공식 판정으로 사용할 수 없다.

목적 등급은 ML의 레이블 완전성, RAG의 청킹 가능 구조, KG의 엔티티 연결 가능성처럼 활용 목적마다 필요한 조건이 다르다는 점을 반영한다. P-Tier는 1~3의 등급으로 구분하며, Q-Tier와 P-Tier의 산정 규칙과 등급 경계는 4-2절과 5-1절에서 설명한다.

3) 신뢰·윤리 기록

신뢰·윤리 기록은 활용 목적과 관계없이 정의되는 공통 기반으로, 라이선스, 개인정보 처리, AI 편향 위험, 책임 추적 가능성(provenance)을 포함한다. Croissant의 RAI(responsible AI) 확장 어휘에서 rai:dataBiases, rai:knownLimitations, rai:dataCollectionMissingData의

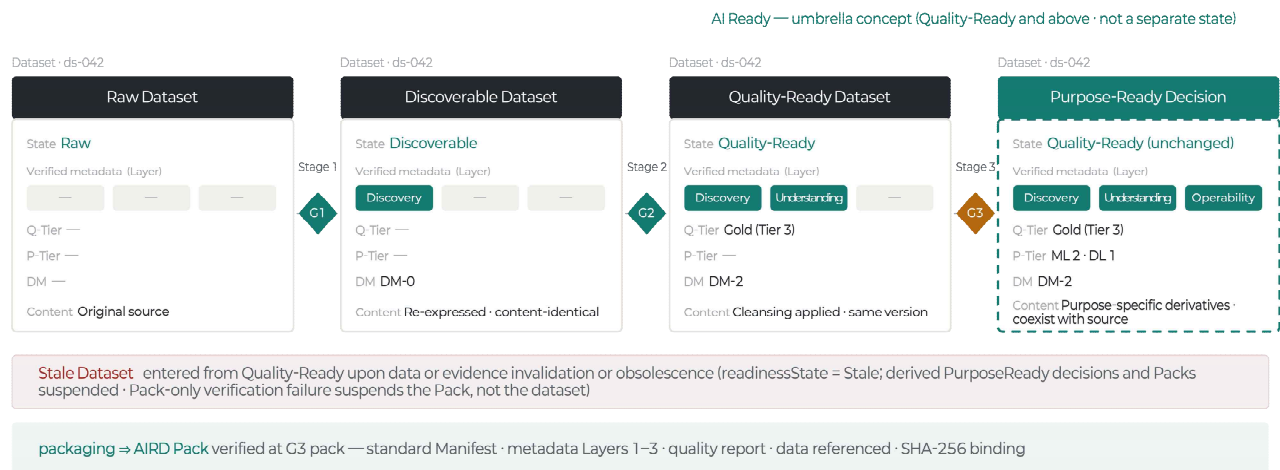


그림 1. AI Ready Data 프레임워크의 구성

Fig. 1. Composition of the AI ready data framework

세 속성을 선별하여 Quality-Ready 전이의 필수 조건으로 규정한다. 세 속성은 최소 신뢰 문서화 집합(minimum trust-documentation set)을 구성한다. 품질 점수는 이 조건을 대신하지 못하므로, 품질 차원의 점수가 높아도 신뢰·윤리 기록이 없으면 전이가 성립하지 않는다.

4) 판정 결과의 배포 단위

AI Ready Data의 준비 상태와 판정 결과는 마니페스트에 기록하고, Quality-Ready와 Purpose-Ready 판정을 통과한 결과는 AIRD Pack으로 구성하여 배포한다. 마니페스트는 데이터셋의 준비 상태와 판정 결과를 기록하며, 상태나 내용이 변경되면 새로운 버전으로 갱신한다. Quality-Ready 팩은 마니페스트, 품질 진단 리포트, 팩 마니페스트로 구성한다.

Purpose-Ready 팩은 Quality-Ready 판정 결과에 활용 목적별 운용 정보를 추가하여 구성한다. 마니페스트, 품질 진단 리포트, 활용 목적별 운용 파일, 팩 마니페스트가 Purpose-Ready 팩에 포함된다. Purpose-Ready 판정은 활용 목적별로 독립적으로 수행하므로 하나의 데이터셋에서 여러 개의 Purpose-Ready 팩을 생성할 수 있다. 데이터 파일은 AIRD Pack에 직접 포함하지 않고 참조한다. AIRD Pack의 정규 직렬화 형식[29], 파일 무결성 확인, 물리적 전달 형식은 표준화 논의에서 규정한다.

AI Ready Data의 구성 요소는 데이터가 충족해야 할 요건과 판정 결과를 기록·배포하는 구조를 규정한다. 데이터가 각 요건을 충족했는지를 판단하려면 명시적인 판정 기준이 필요하다. 이 판정 기준은 2장에서 검토한 기존 프레임워크에도 없으며, 그 공백을 정식화하면 다음과 같다.

명제 2 (공백 명제). 기존 프레임워크는 데이터가 AI 파이프라인에서 자동으로 처리·활용될 수 있는지를 판정하는 명시적 상태 전이 모델과 운용 기준을 갖추지 못한다.

명제 2는 해결해야 할 공백을 규정하고, 명제 1은 모델이 충족해야 할 제약을 규정한다. 두 명제를 바탕으로 준비 상태의 전이 조건과 목적 준비도의 판정 기준을 정의한다.

IV. 결정 게이트 기반 준비 상태 전이 모델(RQ1에 대한 응답)

데이터셋의 준비 상태 모델은 5개의 상태, 3개의 결정 게이트, 전이와 갱신 규칙으로 구성한다. G1은 Raw에서 Discoverable로, G2는 Discoverable에서 Quality-Ready로, G3는 Quality-Ready에서 Purpose-Ready로의 전이를 판정한다. G1과 G2는 데이터셋 단위로 판정하고, G3는 활용 목적 유형마다 독립적으로 판정한다. G2를 통과하면

Quality-Ready Pack을 생성하고, G3를 통과하면 해당 활용 목적의 Purpose-Ready Pack을 생성한다. Pack의 구조와 무결성은 교환·등록 단계에서 검증한다. 메타데이터는 준비 상태에 해당하는 어휘로 기술하며, 상태 전이에 따라 기술하는 정보와 판정 결과를 누적하여 기록한다.

4-1 준비 상태와 운영 활동

AI Ready Data 모델은 준비 상태(readiness state)와 운영 활동(operational activity)을 구분한다.

준비 상태는 데이터셋의 AI 활용 준비 수준을 나타낸다. G1은 원시에서 Discoverable로, G2는 Discoverable에서 Quality-Ready로, G3는 Quality-Ready에서 Purpose-Ready로의 전이를 판정한다. G1과 G2는 데이터셋마다 한 번 판정하고, G3는 활용 목적 유형마다 독립적으로 판정한다. 하나의 활용 목적 유형에서 판정에 실패하더라도 다른 유형의 판정에는 영향을 미치지 않는다.

준비 상태는 원시(Raw), Discoverable, Quality-Ready, Purpose-Ready, 스테일(Stale)의 5개 값으로 구분하며, 두 축으로 부여한다. 데이터셋에는 Raw, Discoverable, Quality-Ready, Stale 가운데 하나를 부여하고, Purpose-Ready와 Stale은 활용 목적 유형마다 별도로 부여한다. Raw는 원천에서 제공된 직후의 진입 상태다. Discoverable은 기계 구문분석 가능한(machine-parsable) 포맷으로 등록하고 발견·접근 메타데이터(Layer 1)를 갖춘 상태다. 데이터의 식별과 접근은 가능하지만 품질과 기계 판독성 수준은 아직 판정하지 않는다. Quality-Ready는 7차원 품질 진단을 통과하여 목적별 변환과 적합도 평가를 수행할 수 있는 공통 품질 기반을 갖춘 상태다. Purpose-Ready는 Quality-Ready 데이터셋이 특정 활용 목적의 요건을 충족한 상태다. 동일한 데이터셋이 ML 목적의 Purpose-Ready 판정을 통과하고 KG 목적의 판정은 통과하지 못할 수 있다. 활용 목적 유형별 충족 수준은 목적 등급(P-Tier)으로 나타내고, 판정 결과는 마니페스트의 목적 준비도 항목에 기재한다.

스테일(Stale)은 Quality-Ready 또는 Purpose-Ready 판정의 유효성이 만료된 상태다. 기존 등급은 유지하지만 효력을 정지하고, 재검증을 통과하기 전까지 유효한 판정으로 사용하지 않는다. 스테일은 준비 수준이 높아진 상태가 아니라 기존 판정의 효력이 정지된 상태다.

운영 활동은 수집, 품질 평가, 변환, 패키징, 교환, 활용의 절차로 구성된다. 운영 활동은 자체적인 준비 상태를 갖지 않으며 데이터의 상태 전이에 필요한 처리와 메타데이터 생성을 수행한다.

상태 전이와 직접 관련되는 운영 활동은 3개 단계(stage)로 구분한다. 단계 1은 데이터를 수집하고 개방 형식으로 변환하여 발견 계층의 정보를 기술한다. 단계 2는 데이터 품질을 진단·정제하고 이해 계층의 정보를 추가한다. 단계 3은 활용 목적에 맞게 데이터를 변환하고 운용 계층의 정보를 추가

한다. 각 단계가 끝나면 결정 게이트가 전이 조건의 충족 여부를 판정한다. 단계 1은 G1, 단계 2는 G2, 단계 3은 G3에 각각 대응한다.

단계와 결정 게이트는 처리와 판정의 역할을 분리한다. 단계는 데이터 처리와 메타데이터 생성을 담당하고, 결정 게이트는 처리 결과가 전이 조건을 충족하는지 판정한다. 처리와 판정을 분리하면 판정 근거와 결과를 기록하고 처리 도구에 종속되지 않는 판정 규칙을 적용할 수 있다. 그림 1은 상태 사이의 구간으로 단계를, 마름모로 결정 게이트를 표시한다.

모든 상태 전이는 결정 게이트의 판정이나 표 2에서 정의한 트리거에 따라 성립한다. 결정 게이트나 명시적 트리거를 거치지 않은 상태 선언은 효력을 갖지 않는다.

4-2 상태전이의 결정 게이트

상태 전이는 명시적인 게이트 조건을 충족할 때에만 허용된다. 결정 게이트는 발행자의 임의적인 상태 선언을 제한하

고 파이프라인 자동화와 감사(audit)에 동일한 판정 기준을 적용한다.

G1(진입). G1은 원천 데이터를 Discoverable 상태로 등록하기 위한 최소 조건을 판정한다. 단계 1에서 변환한 데이터가 Layer 1 필수 항목을 갖추고 발견 계층의 검증 규칙을 통과하면 G1이 성립한다. 진입 조건은 구문분석 가능성으로 한정하고 기계 판독성은 D7에서 평가한다. 기계 판독성이 기준에 미달하면 G2에서 전이를 차단한다. 따라서 스키마가 없는 원시 XLSX는 G1을 통과할 수 있지만 G2는 통과하지 못한다.

G2(품질 검증). G2는 Discoverable에서 Quality-Ready로 전이하기 위한 조건을 판정한다. G2는 품질, 메타데이터, 신뢰·윤리, 판정 근거, 계보의 5개 조건을 모두 충족해야 통과한다. 품질 조건은 최소 차원 점수 QI_{min} , 기계 판독성 D7, 후보 등급의 차원별 통과선으로 판정한다. 메타데이터는 Layer 2 필수 항목, 신뢰·윤리는 관련 정보의 기록 여부, 판정 근거는 품질 진단 결과와 마니페스트의 일치 여부, 계보는 원천

표 2. 상태 전이와 갱신 규칙(Superseded는 데이터셋의 상태 표지가 아니라, 후속 버전의 Pack이 등록될 때 종전 Pack에 부여되는 팩 운영 상태의 값이다)

Table 2. State transition and update rules (Superseded is not a dataset state marker but a value of the pack's operational status, assigned to a previous pack when a successor-version pack is registered)

Transition / update	Condition	Rule
Raw → Discoverable (admission)	G1 = PASS	Admission transition; discovery layer fixed
Discoverable → Quality-Ready (forward)	G2 = PASS	Forward transition; Q-Tier assigned; a quality-ready pack is generated
Quality-Ready → Purpose-Ready (forward, per purpose)	G3 = PASS for purpose p; a registered profile and an active official threshold profile exist for p	The dataset enters Purpose-Ready for purpose p; P-Tier assigned for p; the source dataset remains Quality-Ready; a purpose-specific pack is generated
G3 evaluation from Discoverable (prohibited)	—	Prohibited (SHALL NOT) — blocks purpose evaluation for quality-unverified data
Quality-Ready → Discoverable (manual invalidation)	Publisher declaration	Reason must be recorded; Q-Tier and all derived P-Tiers invalidated
Quality-Ready → Stale (stale entry)	(a) source version (sourceDataVersion) updated; (b) type C (obsolescence) feedback; (c) $dct:accrualPeriodicity \times \alpha$ elapsed — determined and executed by the pipeline operator after confirming the trigger	Derived Purpose-Ready derivatives also enter Stale; the tiers of the preceding state are retained but suspended; the pack's operational status is not changed; the registry returns the dataset's lifecycle state alongside the pack in discovery responses
Purpose-Ready → Stale (stale entry, per purpose)	Trigger affects only the purpose-specific derivative or its operation file	Only the readiness for purpose p enters Stale; the dataset remains Quality-Ready; the P-Tier for p is retained but suspended
Stale → preceding state (recovery 1: re-verification)	Trigger resolvable by restoring metadata or decision evidence	Re-run stage 2 and pass G2 again to return to Quality-Ready; re-run stage 3 and pass G3 again to return a derivative to Purpose-Ready; generate a new pack and reference the previous pack as its earlier version
Stale → version succession (recovery 2)	Trigger affects the data body (source update or type B)	Re-run from stage 1 within the same dataset and increment the dataset version; the previous pack becomes Superseded when the successor pack is registered
Type A (structural error) signal (update)	Pack structure or referenced-file verification failure	Re-adjudicate G2 or G3 according to the pack type; if the error is confirmed, re-run the corresponding stage. States are unchanged
Type B (quality degradation) signal (update)	Consumer feedback	The pipeline operator analyses the scope of impact and decides whether to re-run stage 2

데이터와 전처리 활동의 이력을 확인한다.

5개 조건은 서로 보상되지 않는다. 품질 점수가 높아도 다른 조건의 미충족을 대신할 수 없다. 결정 게이트는 각 조건의 충족 여부를 확인하며 품질 수치를 다시 산출하지 않는다.

품질 등급은 가중치를 적용하지 않은 7개 차원 점수의 최솟값 Q_{min} 을 주판정값으로 사용한다.

$$Q_{min} = \min_i D_i \quad (1)$$

식 (1)에 따라 7개 차원 가운데 가장 낮은 점수가 품질 등급의 기준이 된다. 평균으로 등급을 판정하면 특정 차원의 낮은 점수가 다른 차원의 높은 점수로 상쇄될 수 있다. 7개 차원의 가중 평균 Q_{avg} 는 품질 분포를 확인하는 보조 지표로 사용하며 등급 판정에는 적용하지 않는다.

품질 등급은 Q_{min} 을 기준으로 5단계로 구분한다. Tier 1 Bronze는 θ_{min} 이상 0.70 미만, Tier 2 Silver는 0.70 이상 0.85 미만, Tier 3 Gold는 0.85 이상 0.95 미만, Tier 4 Platinum은 0.95 이상이다. 후보 등급을 정한 후 차원별 통과선을 적용하여 모든 대상 차원이 충족하는 가장 높은 등급을 최종 등급으로 확정한다. 기계 관독성의 면제 불가 조건을 충족하지 못하거나 Q_{min} 이 θ_{min} 에 미달하면 Tier 0 Unqualified로 판정하고 Quality-Ready 전이를 차단한다.

G2는 요구사항의 충족 여부에 따라 PASS와 FAIL로 판정한다. 권고사항은 출처·이력 완전성, 갱신 이력 문서화, 권고 기술 메타데이터의 3종으로 한정하며, 미충족은 판정값을 변경하지 않고 경고로 기록한다. 따라서 요구사항을 모두 충족한 데이터셋은 경고가 있더라도 Quality-Ready로 전이하고 G3의 판정 대상이 된다.

G3(목적 판정). G3는 Quality-Ready 데이터셋이 특정 활용 목적의 요건을 충족하는지 판정한다. 목적 프로파일의 운용 요건과 Layer 3 필수 메타데이터를 확인하고 목적 등급(P-Tier)을 판정한다. 해당 목적이 최소 등급 이상이면 Purpose-Ready 전이가 성립한다. 목적 등급과 전이 조건은 5-1절에서 정의한다.

G3는 활용 목적 유형마다 독립적으로 판정한다. 하나의 데이터셋이 특정 목적의 G3를 통과하지 못하더라도 다른 목적의 판정에는 영향을 미치지 않는다. 등록된 프로파일이 없는 활용 목적은 목적 등급을 판정하지 않는다. G3를 통과하면 판정 결과와 근거를 포함한 Purpose-Ready Pack을 생성한다. Pack의 구조와 무결성은 교환 단계에서 검증한다.

4-3 상태 속성과 갱신 규칙

모델은 상태 변화를 진입, 정방향 전이, 수동 무효화, 금지, Stale 진입, Stale 회복·종결의 여섯 유형으로 구분한다. 전체 전이와 갱신 규칙은 표 2에 요약하고 있다. Quality-Ready까지의 상태 전이는 데이터셋 단위로 적용하고,

Purpose-Ready 전이는 활용 목적 유형별로 적용한다.

Discoverable에서 Purpose-Ready로 직접 전이하는 경로는 허용하지 않는다. Stale은 Quality-Ready와 Purpose-Ready에만 적용한다. Raw와 Discoverable은 아직 등급이 부여되지 않은 상태이므로 판정 효력을 정지하는 Stale의 대상이 아니다. Quality-Ready와 Purpose-Ready는 판정 단위가 다르므로 Stale의 적용 범위도 다르다. 데이터셋의 Quality-Ready 판정이 만료되면 해당 판정을 전제로 성립한 목적별 Purpose-Ready 판정도 함께 효력을 잃는다. 반면 특정 활용 목적의 Purpose-Ready 판정만 만료되면 데이터셋의 Quality-Ready 상태는 유지된다.

표 2에서 정의한 3가지 원인 가운데 하나가 발생하면 파이프라인 운영자가 해당 원인의 발생 여부를 확인하고 Stale로 전환한다. 직전 상태에서 부여한 등급은 삭제하지 않고 효력을 정지한다. Pack의 운영 상태는 Stale 전이에 따라 변경하지 않는다. 구조 오류와 품질 저하 신호는 Stale 전이로 처리하지 않고 해당 단계의 재판정과 재실행 여부를 결정한다.

Stale 상태의 회복은 원인이 영향을 미친 범위에 따라 두 경로로 구분한다. 메타데이터나 판정 근거를 보정하여 유효성을 회복할 수 있으면 재검증을 수행한다. Quality-Ready 판정의 유효성을 회복하려면 단계 2를 다시 실행하고 G2를 통과해야 하며, 특정 활용 목적의 Purpose-Ready 판정만 만료된 경우에는 단계 3을 다시 실행하고 해당 목적의 G3를 통과해야 한다. 데이터 본문이 영향을 받았으면 단계 1부터 다시 실행하고 데이터셋 버전을 증가시킨다. 두 경로 모두 새로운 Pack을 생성하고 이전 Pack을 이전 버전으로 참조한다.

표 2의 수동 무효화는 발행자의 선언에 따라 Quality-Ready 판정을 무효화하고 Discoverable 상태로 전환하는 규칙이다. 이 규칙은 본 연구에서 제한한 것으로, 표준화 논의에는 대응하는 역방향 전이가 규정되어 있지 않다.

모든 상태 전이는 변경 주체의 식별자, 변경 일시, 결정 게이트의 판정 결과, 전이 사유를 PROV-O 온톨로지[30]에 따라 기록한다. 사용한 도구의 명칭과 버전은 활동의 실행 환경에 기록한다. 이 기록은 데이터셋의 상태 전이와 활용 목적별 Purpose-Ready 판정 이력을 추적하는 근거가 된다. 갱신 주기는 $dct:accrualPeriodicity$ 의 빈도 코드를 지속시간으로 환산하여 적용한다.

4-4 적용 범위

형식화의 범위는 준비 상태, 결정 게이트, 전이 규칙의 구조로 한정한다. 품질, 메타데이터, 신뢰·윤리, 판정 근거, 계보의 요구 조건을 구분하고, 품질 등급은 무가중 최솟값을 주판정값으로 사용한다. 요구사항과 권고사항을 구분하여 권고사항의 미충족은 경고로 기록하고 이후 판정에도 승계한다. 판정 구조와 구체적인 임계값은 분리하여 정의한다.

임계값은 파라미터로 정의한다. 결정 게이트는 임계값 θ 를 매개변수로 갖는 $G(\theta)$ 로 형식화하며, θ_{min} , θ_7 , θ_3 , 차원별 t_i ,

Q-Tier 등급 경계, Stale 계수 α 의 구체값은 본 연구의 주장 범위에 포함하지 않는다. 게이트의 구조는 논리적으로 정의할 수 있지만 임계값의 보편적 타당성은 실측을 거쳐 검증해야 한다. 영국 GDS·DSIT 가이드라인도 준비도 점수를 4점 척도의 제안으로 제시하며, 단일한 보편 기준으로 환원할 수 없음을 명시한다[10].

6장의 적용 사례는 표준화 논의에서 예시로 제시한 $\theta_{min} = 0.50$, $\theta_7 = 0.80$, $\theta_3 = 0.65$, $\alpha = 1.5$ 를 사용한다. 이 수치는 모델의 작동을 확인하기 위한 예시 기준값이며, 적용 결과가 임계값 자체의 타당성을 입증하지는 않는다. 임계값의 확장과 경험적 보정은 후속 실증 연구의 과제로 남긴다.

표준화 논의는 이러한 구조를 임계 프로파일로 구현한다. 임계 프로파일은 등급 경계와 차원별 통과선을 별도의 규정 집합으로 관리하고 버전을 부여한다. 활성 상태의 공식 프로파일이 없으면 예시 수치로 산출한 결과는 잠정 결과로 표기한다. 임계값을 파라미터로 두는 $G(\theta)$ 는 이러한 임계 프로파일 구조에 대응한다.

진단 범위는 진단 성숙도로 구분한다. Quality-Ready 상태를 공식적으로 공시(public claim)하려면 해당 데이터 유형의 필수 지표를 모두 측정한 DM-2를 충족해야 한다. 7개 차원의 점수를 모두 산출하더라도 각 차원에 필요한 필수 지표를 모두 측정하지 않았다면 DM-2로 판정하지 않는다. 제3자 적합성 인증은 진단 성숙도와 별개의 절차다. 표준화 논의는 인증 절차와 인증 마크를 시리즈의 범위에서 제외하고 운영 지침에 위임한다. 본 연구도 제3자 적합성 인증을 연구 범위에 포함하지 않는다.

준비 상태 모델은 데이터셋의 현재 준비 상태를 판정한다. Quality-Ready까지의 상태는 데이터셋 단위로 판정하고, Purpose-Ready는 활용 목적 유형마다 독립적으로 판정한다. 따라서 동일한 Quality-Ready 데이터셋도 활용 목적에 따라 Purpose-Ready 판정과 목적 등급이 다를 수 있다.

V. 목적 준비도 벡터(RQ2에 대한 응답)

5-1 목적 준비도 벡터(PRV) 정의

단일 품질 점수만으로는 특정 활용 목적에 대한 데이터의 준비 수준을 판정하기 어렵다. 목적 준비도는 데이터셋 버전 D_v 와 활용 목적 p 의 관계 $R(D_v, p)$ 로 표현하며, p 는 표준화 논의에서 관리하는 활용 목적 유형이다. $R(D_v, p)$ 는 활용 목적별 P-Tier를 결정하고 Purpose-Ready 전이 여부를 판정하는 데 사용한다.

P-Tier는 P-Tier 1, P-Tier 2, P-Tier 3의 3개 등급이며 $1 < 2 < 3$ 의 순서 관계를 갖는다. 등급을 순서값으로 변환한 $tier(R)$ 를 판정에 사용한다. 활용 목적을 적용할 수 없거나 해당 목적의 요건을 충족하지 못하면 P-Tier를 부여하지 않

고 목적 평가 상태와 사유 코드를 기록하며, $tier(R)$ 는 0으로 처리한다. N/A와 Not-Ready는 목적 평가 상태를 나타내며 P-Tier에 포함하지 않는다.

활용 목적 p 가 요구하는 필수 입력 구조가 데이터 유형에 존재하지 않으면 전제 미충족으로 기록하고 N/A로 표기한다. DL의 비정형 본문, RAG의 청킹 가능한 텍스트, FineTuning의 instruction-response 페어가 그러한 필수 입력 구조에 해당한다. 필수 입력 구조가 존재하지만 목적 프로파일의 임계값을 충족하지 못하면 Not-Ready로 표기한다. 따라서 적용할 수 없는 목적과 평가 결과 요건을 충족하지 못한 목적을 구분하여 기록한다.

G3는 활용 목적별 P-Tier를 기준으로 Purpose-Ready 전이를 판정한다. G3의 판정 조건은 다음과 같다.

$$G3(D_v, p) = PASS \Leftrightarrow S(D_v) = Quality-Ready \wedge tier(R(D_v, p)) \geq 1 \quad (2)$$

G3가 PASS이면 해당 활용 목적에 대한 데이터셋의 준비 상태를 Purpose-Ready로 판정한다. 활용 목적별 판정 결과는 목적 유형별로 독립하여 기록한다.

목적 준비도 벡터(PRV)는 판정을 완료한 활용 목적과 그 목적 준비도의 쌍을 모은 표기다. PRV는 고정된 길이를 갖지 않고 활용 목적 유형으로 색인하며, 판정을 완료한 목적만 포함한다. 활용 목적 유형이 추가되거나 통합되어도 해당 목적의 판정 결과를 추가하거나 조정할 수 있으므로 PRV의 형식은 유지된다. Q-Tier와 P-Tier는 서로 다른 판정값이다. Q-Tier는 모든 목적 프로파일의 선행조건이 되는 공통 기반 적격 수준(baseline qualification)을 나타내고, P-Tier는 특정 활용 목적에 대한 준비 수준을 나타낸다. P-Tier는 Quality-Ready 판정을 전제로 목적별로 산출하며, 두 등급은 구분하여 기록한다.

Q-Tier는 Tier 0 Unqualified부터 Tier 4 Platinum까지 5개 등급으로 구분한다. Bronze, Silver, Gold의 명칭은 GDS·DSIT(2026)의 메달리온 구조에서 차용하였으나[10], 등급의 의미는 다르다. 메달리온 구조가 데이터의 가공 단계를 구분하는 반면, Q-Tier는 품질 지수와 차원별 통과선을 적용하여 판정한 공통 품질 수준을 나타낸다.

상세 카탈로그, 적합성 선언, 자동 에이전트 인터페이스는 품질 등급과 목적 등급을 구분하여 제공한다. 단일 등급만 표시해야 할 경우 Q-Tier를 상한으로 P-Tier를 보수적으로 표시하는 규칙을 확장 항목으로 제안한다. 이 규칙은 Q-Tier와 P-Tier의 판정값을 변경하지 않고 표시 단계에만 적용한다. 구체적인 표기 방식과 사용자 이해도는 후속 실증 연구에서 평가한다.

표 3. 목적 유형별 핵심 운용 기준(기준값은 예시 기본값)

Table 3. Key operational criteria by purpose type (criteria are illustrative defaults)

Purpose type	Key operational criteria	Key metadata standards
ML-DL	Label conformance ≥ 0.95 ; class balance; feature completeness ≥ 0.90 ; label consistency (Cohen's κ)	Croissant 1.0
RAG	Chunkability; maximum chunk tokens; embedding-model compatibility	aird-rag profile
KG	Entity linking ratio ≥ 0.80 ; complete ontology mapping; identifier consistency	VoID + SHACL, SPARQL
Stats	Multidimensional cube structure; complete code-list linkage; explicit aggregation rules	SDMX 2.1
FineTuning	Instruction-response pair format; DPO data structure; domain vocabulary	aird-finetuning profile

5-2 PRV의 산출과 활용

각 활용 목적은 목적 유형별 운용 기준에 따라 판정한다. 현재 등재된 6개 유형의 핵심 기준과 예시값은 표 3과 같다.

세부 임계값과 P-Tier 등급 경계는 후속 실증과 표준화 과정에서 확정한다. 6개 유형은 하나의 데이터셋에 중첩하여 적용할 수 있으며, 각 목적의 준비도를 독립적으로 판정한다. 새로운 목적 유형은 해당 목적의 판정 기준을 추가하여 확장할 수 있으며, 표준화 논의의 등록과 활성화 절차를 거쳐 제어 어휘로 사용한다.

목적별 준비도는 표 3의 운용 기준에 따라 규칙 기반(rule-based) 방법이나 ML 보조 방법으로 측정한다. ML 목적은 레이블 완전성, 클래스 균형 비율, 피쳐 완전성을 진단 프로브로 측정하고, KG 목적은 엔티티 연결 비율 등을 SPARQL 질의로 측정한다. 원본 데이터의 변경으로 판정의 유효성이 만료되면 Stale로 전이하고 목적별 준비도를 다시 판정한다. 재판정이 필요한 조건은 자동으로 탐지하며, 운영자의 명시적 승인 후 평가를 다시 실행하고 결정 게이트를 재통과해야 한다.

마니페스트는 목적별 준비도와 품질 등급을 구분하여 기록한다. 목적별 준비도는 활용 목적 유형별 P-Tier와 평가 상태, 적용한 목적 프로파일의 식별자와 버전, 임계 프로파일의 식별 정보를 기록한다. 품질 등급은 등급명과 등급 번호, 데이터 유형을 병기한 라벨 형식으로 기록한다. 항목명과 값에 사용하는 어휘는 표준화 논의에서 규정한다.

목적별 준비도는 데이터 탐색, AI 에이전트의 데이터 선택, 목적별 Pack 생성의 우선순위 결정에 활용할 수 있다. 데이터 탐색은 P-Tier 2 이상과 같은 조건을 적용할 수 있으며, AI 에이전트는 목적 등급과 품질 등급을 구분하여 판독한다.

목적별 준비도 판정만으로 데이터의 자동 적재를 허용하지는 않는다. 운용 계층을 갖추지 않은 Pack은 자동 적재 대상에서 제외하며, 레지스트리는 규정된 조건의 충족 여부를 판정하여 허용 여부를 결정한다. 목적별 준비도는 이러한 운용 조건과 함께 AI가 데이터를 선택하고 활용하기 위한 판정 정보로 사용한다.

VI. 적용 시연: 전국등록공장현황 공공데이터

6-1 사례 설정

전국등록공장현황은 전국 산업단지 내 공장의 등록번호, 업종코드, 행정구역, 고용인원, 생산가동률을 포함하는 약 125만 건 규모의 정형(STRUCT) 공공데이터다. 이 데이터셋은 행정구역코드와 같은 식별 정보를 이용하여 산업단지, 고용통계 등 여러 데이터셋과 연계할 수 있다.

전국등록공장현황은 정형 데이터의 품질 진단뿐 아니라 서로 다른 활용 목적에 따른 준비도의 차이를 확인할 수 있어 적용 사례로 선정하였다. 공장의 위치와 산업 특성을 이용한 예측, 다른 산업 데이터와의 연계, 지역별 통계 분석처럼 동일한 데이터셋을 여러 목적으로 활용할 수 있다. 또한 원시 XLSX와 전처리한 CSV를 비교할 수 있어 데이터의 내용이 유사하더라도 구조와 기계 판독성에 따라 준비 상태가 달라지는 과정을 확인할 수 있다. 준비 상태 모델, 7차원 품질 진단, 결정 게이트 조건을 원시 XLSX와 전처리한 CSV에 적용하여 판정 결과를 비교한다. 전처리한 CSV는 UTF-8 인코딩과 명시적 스키마를 갖추고 있으며, 원시 XLSX는 다중 시트와 병합 헤더를 포함하고 명시적 스키마가 없다. 활용 목적은 ML 기반 입지 예측, KG 기반 산업 데이터 연계, Stats 기반 지역별 산업 통계로 설정한다. 세 활용 목적은 동일한 데이터셋에 서로 다른 목적별 요건을 적용했을 때 P-Tier와 Purpose-Ready 판정이 어떻게 달라지는지를 확인하기 위한 시나리오다. 공공데이터는 데이터 접근성과 정책 연계 가능성을 고려하여 사례 대상으로 선택하였으며, 준비 상태 모델의 적용 범위는 특정 도메인으로 한정하지 않는다.

7차원 품질 점수는 실측값이 아니라 판정 과정을 확인하기 위해 설정한 시나리오의 예시값이다. 따라서 사례는 모델의 경험적 타당성을 검증하는 실증 분석이 아니라, 주어진 조건에서 준비 상태와 등급이 어떻게 판정되는지를 확인하는 적용 예시(worked example)다.

6-2 상태 전이 적용 결과: 전처리 후 CSV(PASS 케이스)

사례의 단계별 처리와 판정 과정은 그림 2와 같다. 단계 1은 원시 XLSX를 CSV로 변환하고 인코딩을 UTF-8로 정규

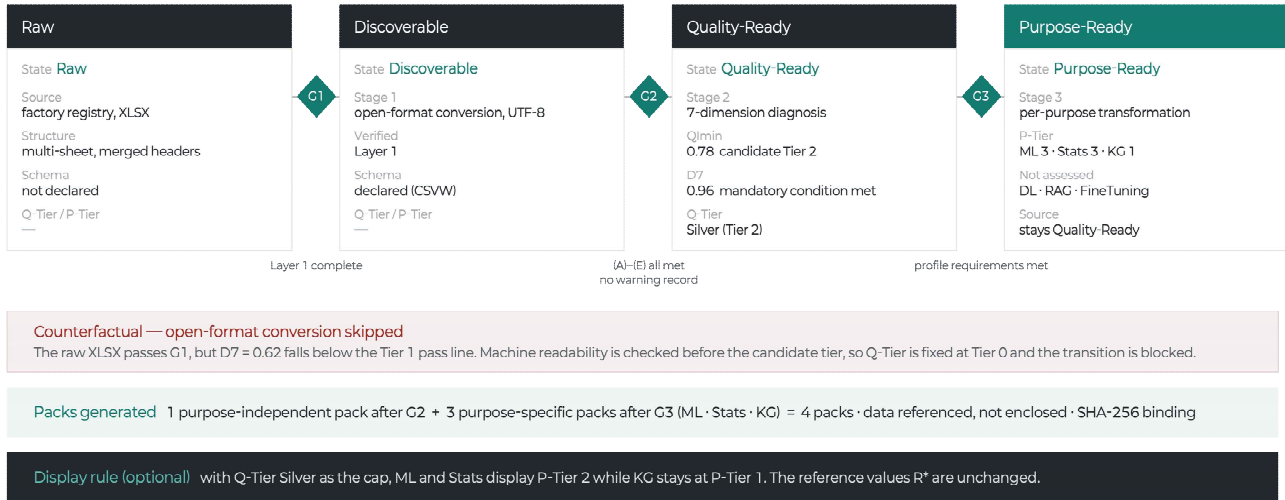


그림 2. 공장현황 데이터의 판정 경로

Fig. 2. Decision path for the factory dataset

화한 뒤 Layer 1 필수 항목을 생성한다. 처리 결과가 G1 조건을 충족하면 데이터셋은 Discoverable 상태로 전이하고, 해당 상태와 판정 결과를 마니페스트에 기록한다. 단계 2는 7개 품질 차원을 진단한다. 시나리오의 차원별 점수는 표 4의 CSV 열과 같다. 기계 판독성 D7은 개방 포맷, 명시적 스키마, CSVW[31] 기반 식별 가능성을 정량화하며 이 사례에서는 0.96으로 설정하였다.

G2는 품질, 메타데이터, 신뢰·윤리, 판정 근거, 계보의 요구 조건을 확인하고 권고사항의 미충족 여부를 함께 판정한다. 표 4의 CSV 열과 시나리오 가정값은 모든 요구사항과 권고사항을 충족한다. 따라서 G2는 경고가 없는 PASS로 판정되고 데이터셋은 Quality-Ready 상태로 전이한다. 새로운 상태와 판정 결과는 마니페스트에 기록한다.

품질 등급은 두 단계로 확정한다. 최소 차원 점수 Q_{min} = 0.78은 Tier 2 구간에 해당하므로 후보 등급은 Tier 2다. Tier 2의 차원별 통과선을 적용하면 모든 대상 차원이 기준을 충족하므로 최종 등급은 Tier 2 Silver로 확정된다. 가장 낮은 차원은 적시성 D4 = 0.78이며, 다른 차원의 높은 점수

가 적시성의 낮은 점수를 보상하지 않으므로 최솟값 0.78이 등급 판정에 적용된다.

단계 3은 활용 목적별 프로파일을 적용하여 목적 준비도를 판정한다. ML 프로파일은 Train 70%·Valid 15%·Test 15%의 stratified 분할과 Croissant 1.0 메타데이터를 구성한다. 레이블 완전성 0.97, 클래스 균형 2.3:1, 피쳐 완전성 0.92의 가정값이 요건을 충족하므로 ML 목적은 P-Tier 3으로 판정한다. Stats 프로파일은 시점·지역·업종의 SDMX 다차원 구조, 코드 목록 연계, 집계 규칙을 갖추어 P-Tier 3으로 판정한다. KG 프로파일은 공장 식별자로 엔티티를 추출하고 RDF 트리플을 생성한다. 엔티티 연결 비율 0.82는 하한을 충족하지만 한국표준산업분류와 행정구역의 온톨로지 매핑이 부분적으로 완료된 상태이므로 P-Tier 1로 판정한다.

DL, RAG, FineTuning은 P-Tier를 판정하지 않는다. DL은 비정형 본문 구조가 없어 전제를 충족하지 못하고, RAG와 FineTuning은 필요한 입력 구조가 없어 판정 대상에서 제외한다. 세 목적에는 P-Tier를 부여하지 않고 목적 평가 상태와 사유 코드를 기록한다.

표 4. 공장현황 데이터의 7차원 품질 진단: 전처리 후 CSV(PASS)와 원시 XLSX(반사실 FAIL) 비교

Table 4. Seven-dimension quality diagnosis of the factory dataset: preprocessed CSV (PASS) versus raw XLSX (counterfactual FAIL)

Dimension	Mandatory	$D_{\text{—}}$ CSV (PASS)	$D_{\text{—}}$ Raw XLSX (FAIL)
D1 Completeness	—	0.92	0.61
D2 Consistency	—	0.85	0.54
D3 Accuracy	—	0.91	0.72
D4 Timeliness	—	0.78	0.80
D5 Validity	—	0.89	0.58
D6 Uniqueness	—	0.88	0.76
D7 Machine-readability	✓	0.96	0.62 (gate failed)
Q_{min} (primary)	—	0.78 → candidate T2 → Silver (Tier 2)	0.54 → Tier 0 (by D7 gate)
Q_{avg} (reference, weighted)	—	0.89	0.66

식 (2)의 조건에 따라 ML, Stats, KG 목적의 G3는 각각 PASS로 판정되며, 세 활용 목적은 Purpose-Ready에 도달한다. 데이터셋의 Quality-Ready 판정은 유지된다. G2를 통과하면 Quality-Ready 팩 1건을 생성하고, G3를 통과한 ML, Stats, KG에는 각각 Purpose-Ready 팩을 생성한다. 따라서 사례에서 생성되는 AIRD Pack은 총 4건이다. 목적별 판정과 품질 등급을 마니페스트에 기록하고 Pack 구조와 참조 파일의 해시가 유효하다고 가정하면 등록 절차의 검증을 거쳐 교환 대상이 된다.

6-3 G2 FAIL 케이스: 원시 XLSX

동일한 전국등록공장현황 데이터를 공공데이터포털에서 제공하는 원본 XLSX 형태로 처리한다고 가정한 결과는 표 4의 XLSX 열과 같다. 진단 파이프라인은 단계 1에서 개방 형식으로 변환하므로 원시 XLSX가 그대로 G2에 도달하지 않는다. 이 사례는 개방 형식 변환을 생략했을 때 상태 전이가 어느 단계에서 차단되는지를 확인하기 위한 반사실(counterfactual) 시나리오다.

XLSX는 사양이 공개된 기계 구분분석 가능 포맷(ISO/IEC 29500, OOXML)이므로 발견 계층의 필수 항목을 갖추면 G1을 통과하여 Discoverable 상태로 전이할 수 있다. 그러나 사례의 원시 XLSX는 다중 시트, 병합 셀, 수식을 포함하여 논리 구조의 해석이 필요하고 명시적 스키마가 결합되어 있지 않다. 구조 명료성과 스키마 명시성은 G1의 진입 조건이 아니라 기계 판독성 D7에서 평가한다. D7 = 0.62는 XLSX 포맷 자체에 일률적으로 부여한 점수가 아니라 해당 파일의 구조적 문제를 반영하여 설정한 예시값이다.

G2는 기계 판독성을 후보 등급 산정보다 먼저 확인한다. Q_{min} = 0.54는 Tier 1 구간에 해당하지만 D7 = 0.62가 기계 판독성의 Tier 1 통과선에 미달한다. D7은 면제할 수 없는 조건이므로 다른 품질 차원의 점수와 관계없이 Tier 0 Unqualified로 판정하고 Quality-Ready 전이를 차단한다. 이 경우 후보 등급은 산출하지 않는다.

Q_{avg} = 0.66과 Q_{min} = 0.54가 Tier 1의 최소 기준을 충족하더라도 D7이 면제 불가 조건을 충족하지 못하면 G2는 FAIL로 판정한다. 평균이나 다른 차원의 높은 점수는 D7의

미달을 보상하지 않는다. 동일한 데이터셋에 설정한 두 시나리오는 원시 XLSX에서 Tier 0, 전처리한 CSV에서 Tier 2로 서로 다른 판정 결과를 산출한다. 모델은 데이터셋의 준비 상태를 고정된 속성이 아니라 진단과 변환에 따라 달라지는 상태로 판정한다.

6-4 프레임워크 유형별 판정 정보의 비교

동일한 입력에 기존 프레임워크와 본 연구의 모델을 적용했을 때 산출되는 판정 정보의 범위를 표 5와 같이 비교하였다. 표 5의 세 기존 방식은 데이터셋의 준비 상태, 활용 목적별 준비도, 상태 전이의 차단 지점을 하나의 판정 체계에서 함께 산출하지 않는다. 단일 품질 점수 방식은 여러 품질 차원을 하나의 값으로 집계하므로 특정 차원의 낮은 점수가 다른 차원의 높은 점수로 상쇄될 수 있으며, 활용 목적별 준비 수준의 차이도 구분하지 않는다.

본 연구의 모델은 동일한 입력에 대하여 준비 상태, Q-Tier, 활용 목적별 P-Tier, 상태 전이의 차단 지점을 구분하여 산출한다. 이 비교는 각 프레임워크를 동일한 조건으로 구현하여 재실행한 성능 비교가 아니라, 각 프레임워크가 정의하는 산출 정보의 범위를 사례의 판정 항목에 대응하여 비교한 것이다. 따라서 비교 대상은 측정 정확도가 아니라 각 방식이 산출하는 판정 정보의 범위이며, 측정 정확도의 정량적 비교는 동일 데이터셋의 실측과 하류 AI 성능 관측을 포함하는 후속 연구가 필요하다.

6-5 논의

사례의 가정값을 적용한 결과, 제안한 판정 구조에서 최솟값에 따른 품질 판정, 품질 등급과 목적 등급의 분리, 차원별 통과선에 따른 등급 조정이 각각 적용되었다.

성공 경로에서 가중 평균은 Gold 구간에 해당하지만, 주판정값인 최소 차원 점수에 따라 Q-Tier는 Silver로 판정되었다. 실패 경로는 최소 차원 점수가 Tier 1 구간에 해당하더라도 기계 판독성이 면제 불가 조건에 미달하여 G2가 FAIL로 판정되었다. 두 결과는 특정 차원의 미달을 다른 차원의 높은 점수로 보상하지 않는 판정 원칙이 적용된 사례다.

표 5. 동일 사례에서 프레임워크 유형별로 제공 가능한 판정 정보의 비교

Table 5. Comparison of the decision information that each type of framework can provide for the same case

Framework type	Decision information provided	Information not provided
Self-assessment checklist	Per-item compliance status	Readiness state; whether the data can be consumed for a given purpose
Single quality score	An aggregate score over multiple quality items	Whether a minimum dimension blocks the transition; differences across purposes
ML metadata format	Machine-readable description of the dataset	Tier decision; comparison across purposes
This study	State (Quality-Ready), tier (Tier 2), per-purpose decisions, and the blocking point (D7 in the counterfactual case)	—

품질 등급과 목적 등급은 서로 다른 기준에 따라 판정된다. 동일한 데이터셋에서 ML과 Stats는 P-Tier 3, KG는 온톨로지 매핑이 일부만 충족되어 P-Tier 1로 판정되었다. 적시성을 보장하여 Q-Tier가 Gold로 높아지더라도 KG의 목적 요건이 달라지지 않으면 P-Tier는 바뀌지 않는다. 품질 보장과 목적별 요건의 보장은 각각 Q-Tier와 P-Tier의 판정 기준에 따라 구분하여 적용한다.

사례는 단일 데이터셋을 대상으로 하며, 7개 차원의 점수도 실측값이 아닌 예시값을 사용하였다. 따라서 확인한 범위는 주어진 조건에서 모델이 산출하는 준비 상태와 등급, 게이트의 판정 결과로 한정된다. 목적별 준비도의 자동 산출, 진단 도구 간 정합성, 시간 경과에 따른 상태 전이의 실제 운용은 검증하지 않았다. 다수 데이터셋에 대한 자동 진단, 표기 방식에 대한 사용자 이해도, 진단 도구에 따른 결과 변동성도 연구 범위에 포함하지 않았다.

VII. 표준화와 정책 연계 방향(RQ3에 대한 응답)

형식 모델을 표준화 기능 계층으로 구체화하고 국제 표준과의 정합 관계를 정리한다. 공공데이터는 대표 적용 분야로 정책 연계 방향을 제시하되, 정책 적용의 실증 평가는 연구 범위에 포함하지 않는다.

7-1 표준화 기능 계층과 국제 표준 정합

AI Ready Data 구현에 필요한 기능은 역할에 따라 계층으로 구분한다. 표 6은 각 기능 계층과 기반 국제 표준의 대응 관계를 정리한다.

기능 계층은 서로 다른 역할을 분리하고, 동일한 수치와 기준을 여러 계층에서 중복하여 정의하지 않는다. 기존 국제 표준으로 표현할 수 있는 항목은 해당 표준을 적용하고, 표현할

수 없는 항목에 한하여 독자 어휘를 정의한다.

형식 모델의 일부는 표준화 논의에서 규범으로 구체화되었다. 신뢰·윤리 기록 3종을 전이 요건으로 적용하고, 최솟값을 품질 등급의 주관정값으로 사용하며, 기계 판독성을 독립된 품질 차원이자 면제 불가 조건으로 규정한다. 임계값은 별도의 규정 집합으로 관리하고, 목적별 판정과 배포 단위의 생성을 분리한다.

표준화 논의에 반영되지 않은 내용으로 G1·G2·G3의 형식 구조와 요구 조건 집합의 분리, 활용 목적별 독립 판정, 공통 준비 상태와 활용 목적별 준비 상태의 분리, 평가 미달과 미평가의 구분, 권고사항 미충족 경고의 소비자 노출을 제안한다. 임계 파라미터의 구체값은 후속 실증 연구에서 보정하고, 그 결과는 표준화 과정에서 임계값을 확정하는 근거로 활용할 수 있다.

3계층 메타데이터는 FAIR 원칙과 대응한다. FAIR 원칙이 데이터의 발견, 접근, 상호운용, 재사용을 위한 원칙을 제시한다면, 본 연구의 모델은 관련 요건의 충족 여부를 준비 상태 전이와 결정 게이트에 연결하여 판정한다.

품질 측정 계층은 ISO/IEC 25012에서 선택하여 재구성한 6개 차원에 ISO/IEC 25024:2015의 측정 함수를 결합하고 D7 기계 판독성을 추가한다[2], [28]. Layer 1은 CSVW를 컬럼 스키마 기술의 기준으로 사용하고, Layer 2는 RAI 어휘를 신뢰·윤리 기록 3종의 기준으로 사용한다. 컬럼 스키마의 측정 결과인 스키마 적합률은 Layer 2에 기록한다. KG 프로파일은 SHACL[32]로 무결성 제약을 표현한다. 변환·거버넌스 계층과 교환 계층에 적용하는 기반 표준은 표 6에서 요약하고 있다.

7-2 공공데이터 정책 연계 방향

공공데이터는 도메인 중립적 기능 계층의 적용 가능성을 설명하는 대표 사례로 사용한다. 제안 체계는 AI 활용을 위한

표 6. 표준화 대상 기능 계층과 기반 국제 표준의 매핑

Table 6. Functional layers for standardization and their mapping to international standards

Functional layer	Core question	Main components	Base international standards
Discovery	How is data found and accessed?	Layer 1 metadata; rights and licenses; column schema	DCAT v3, Dublin Core, SPDX, CSVW
Quality & Grading	What is measured and decided?	7-dimension QI; QI_{min} primary decision; Q-Tier; mandatory gates; DM	ISO/IEC 25012:2008, ISO/IEC 25024:2015, ISO/IEC 5259, W3C DQV
Metadata & Profile	How is it recorded and represented?	Readiness state model; Layers 1/2/3; purpose-type profiles	Croissant 1.0, VoID + SHACL/SPARQL, SDMX, aird-rag/aird-finetuning profiles
Transformation & Governance	How is it executed and operated?	3-stage pipeline; Gates 1/2/3; warning records; AIRD Pack; RAI documentation	PROV-O [30], RAI vocabulary, JCS (RFC 8785)
Exchange & Consumption	How is it discovered, exchanged, and consumed?	Pack exchange specification; registry; conditions for agent consumption	SHA-256, DCAT v3, PROV-O
Conformance	How is it certified?	Conformity assessment procedure; certification mark	— (future work)

판정 기준과 운용 요건을 추가하는 방식으로 연계할 수 있다.

Layer 3 운용 메타데이터는 기존 공공데이터 메타데이터 체계에 AI 도구가 데이터를 처리할 때 필요한 정보를 추가하는 방식으로 적용할 수 있다. Q-Tier는 Q_{min} 에 따른 주관정, 기계 판독성의 면제 불가 조건, 등급별 통과선과 강등 절차, 권고사항 미충족에 대한 경고 기록을 공공데이터 품질 관리에 적용하는 기준으로 활용할 수 있다. AIRD Pack은 품질과 목적별 준비도 판정 결과, 판정 근거, 메타데이터를 데이터와 함께 제공하는 배포 단위로 적용할 수 있다. 초기 적용 대상은 여러 데이터셋과 반복적으로 연계되는 허브 성격의 데이터셋을 고려할 수 있으며, 전국등록공장현황, 인구, 도로 데이터 등이 이에 해당한다.

정책 적용에 앞서 메타데이터 확장에 따른 표준 준수 비용, 기계 판독성의 면제 불가 조건이 기존 공개 데이터셋에 미치는 영향, AIRD Pack 적용에 필요한 기관의 업무와 비용을 평가해야 한다. 이러한 정책 효과와 적용 가능성의 평가는 본 연구의 범위에 포함하지 않는다.

판정 기준을 법적 의무로 적용하려면 법령과 하위 규정, 외부 표준의 관계를 별도로 검토해야 한다. 다만, RQ3는 AI Ready Data의 판정에 필요한 기능 계층과 국제 표준의 대응 관계를 제시하고, 공공데이터에 적용할 수 있는 정책 연계 방향을 제안하는 범위로 한정한다.

VIII. 결 론

본 연구는 AI Ready Data를 AI 처리 파이프라인이 자동으로 적재·검증·활용할 수 있는 동적 준비 상태로 정의하고, 준비 상태의 전이와 활용 목적별 준비도를 판정하는 모델을 제안하였다. 기존 접근이 데이터 준비에 필요한 요소를 제시하는 데 중점을 두었다면, 이 연구는 데이터가 언제 AI Ready 상태에 도달하고 그 판정이 언제까지 유효한지를 결정하는 기준을 구체화하였다.

연구 결과는 세 가지 연구 질문에 대응한다. RQ1은 준비 상태와 운영 활동을 분리하고, 결정 게이트를 통해 상태 전이를 판정하는 구조로 구체화하였다. 품질 검증을 통과한 데이터만 목적별 판정의 대상으로 허용하고, 품질 조건과 메타데이터·신뢰·윤리·판정 근거·정보 조건의 상호 보상을 제한하였다. 판정의 유효성이 만료된 경우 스테일 전이와 회복 규칙을 적용하여 준비 상태를 지속적으로 관리할 수 있도록 하였다.

RQ2는 활용 목적마다 준비도를 독립적으로 판정하는 체계로 구체화하였다. Q-Tier는 공통 품질 수준을 나타내고, P-Tier는 특정 활용 목적의 준비 수준을 나타내도록 분리하였다. 따라서 동일한 Quality-Ready 데이터셋도 활용 목적에 따라 서로 다른 P-Tier와 Purpose-Ready 판정을 가질 수 있다. 전국등록공장현황 적용 사례는 ML과 Stats에서 P-Tier 3, KG에서 P-Tier 1이 산출되는 과정을 통해 이러한 목적 의존적 판정 구조의 작동을 확인하였다.

RQ3은 모델을 개념·품질·메타데이터·파이프라인·교환의 표준화 기능 계층으로 구성하고 국제 표준과의 정합 관계를 정리하였다. 준비 상태, 목적별 준비도, 판정 결과와 근거를 함께 기록하고 전달하는 구조를 제시하여 AI Ready Data를 개념적 요구사항에서 기계가 판독할 수 있는 판정 체계로 구체화할 수 있는 기반을 마련하였다.

적용 사례는 예시값을 사용한 단일 데이터셋의 시연이며 모델의 경험적 타당성을 검증한 결과가 아니다. 임계값은 파라미터로 정의하였으며 구체값의 타당성은 검증하지 않았다. 공공데이터 정책 연계의 비용과 효과도 연구 범위에 포함하지 않았다. 후속 연구는 실제 데이터에 자동 진단 도구를 적용하여 판정 결과와 AI 활용 결과의 관계를 검증하고, 다양한 데이터 유형과 활용 목적에 대한 실측을 바탕으로 임계값의 타당성을 검토할 필요가 있다.

감사의 글

이 논문은 2026년도 중앙대학교 연구년 결과물로 제출됨 (This research was supported by the Chung-Ang University research grant in 2026).

참고문헌

- [1] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, pp. 1-15, May 2021. <https://doi.org/10.1145/3411764.3445518>
- [2] International Organization for Standardization, Software Engineering — Software Product Quality Requirements and Evaluation (SQuaRE) — Data Quality Model, ISO/IEC, Geneva, ISO/IEC 25012:2008, 2008.
- [3] Open Data Institute. A Framework for AI Ready Data [Internet]. Available: <https://theodi.org/insights/reports/a-framework-for-ai-ready-data/>.
- [4] U.S. Department of Commerce. Generative Artificial Intelligence and Open Data: Guidelines and Best Practices [Internet]. Available: <https://www.commerce.gov/news/blog/2025/01/generative-artificial-intelligence-and-open-data-guidelines-and-best-practices>.
- [5] OECD. Government at a Glance 2025 [Internet]. Available: https://www.oecd.org/en/publications/2025/06/government-at-a-glance-2025_70e14c6c.html.
- [6] N. Majithia, T. Carey-Wilson, E. Simperl, and N. Shadbolt, "An Actionable Framework for AI-Ready Data," *AI*

- Magazine*, Vol. 47, No. 1, e70054, 2026. <https://doi.org/10.1002/aaai.70054>
- [7] S. Verhulst, A. J. Zahuranec, and H. Chafetz, "Moving Toward the FAIR-R Principles: Advancing AI Ready Data," SSRN Scholarly Paper No. 5164337, March 2025. <https://doi.org/10.2139/ssrn.5164337>
- [8] Ministry of the Interior and Safety and National Information Society Agency, Guidelines for AI-Friendly Management of Public Data, 2026.
- [9] J. Lee and H. Kim, "Conceptualizing AI Ready Data and Analyzing South Korea's AI Ready Public Data Policy Framework," *Informatization Policy*, Vol. 33, No. 2, pp.67-86, 2026. <https://doi.org/10.22693/NIAIP.2026.33.2.067>
- [10] Government Digital Service and Department for Science, Innovation and Technology. Guidelines and Best Practices for Making Government Datasets Ready for AI [Internet]. Available: <https://www.gov.uk/government/publications/making-government-datasets-ready-for-ai/guidelines-and-best-practices-for-making-government-datasets-ready-for-ai>.
- [11] N. D. Lawrence, "Data Readiness Levels," arXiv:1705.02245, May 2017. <https://arxiv.org/abs/1705.02245>
- [12] L. A. Castelijns, Y. Maas, and J. Vanschoren, "The ABC of Data: A Classifying Framework for Data Readiness," in *Proceedings of the International Workshops of ECML PKDD 2019 on Machine Learning and Knowledge Discovery in Databases*, Würzburg, Germany, pp. 3-16, 2020. https://doi.org/10.1007/978-3-030-43823-4_1
- [13] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel, "Data Readiness Report," arXiv:2010.07213, October 2020. <https://arxiv.org/abs/2010.07213>
- [14] M. Ahangaran, H. Zhu, R. Li, L. Yin, J. Jang, A. P. Chaudhry, ... and V. B. Kolachalama, "DREAMER: A Computational Framework to Evaluate Readiness of Datasets for Machine Learning," *BMC Medical Informatics and Decision Making*, Vol. 24, No. 1, 152, June 2024. <https://doi.org/10.1186/s12911-024-02544-w>
- [15] K. Hiniduma, S. Byna, and J. L. Bez, "Data Readiness for AI: A 360-Degree Survey," *ACM Computing Surveys*, Vol. 57, No. 9, April 2025. <https://doi.org/10.1145/3722214>
- [16] W. Brewer, P. Widener, V. Anantharaj, F. Wang, T. Beck, A. Shankar, and S. Oral, "Data Readiness Pipeline Patterns for Scientific AI at Scale: Insights from Climate, Fusion, Life Sciences, and Materials," *AI Magazine*, Vol. 47, No. 1, e70056, 2026. <https://doi.org/10.1002/aaai.70056>
- [17] Big Data Value Association. AI-Ready Data Products [Internet]. Available: <https://bdva.eu/blog/ai-ready-data-products/>.
- [18] S. Al Manir, M. A. Levinson, J. Niestroy, C. Churas, N. C. Sheffield, B. Sullivan, ... and T. Clark, "The FAIRSCAPE AI-Readiness Framework for Biomedical Research," bioRxiv preprint 2024.12.23.629818, March 2026. <https://doi.org/10.1101/2024.12.23.629818>
- [19] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, ... and B. Mons, "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data*, Vol. 3, 160018, March 2016. <https://doi.org/10.1038/sdata.2016.18>
- [20] L. Schreiber, D. S. Katz, Y. Rao, and C. R. Kirkpatrick, "Introduction to Special Issue on FAIR Principles and Machine Learning, AI Readiness and AI Reproducibility," *AI Magazine*, Vol. 47, No. 1, e70063, 2026. <https://doi.org/10.1002/aaai.70063>
- [21] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," *Communications of the ACM*, Vol. 64, No. 12, pp. 86-92, 2021. <https://doi.org/10.1145/3458723>
- [22] M. Akhtar, O. Benjelloun, C. Conforti, L. Foschini, J. Giner-Miguel, P. Gijsbers, ... and L. Zhang, "Croissant: A Metadata Format for ML-Ready Datasets," arXiv:2403.19546, March 2024. <https://arxiv.org/abs/2403.19546>
- [23] World Wide Web Consortium. Data Catalog Vocabulary (DCAT) — Version 3 [Internet]. Available: <https://www.w3.org/TR/vocab-dcat-3/>.
- [24] World Wide Web Consortium. Data on the Web Best Practices: Data Quality Vocabulary [Internet]. Available: <https://www.w3.org/TR/vocab-dqv/>.
- [25] International Organization for Standardization, Artificial Intelligence — Data Quality for Analytics and Machine Learning (ML) — Part 2: Data Quality Measures, ISO/IEC, Geneva, ISO/IEC 5259-2:2024, 2024.
- [26] H. Chafetz, A. J. Zahuranec, and S. Verhulst, A Blueprint to Unlock New Data Commons for AI, Open Data Policy Lab (The GovLab), New York: NY, 2025.
- [27] UK Government. AI Opportunities Action Plan [Internet]. Available: <https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan>.
- [28] International Organization for Standardization, Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — Measurement of Data Quality, ISO/IEC, Geneva, ISO/IEC 25024:2015, 2015.
- [29] Internet Engineering Task Force, JSON Canonicalization Scheme (JCS), RFC 8785, June 2020. <https://doi.org/10.17487/RFC8785>
- [30] World Wide Web Consortium. PROV-O: The PROV

Ontology [Internet]. Available: <https://www.w3.org/TR/pr-ov-0/>.

[31] World Wide Web Consortium. Model for Tabular Data and Metadata on the Web [Internet]. Available: <https://www.w3.org/TR/tabular-data-model/>.

[32] World Wide Web Consortium. Shapes Constraint Language (SHACL) [Internet]. Available: <https://www.w3.org/TR/shacl/>.



김학래 (Haklae Kim)

2010년 : 아일랜드 국립대학교
(공학박사)

2004년 ~ 2009년: Digital Enterprise Research Institute,
Ireland

2009년 ~ 2016년: 삼성전자

2017년 ~ 2019년: 한국과학기술정보연구원

2019년 ~ 현 재: 중앙대학교 문헌정보학과 교수

※ 관심분야 : 지식그래프, 온톨로지, 인공지능 등