

AI READY DATA 2026 · SPECIAL SESSION

AI Ready Data

데이터를 AI 활용 가능 상태로 만드는 표준 체계

국립군산대학교

2026. 09. 03. (수) 16:45 – 17:05

발표자 : 김 장 원

목 차

01 설명은 있으나 확인 기준이 미흡하다

02 접근과 기준이 통일되어 있지 않다

03 AIRD 표준 체계 — 정의 · 어휘 · 형식 · 관리

문제 인식 · 첫째

01

「AI Ready 데이터」, 설명은 있으나 확인 기준이 미흡하다

가트너를 비롯해 「AI Ready 데이터」가 무엇인지에 대한 설명은 이미 나와 있다.
그러나 그 설명은 무엇을 지향해야 하는지를 말할 뿐,
그것을 갖추었는지 무엇으로 확인하는지는 말하지 않는다.

가트너는 AI Ready 데이터를 이렇게 설명한다

데이터는 특정 용도로 AI 모델을 학습하거나 구동하는 데 필요한 모든 패턴·오류·이상치·예기치 못한 출현을 포함하여, 그 활용 사례를 대표해야 한다.

AI 레디 데이터는 메타데이터의 가용성에 기반하여 데이터를 정합·자격 부여·거버넌스하는 프로세스이자 실무 체계이다.

※ 위 정의들은 방향성을 제시하는 선언적 수준에 머물러 있어, 실제 데이터가 이를 충족하였는지 판별할 측정 기준이 없다. 즉, 시장 조사 기관의 분석적 제안일 뿐 공식적인 검증 표준이 아니다.

63%

AI에 맞는 데이터 관리 체계를 갖추지 못하였거나 확신하지 못함
데이터 관리 리더 1,203명 대상 조사 · 2024. 7. ²

고품질 ≠ AI 레디

같은 데이터라도 AI 활용 용도에 따라 남겨야 하는 값이 다르기 때문이며,
전통적 품질 관리는 오히려 학습에 필요한 오류·이상치를 제거해 버린다¹

¹ Gartner, "AI-Ready Data Essentials to Capture AI Value" ² Gartner, "Lack of AI-Ready Data Puts AI Projects at Risk" (2025. 2. 26.)

가트너의 설명을 네 구성 요소로 나누어 본다

메타데이터의 가용성

전제. 판단에 필요한 설명 정보가 활용 가능한 상태로 존재하는가 — 나머지 셋의 전제

정합

이 데이터가 **활용하려는 용도에 부합하는가** — 용도가 먼저 정해져야 답할 수 있다

자격 부여

현재 **활용 가능한 상태인지 판별한다** — 요건을 충족하였는지, 그 판단을 얼마나 신뢰할 수 있는지

거버넌스

그 상태를 **유지하고 관리한다** — 원본이 변경되면 앞의 판단은 어떻게 되는가

※ 무엇을 갖추어야 하는지는 이 네 요소가 제시한다. 그러나 무엇을 측정하는지 · 어떤 방법으로 측정하는지 · 어느 수준이면 충족인지 · 누가 확인하는지는 제시하지 않는다.

전제가 성립하지 않으면 나머지 셋은 수행되더라도 그 근거가 남지 않는다

「메타데이터의 가용성」은 무엇을 뜻하는가

여기서 가용성은 시스템 가동률이 아니라, **판단에 필요한 설명 정보가 실제로 활용 가능한 상태로 존재하는가**를 뜻한다. 네 가지 조건으로 나뉜다.

1 존재하는가 이 항목이 무엇을 의미하는지 문서에 기재되어 있는가 — 담당자의 기억에만 있다면 부재한 것과 같다

2 탐색 가능한가 그 기재가 어디에 있는지 확인 가능한가 — 개인 파일에만 있다면 탐색이 불가능하다

3 기계가 판독하는가 사람이 열람하는 문서인가, 프로그램이 직접 읽어 들이는 형태인가

4 신뢰할 수 있는가 누가 · 언제 · 무엇에 근거하여 기재하였는지가 함께 있는가

현장에서는 첫 번째 조건부터 충족하기 어렵다

AI Ready 데이터가 기관마다의 실무로만 남을 때의 한계

조직마다 다르다

동일한 업무를 부서마다 다르게 수행한다. 비교의 기준이 존재하지 않는다.

사람이 바뀌면 사라진다

담당자가 교체되면 판단의 근거도 함께 소실된다.

준수 여부를 확인할 수 없다

준수하였는지를 사후에 확인할 방법이 없다.

밖에서 인정되지 않는다

「준비되었다」는 선언이 다른 기관에 전달되지 않는다.

※ 가트너의 진단은 현재의 실무 관행을 기술한 **현상 분석**일 뿐, 조직이 이행하여야 할 **구체적 검증 규범(처방)**을 담고 있지 않다.

기관별 실무를 공통 절차로, 절차를 기계가 검증할 수 있는 규칙으로 — 표준의 역할

02

AI Ready 데이터에 이르는 접근과 기준이 통일되어 있지 않다

AI Ready 데이터에 이르는 출발점은 조직마다 다르다.

참고할 지침과 기준도 이미 여럿 나와 있으며, 각각은 저마다 타당하다.

다만 서로 다른 출발점과 기준이 내놓은 결과를 하나로 잇는 방법은 아직 없다.

출발점에 따라 「AI Ready」가 가리키는 것이 다르다

출발점	준비되었다고 보는 상태	완료로 보는 시점	그 작업만으로 알 수 없는 것
수집 · 정제	결측과 오류가 정리된 상태	정제 절차 통과	이 데이터를 무엇에 쓸 수 있는지
데이터 자체의 조건	형식과 구조 요건을 갖춘 상태	스키마 검증 통과	값의 의미가 맞는지
지식그래프 · 온톨로지	개체와 관계로 연결된 상태	그래프 적재 완료	그 그래프가 얼마나 정확한지
검색 · 맥락 구성	질의로 꺼낼 수 있는 상태	검색 결과가 나온다	꺼낸 것을 믿어도 되는지
에이전트 활용	호출해 쓸 수 있는 상태	도구가 읽어 들인다	쓸지 말지를 무엇으로 판단하는지

※ 각 출발점은 데이터의 구축 및 활용 목적이 서로 다르므로, 「준비되었다」가 의미하는 바 또한 상이하다. 특히 각 출발점마저 해결하지 못하는 **마지막 열의 공백**은, 다른 영역의 작업으로도 대신 채워질 수 없다.

품질 차원의 이름은 같아도 측정 방법이 다르다

여기서 「기준」이란

과기정통부 DQ 인증 기준³ · NIA 가이드라인 v3.5⁴ · ISO/IEC 25024⁵ · ISO/IEC 5259-2⁶ 등 지금 참고되고 있는 지침과 기준. 이 밖에도 기관·사업별 지침이 다수 존재한다.

완전성

어디까지를 결측으로 볼 것인가

일관성

무엇과 무엇을 대조하는가

정확성

무엇을 정답으로 삼는가

이름이 같다고 뜻까지 같지는 않다. 산정의 모집단이 무엇인지, 어느 수준이면 충족인지가 기준마다 다르므로 **여러 기준을 모두 충족하더라도 그 결과가 하나의 값으로 모이지 않는다.**

「우리 데이터는 준비되었습니다」가 기관 사이에서 통용되지 않는 이유

3 과기정통부 데이터 품질인증(DQ) 기준 · 데이터산업법 제20조 4 한국지능정보사회진흥원(NIA) 가이드라인 v3.5 5 ISO/IEC 25024:2015 6 ISO/IEC 5259-2:2024

데이터 품질 진단 도구가 보는 것과 보지 못하는 것

지금 도구가 보는 것 — 형식

- 값이 정해진 범위 안에 있는가
- 결측이 몇 퍼센트인가
- 허용된 코드값 목록에 들어 있는가
- 다른 표와의 참조가 일치하는가



보지 못하는 것 — 의미

- 이 항목이 무엇을 가리키는가
- 이 코드값이 무엇을 뜻하는가
- 이 값이 이 맥락에서 타당한가
- 다른 기관의 비슷한 항목과 같은 것인가

예 · 기관 내부 코드값

정의가 함께 공개되지 않은 경우

기관이 내부 업무를 위해 정한 코드값은 그 안에서는 통한다. 그러나 데이터를 공개할 때 **그 정의가 함께 공개되지 않으면 외부 이용자나 AI에게는 값의 뜻을 확인할 근거가 없다**. 형식 검사는 모두 통과하지만 의미로는 활용할 수 없으며, 「가용성」의 첫 번째 조건이 충족되지 않은 경우이다.

데이터를 준비하는 절차의 공통 기준도 미흡하다

달라야 하는 것 — 당연하다

- 도메인마다 다루는 **데이터의 성격**이 다르다
- 용도마다 요구하는 **정밀도와 형태**가 다르다
- 기관마다 쓰는 **어휘와 코드 체계**가 다르다

공통으로 지켜야 하는 것 — 정해져 있지 않다

- 데이터가 **지금 어느 상태에 있는지**를 가리키는 공통 기준이 없다
- 데이터를 **다음 상태로 넘겨도 되는지**를 확인하는 공통 조건이 없다
- 그 **확인 결과를 어디에 어떤 형식으로 기재할지**가 정해져 있지 않다

여기서 「**상태**」는 데이터가 지금 어디까지 준비되었는지를 가리키는 말이다. 그 상태를 몇 가지로 나눌지, 무엇을 확인하여야 다음으로 넘어가는지는 정해져 있지 않다.

모든 절차를 같게 만드는 것은 표준의 목표가 아니다 — 공통으로 지켜야 할 항목부터 정하여야 한다

정리 — 정의 · 어휘 · 형식 · 관리가 서로 어긋난다

1 **정의** 무엇을 「AI Ready」라 부르는가 — 방향을 제시하는 서술은 있으나 확인 기준으로 구체화되지 않는다

2 **어휘** 그 정의를 어떤 용어로 기술하는가 — 이름은 겹치지만 측정 방법이 다르다

3 **형식** 그 어휘를 어떤 구조로 기계가 읽게 하는가 — 형식은 검증되지만 의미는 검증되지 않는다

4 **상태 관리** 그 값이 언제 어떻게 바뀌고 누가 확인하는가 — 부여하는 규칙은 있으나 철회하는 규칙이 없다

앞의 요소가 정해져야 뒤의 요소가 정해진다 — 네 요소는 하나의 체계로 함께 규정하여야 한다

03

정의 · 어휘 · 형식 · 관리를 하나의 체계로 잇는다

AIRD 표준은 모든 항목을 같게 만들지 않는다.

공통으로 지킬 항목과 기관이 정할 항목을 나누어 규정하고,
그 판단의 근거를 기계가 확인할 수 있도록 남긴다.

AI Ready Data는 정제된 파일이 아니라 지속되는 상태다

실제로 저장되는 것

- 데이터 그 자체
- 그 데이터를 설명하는 정보 — 항목 단위의 기재
- 마니페스트 — 그 설명 정보를 하나로 묶은 문서
- 진단 리포트 — 무엇을 어떻게 측정하였는지의 기록

「AI Ready」라는 고정된 정제 파일은 없다

착각

한 번 정제하여 만들어 두는 고정된 형태의 완성 파일이 따로 존재한다

본질

데이터는 계속 변하므로, 왼쪽의 기록이 그 상태를 표현하고 증명하며 함께 유지될 때 성립하는 지속되는 상태이다

「AI Ready」는 별도로 제작하는 독립된 산출물이 아니라, 데이터셋이 갖추어야 할 객관적 상태 속성이다

준비 상태는 용도와 시간에 따라 변한다

용도에 따라 달라진다

같은 데이터셋이라도 용도에 따라 적합한 정도가 다르다. 목적 등급을 하나로 부여하지 않고 **용도별로 구분하여** 부여하는 이유이다. 「이 데이터는 준비되었는가」에 답하려면 「무엇에 활용할 것인가」가 먼저 정해져야 한다.

시간에 따라 변한다

확인한 뒤 **원본이 변경되면 스테일 상태가 되어 이미 부여된 품질 등급의 효력이 정지된다**. 앞서 비어 있던 「값을 철회하는 규칙」이 여기에서 채워진다. 준비도는 한 번 받아 두는 인증서가 아니다.

비유 · 식재료의 신선함 따로 구매하는 대상이 아니다

신선함은 **재료가 지닌 상태**이다. 보관 조건이 무너지면 사라지고, 용도에 따라 충분한지가 달라진다 — **생식용으로는 부족하더라도 가열용으로는 충분할 수 있다**.

따라서 준비도는 **한 번 부여하고 끝나는 표시가 아니라, 조건이 바뀌면 함께 바뀌는 값**으로 다루어야 한다. 부여하는 규칙과 철회하는 규칙을 함께 정하는 것이 이 표준의 출발점이다.

퀄리티 레디와 펄포즈 레디, 그 사이의 게이트

퀄리티 레디

용도와 무관하게 품질 기반을 충족

→

G3

펄포즈 레디

특정 용도의 정량 요건까지 충족

퀄리티 레디

용도를 정하지 않은 상태에서도 평가할 수 있는 **준비 상태**를 뜻한다. 품질 등급과 진단 성숙도 수준이 부여되어 있다.

펄포즈 레디

용도를 정한 뒤에야 비로소 판단할 수 있는 **준비 상태**이다. 앞 상태를 거쳐야 하며, **원본은 퀄리티 레디로 남고 별개의 산출물이 함께 존재한다.**

「AI Ready」는 이 둘을 포괄하는 용어이며 별도의 세 번째 상태가 아니다

데이터의 다섯 가지 상태와 세 곳의 결정 게이트



단계

데이터를 다음 준비 상태로 **전환하는 변환 작업**이다. 실행 결과로 한 묶음의 설명 정보를 생성한다.

결정 게이트

현재 데이터가 다음 단계의 준비 조건을 충족하였는지 **검증하는 통과 기준**이다. 표준은 이를 G1·G2·G3 셋으로 정의한다.

스테일

원본 데이터가 변경되어 **기존 준비 상태가 유효하지 않게 된 상태**이다. 이로 인해 이미 부여된 품질 등급의 효력이 일시 정지된다.

데이터가 지금 어느 준비 상태에 있는지를 가리키는 공통 기준 — AIRD 표준은 이를 다섯 가지 상태로 정의한다

가트너의 네 구성 요소와 이 표준이 규정하는 것

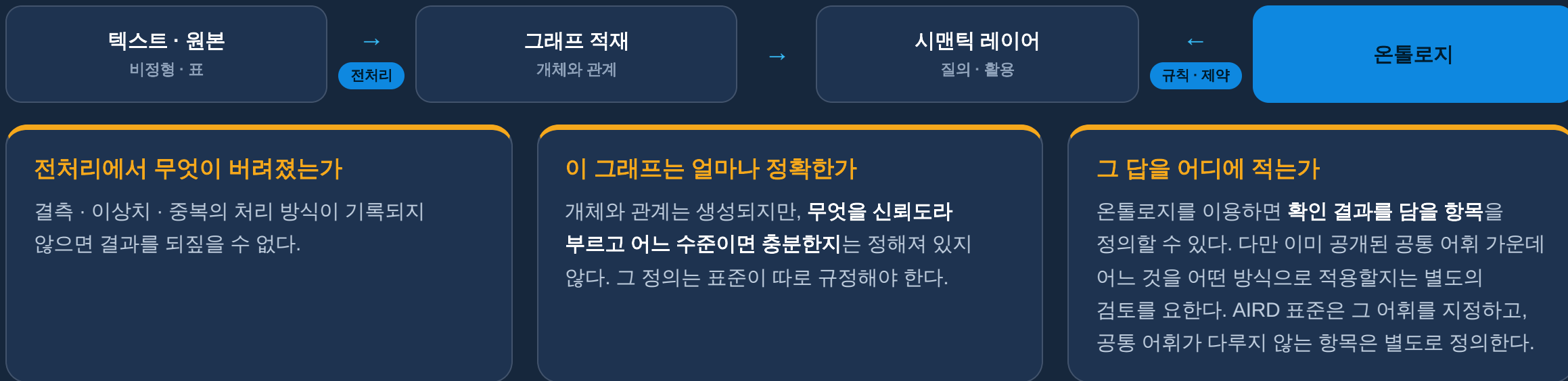
가트너의 구성 요소 ¹	이 표준이 제시하는 것	해당 부(Part)
메타데이터의 가용성	탐색 · 이해 · 즉시 활용을 위한 설명 정보 묶음	제3부 데이터셋 기술 스키마 및 프로파일
정합	활용 용도의 구분과 목적 등급	제3부 · 제4부 변환 파이프라인·거버넌스
자격 부여	품질 등급 · 진단 성숙도 수준 · 필수 통과선	제2부 품질 측정·등급화
거버넌스	마니페스트 · 진단 리포트 · 이력 · 결정 게이트	제4부 · 제5부 탐색·교환·활용

가트너가 프로세스이자 실무 체계로 서술한 것을 기계가 검증할 수 있는 규칙으로 구체화한 것이 이 표준이다

제1부 개요 및 프레임워크는 다섯 부가 공유하는 개념과 용어를 고정한다.

¹ Gartner, "AI-Ready Data Essentials to Capture AI Value"

지식그래프를 만드는 과정에서 온톨로지와 어휘가 맡는 역할



※ 기업 안에서도 원본이 아니라 **추출된 데이터**를 쓰는 경우가 많다. 추출과 정제의 이력이 끊기면 품질을 판단할 근거가 남지 않는다 — 앞서 본 「가용성」의 **네 번째 조건**이 성립하지 않는 지점이다.

표준 어휘를 이름만 빌려 쓰는 문제

같은 뜻을 세 기관이 이렇게 적는다

A 기관 · 자체 필드명

`"title": "서울시 연도별 인구 변화"`

B 기관 · DCTERMS 어휘

`"dct:title": "서울시 연도별 인구 변화"`

C 기관 · schema.org 어휘

`"name": "서울시 연도별 인구 변화"`

데이터셋명은 위 지침의 예시이며, 세 가지 표기는 설명을 위하여 구성하였다

어휘를 따르면 이름과 함께 세 가지 약속을 받는다

어휘는 이름의 목록이 아니라 기재 방식까지 정해 둔 규정이다.

- 필수 여부 — 어휘가 필수로 정한 항목은 비워 둘 수 없다
- 값의 개수 — 하나만 둘지 여럿을 나란히 둘지가 정해져 있다
- 값의 범위 — 그 자리에 쓸 수 있는 값의 목록이 정해져 있다

※ 세 기관 모두 「제목」을 적었다. 그러나 어느 어휘를 따랐는지 밝히지 않으면 기계는 셋이 같은 항목인지 알지 못한다. 이름을 빌리는 것과 어휘를 따르는 것은 다른 일이다.

이름이 같다고 뜻까지 같지는 않다 — 앞서 살펴본 「완전성」이 이에 해당한다

온톨로지란 무엇인가

단순한 용어 목록이 아니라, **용어의 의미와 이들이 준수하여야 할 관계 규칙을 함께 정의한 체계**이다. 필수 속성, 개체 간의 연계 관계, 그리고 허용되는 데이터 값의 제약을 규정함으로써, **사람은 이해하고 기계는 논리적으로 검증할 수 있게 한다**.

앞의 예시를 어휘에 연결하면

```
{
  "@context": {
    "aird": "http://datahub.kr/ns/aird#",
    "dcat": "http://www.w3.org/ns/dcat#",
    "dct": "http://purl.org/dc/terms/",
    "prov": "http://www.w3.org/ns/prov#"
  },
  "@type": "dcat:Dataset",
  "aird:metadataStatus": "Quality-Ready",
  "dct:title": "서울시 연도별 인구 변화",
  "prov:wasDerivedFrom": ".../kr-seoul-pop-raw"
}
```

AIRD 마니페스트 JSON 구성 지침(안) 발췌 — 일부 항목 생략

이 한 줄씩이 가져오는 것

- **@context** — 각 이름이 어느 어휘의 것인지 밝힌다. 세 기관의 표기가 여기에서 하나로 모인다
- **aird:** — 우리 표준의 자체 어휘. 표준 어휘로 표현되지 않는 것만 정의한다
- **dcat:Dataset** — 그 어휘가 정한 필수 항목을 함께 요구한다
- **prov:wasDerivedFrom** — 원본과의 연결을 기계가 따라갈 수 있게 한다
- **어휘를 가져오면** 그 어휘에 딸린 검증 규칙까지 함께 따라온다

AIRD가 채택한 표준 어휘와 그 채택 이유

탐색을 위한
기술

DCAT · DCTERMS⁷ — 공공데이터 포털이 이미 쓰는 어휘이므로 대응표가 필요 없다

출처와
이력

PROV-O⁸ — 원본에서 무엇이 어떤 처리를 거쳐 나왔는지를 기록한다

품질 측정
결과

DQV⁹ — 측정 결과를 진단 도구가 아니라 데이터 쪽에 기재한다

즉시
활용

CSVW · Croissant¹⁰ — 표와 학습 자료를 AI 도구가 그대로 적재할 수 있게 한다

기재
형식

JSON-LD¹¹ — 기존 JSON 구조를 유지하면서 위 어휘를 함께 표기한다

앞의 네 요소가 요구하는 것을 어휘별로 나누어 말기고, JSON-LD는 그 넷을 한 파일에 담는 형식이다. 표현력이 아니라 각 요소가 요구하는 것에 따라 정하였다.

⁷ W3C DCAT · DCMI Metadata Terms ⁸ W3C PROV-O ⁹ W3C DQV ¹⁰ W3C CSVW · MLCommons Croissant ¹¹ W3C JSON-LD 1.1

표현할 수 있다와 반드시 해야 한다는 다르다

표현력 — 어휘는 무엇을 적을 수 있게 하는가

- 해당 항목을 답을 수 있는 스키마 필드가 존재하면 된다
- 선택적 항목이므로 값을 비워 두어도 문법적 오류는 발생하지 않는다
- 해당 항목의 유효성 검증은 기계적 규칙이 아닌 사람의 해석에 의존한다

준수 — 표준은 무엇을 반드시 하도록 요구하는가

- 상태별로 기재해야 할 항목과 기재해서는 안 되는 항목이 정해져 있다
- 마니페스트에 상태와 등급 값을 적고, 진단 리포트를 정본으로 가리킨다
- 허용 값 목록을 벗어난 값은 이 표준에 적합하지 않다
- 원본이 바뀌면 이미 부여된 품질 등급의 효력을 정지시킨다

① 구조
필수·금지 항목

→

② 무결성
리포트 해시 일치

→

③ 값 일치
마니페스트 = 리포트

→

④ 규범 값
허용 목록 안인가

※ 「기계가 검증한다」는 이 네 단계를 뜻한다. 어휘를 이름만 빌리면 ①과 ④가 대조할 규격을 갖지 못한다 — 무엇이 필수인지, 어떤 값이 허용되는지를 가져오지 않았기 때문이다.

조직에 쌓인 암묵적 지식을 규칙으로 옮긴다

어떤 누락이 치명적인가

동일한 결측치이라도 **필수 속성**의 누락 여부에 따라 데이터의 활용 적합성이 갈린다.

데이터의 유효 기한은 언제까지인가

업무 도메인별로 데이터의 **생명주기**가 다르나, 그 판단 기준이 개인의 암묵지에 머물러 있다.

특정 목적에 부합하는가

동일한 이상치라도 **어떤 분석 목적**에서는 **보존**하고 **다른 학습**에는 **제거**하여야 하는 기준이 정의되어야 한다.

셋 모두 조직의 머릿속에만 남아 있는 **암묵적 지식**이다. 문서화되거나 규칙으로 명시되지 않으면, 사람은 물어보아 유연하게 대처하여도 **에이전트는 맥락을 가져올 곳이 없어 잘못된 판단에 이른다**. 온톨로지는 이 지식을 명시적 규칙과 값으로 전환하는 수단이며, 가트너가 지적한 「고품질 ≠ AI 레디」의 원인도 여기에 있다.¹

AIRD 표준은 이 지식을 **세 가지 영역**으로 구분하여 체계화하며, 이를 바탕으로 각 출발점이 제기한 질문에 유기적으로 대응한다.

표준 어휘와 허용 값 목록

의미 · 필수 속성 · 허용 값을 정의한다 · 제3부

품질 등급과 진단 리포트

사람이 확인하던 것을 기계가 검증하게 한다 · 제2부

다섯 상태와 세 결정 게이트

언제 어느 목적에 쓸 수 있는지를 정한다 · 제4부

¹ Gartner, "AI-Ready Data Essentials to Capture AI Value"

각 출발점이 남긴 질문을 어느 부가 답하는가

출발점	그 작업만으로 알 수 없던 것	이 표준이 답하는 자리
수집 · 정제	무엇에 쓸 수 있는지	활용 목적 유형과 목적 등급 — 제3부 · 제4부
데이터 자체의 조건	값의 의미가 맞는지	표준 어휘와 허용 값 목록 — 제3부
지식그래프 · 온톨로지	얼마나 정확한지	품질 등급과 진단 리포트 — 제2부
검색 · 맥락 구성	믿어도 되는지	진단 성숙도 수준 — 그 판단을 얼마나 신뢰할 수 있는가 · 제2부
에이전트 활용	쓸지 말지 무엇으로 판단하는지	다섯 상태와 세 결정 게이트 — 제4부

상태와 게이트의 **개념 정의**는 제1부가 고정하며, 위 표의 「답하는 자리」는 **실행 규범을 규정하는 부**를 가리킨다.

출발점을 하나로 만들지 않는다 — 어느 작업도 답하지 못한 것에 답할 자리를 정한다

정 리

설명이 아니라, 판독할 수 있는 상태

AIRD 표준은 「AI Ready」를 사람마다 다르게 해석하는 **모호한 설명이나 주관적 정의**로 두지 않습니다. 엑스레이 판독처럼 **동일한 기준을 적용하고 표준 어휘로 명시하도록** 규정합니다.

진단 — 다섯 가지 상태와 세 곳의 결정 게이트로 데이터가 지금 어느 상태에 있는지를 확인합니다.

추적 — 무엇이 어떤 처리를 거쳐 나왔는지가 마니페스트와 진단 리포트에 명시됩니다.

관리 — 원본 데이터가 변경되면 연계된 변환 파이프라인을 거쳐 스테일 상태로 자동 전환되며, 기존에 부여된 품질 등급의 효력이 정지됩니다.

이는 마니페스트와 진단 리포트 이력을 통제하는 **동적 거버넌스 체계**에 의해 보장됩니다.

사람과 에이전트가 동일한 값을 일관된 의미로 상호 해석할 수 있습니다.

참고문헌 및 인용 출처

정의 · 시장 조사

- 1 Gartner, “AI-Ready Data Essentials to Capture AI Value”**
gartner.com/en/articles/ai-ready-data · 본 발표의 AI 레디 데이터 관련 서술 인용
- 2 Gartner, “Lack of AI-Ready Data Puts AI Projects at Risk” (2025. 2. 26.)**
데이터 관리 리더 1,203명 조사(2024. 7.) · 63% 수치 출처

국내 기준 · 지침

- 3 과학기술정보통신부 데이터 품질인증(DQ) 기준**
근거 : 데이터산업법 제20조 · 한국데이터산업진흥원 시행
- 4 한국지능정보사회진흥원(NIA) 가이드라인 v3.5**
AI 구축사업 수행 기준

국제 표준

- 5 ISO/IEC 25024:2015**
Systems and software engineering — SQuaRE — Measurement of data quality
- 6 ISO/IEC 5259-2:2024**
Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures

채택한 표준 어휘

- 7 W3C DCAT · DCMI Metadata Terms**
데이터 카탈로그 및 자원 기술 어휘 — 탐색을 위한 설명 정보
- 8 W3C PROV-O : The PROV Ontology**
생성·변환 이력(출처) 기술 어휘
- 9 W3C DQV : Data Quality Vocabulary**
품질 측정 결과를 데이터에 부착하여 기술하는 어휘
- 10 W3C CSVW · MLCommons Croissant**
표 형식 데이터 및 기계학습 데이터셋 기술
- 11 W3C JSON-LD 1.1**
기존 JSON 구조를 유지하며 의미를 얻는 직렬화 형식

본 표준 시리즈

- 12 AI 레디 데이터 표준 제1부 ~ 제5부**
개요 및 프레임워크 / 품질 측정·등급화 / 데이터셋 기술 스키마 및 프로파일 / 변환 파이프라인·거버넌스 / 탐색·교환·활용 (제4부·제5부 작성 중)



감사합니다

Q & A

