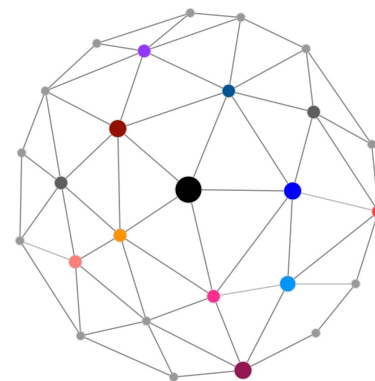


AI READY DATA 2026

공공데이터로 확인하는 AI Ready Data 진단과 구축

이정운 · 중앙대학교 HIKE Lab.

2026. 9. 3.



온톨로지는 목적에 따라 표현 수준이 다릅니다

온톨로지는 수준에 따라 다양한 층위로 표현되며, 목적에 따라 정할 수 있습니다.

01

어휘 · 분류

용어를 맞추고 묶는다
용어집 · 시소러스 · SKOS

02

개념 모델

의미로 잇는다
개념 모델 · 지식그래프

03

제약 · 검증

위반을 판정한다
품질 규칙 · SHACL

04

액션 · 실행 경계

무엇을 해도 되는지 정한다
사전조건 · 권한

이 층위가 없으면 못 하는 일

같은 것을 같다고
선언할 수 없다

흠어진 데이터를
이어서 볼 수 없다

틀린 값이
그대로 통과한다

실행해도 되는지
정할 수 없다

관련 기술 · 작업

✓ 스키마 설계

✓ 지식그래프 · GraphRAG

✓ 제약조건 · 규칙(SHACL)

한쪽만으로는 작동하지 않습니다

무게중심은 오른쪽으로 옮겨갔지만, 오른쪽은 왼쪽 위에서만 성립합니다.

어휘 · 분류 · 연결



규칙 · 제약 · 실행 경계

지난 20년의 무게중심

지금 이동하는 방향

왼쪽이 없으면 규칙과 제약을 적용할 대상이 없습니다

오른쪽이 없으면 에이전트가 실행해도 되는지 판단할 수 없습니다

의미 요소

객체 · 속성 · 링크

동적 요소

액션 · 함수 · 보안

운영 구조

권한 · 감사 · 결과 반영

한쪽 방향에서만 진행하면, 적용 대상이 없는 규칙만 남습니다.

사람은 데이터 밖에서 채우고, 에이전트는 만들어냅니다

사람

데이터 밖에서 보충한다



에이전트

데이터에 없으면 만들어낸다



같은 데이터에서 출발하지만, 빈칸을 채우는 방식이 다릅니다.

온톨로지는 그 빈칸을 미리 채워 둡니다

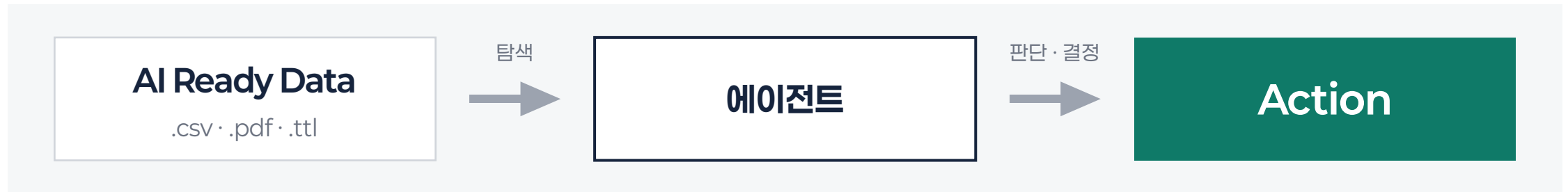
에이전트가 빈칸을 스스로 만들지 않게 하려면, 데이터에 미리 적혀 있어야 합니다.

에이전트가 만들어내는 가정	온톨로지가 대신 주는 것	
“이 컬럼이 무슨 뜻인가”	의미	개념과 속성의 정의
“이 값이 타당한가”	제약	허용 범위와 규칙
“무엇을 해도 되는가”	경계	액션과 사전조건

이 상태를 정의한 것이 AI Ready Data입니다

온톨로지 기반 위에서 에이전트가 동작합니다

② 활용 · 에이전트가 동작합니다



① 기반 · 이 두 층이 없으면 위가 성립하지 않습니다

01 도메인 온톨로지

도메인별 참조데이터 — 식별자 · 분류체계 · 통제어휘

주소

기관명

업종 분류

02 구조 온톨로지

전 범위 공통 — DCAT · DQV · PROV-O · SHACL

의미

제약

경계

PART 2 · IN PRACTICE

공공데이터를 활용한 AI Ready Data 실증

방대하고 품질 비판을 받는 공공데이터를
에이전트가 바로 쓸 수 있는 상태로 만든 과정

산업 분야 공공데이터 약 1,800건

공공데이터포털(data.go.kr)에 개방된 산업 분야 데이터를 대상으로 실증했습니다.

대상 규모

기관	데이터셋	컬럼	행
21 개	1,773 건	18,788 개	약 8,300만 행

파일 형태

정형 테이블 데이터 컬럼과 행이 정의된 표 형식	CSV	XLSX	...	스키마와 값을 기계가 읽을 수 있는 형식
-------------------------------	-----	------	-----	---------------------------

맥락은 데이터 공유 과정에서 사라집니다

① 컬럼명

시스템 내부의 맥락 소실

② 정의·출처

불충분한 컬럼 정의서

③ 시점·버전

스냅샷 제공

DB

값 — 세 곳 모두 동일

mkt_nm · reg_dt

스키마의 영문 컬럼명

설명 · 분류체계 · 출처

DB 안에는 존재한다

운영 원본 — 계속 갱신

오늘의 값

CSV

값 — 세 곳 모두 동일

「역사명」

「역사 이름」

제공처마다 다른 한글명 — 이질성

정의서 미제공

사용자는 알 방법이 없다

시점 미상의 스냅샷

삭제된 컬럼 · 시점별 차이

AI Ready

값 — 세 곳 모두 동일

역사명 = 역사 이름

표준 어휘로 의미를 선언

정의 · 출처를 함께 발행

기계가 읽는 형식으로

스냅샷 시점 · 버전 명시

스냅샷의 버전을 계보로 관리

AI Ready Data는 소실된 맥락을 채워줍니다

목표 — 기관마다 따로 공개된 데이터셋을 에이전트가 스스로 찾고 이어붙일 수 있게

메타데이터

메타데이터를 RDF 어휘로 표현

표현 방식이 통일된다

컬럼(명)

컬럼 수준까지 표준을 적용

연계 조건이 드러난다

데이터값

참조데이터 연결 · URI 부여

시맨틱 데이터가 된다

참조데이터: 식별자(주소, 기관) · 분류체계(한국표준산업분류) · 통제어휘(코드1, 2) → 도메인 온톨로지

온톨로지부터 설계하지 않았습니다

순서를 뒤집으면 같은 자원으로 끝낼 수 있는 범위가 달라집니다.

흔한 순서

- 1 데이터 현실을 모른 채 온톨로지부터 설계
- 2 설계한 모델에 맞는 데이터가 없다
- 3 품질 확보에 자원이 소모된다
- 4 범위가 축소된 PoC로 종료

우리가 한 순서

- 1 무엇에 답할 데이터인지 먼저 적는다
- 2 데이터가 감당할 범위가 거기서 보인다
- 3 카탈로그와 품질부터 — 어느 목적에나 쓰인다
- 4 목적이 정해진 다음에 도메인 온톨로지

도메인 온톨로지부터 설계했다면, 1,800건은 끝내지 못했습니다.

수집에서 목적별 팩까지, 상태가 한 단계씩 올라갑니다

각 단계를 통과할 때마다 데이터의 상태가 판정됩니다.



AI Ready Data

품질이 검증되고, 목적이 부여된 상태

(1) 수집 · 보강: 사람이 아니라 에이전트가 읽는 형식으로

같은 데이터셋이 두 형태로 존재합니다. 오른쪽만 에이전트가 그대로 읽습니다.

사람이 읽는 화면 · 공공데이터포털

파일데이터 정보			
데이터 오류·개선 의견이 있으신가요? 데이터 개선요청 > 오류신고 및 문의 >			
기본 정보			
파일데이터명	산업통상부_외국인투자기업_20251209		
분류체계	산업·통상·중소기업 - 무역및투자유치	제공기관	산업통상부
관리부서명	투자유치과	관리부서 전화번호	044-203-4087
보유근거	외국인투자촉진법 제24조	수집방법	
업데이트 주기	수시 (자동 갱신)	차기 등록 예정일	
상세 및 제공정보			
설명	해당 자료는 외국인투자촉진법에 근거하여 등록된 외국인 투자기업에 관한 자료로 회사명, 표준산업중분류명, 분류명, 분류코드 등에 관한 정보를 제공하고 있음 - 해당 데이터는 외투기업의 투자 등록기업에 대한 상세현황 자료로 주요투자국가(제1지분투자자)와 해당기업의 주활동 업종과 분야를 파악할 수 있음 - 해당 기업의 위치 정보를 파악할 수 있어 지역 사회 내의 외투기업의 존재와 경제 참여 규모등을 간접적으로 파악할 수 있음 - 국내 산업 내 주요 산업별 투자 현황 등을 파악할 수 있어 투자시책개발에 도움		
매체유형	텍스트	확장자	XLS
전체 행	19651	키워드	외국인직접투자기업, 투자기업정보, 투자자, 투자국가, 주업종, 투자 업종
데이터 한계		공간범위	
시간범위		제공형태	기관자체에서 다운로드(제공데이터URL기재)
다운로드(바로그기)	4492	URL	https://www.motir.go.kr/kor/contents/105
등록일	2020-01-14	수정일	2025-12-09
기타 유의사항			

① Discoverable Metadata

에이전트가 이 데이터에 대해 갖는 최소한의 정보이자, 데이터로 들어가는 진입점입니다.

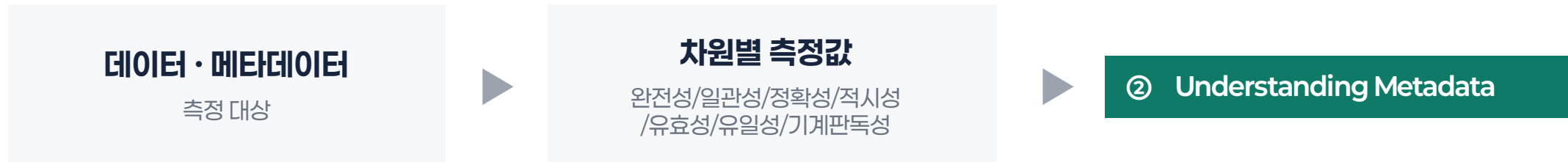
무엇에 관한 데이터인가 · 어디서 어떻게 받는가 · 어떤 컬럼으로 구성되는가

```
{
  "@id": "aird:dataset/15054398",
  "@type": "dcat:Dataset",
  "dct:title": "산업통상부_외국인투자기업",
  "dcat:theme": "무역및투자유치",
  "dct:publisher": "aird:organization/1451000",
  "dct:accrualPeriodicity": "freq:IRREG",
  "dcat:distribution": { "mediaType": "text/csv", "encoding": "cp949" },
  "csvw:tableSchema": {
    "csvw:column": [
      { "csvw:name": "연도", "csvw:datatype": "xsd:date" },
      { "csvw:name": "신규등록", "csvw:datatype": "xsd:string" }
    ], "aird:metadataStatus": "Discoverable"
  }
}
```

에이전트는 사람에게 묻지 않고, 이 문서만 읽고 쓸 수 있는지 판정합니다.

(2) 품질진단: 얼마나 좋은가와, 그 판단을 믿을 수 있는가

데이터와 메타데이터를 함께 측정해 차원별 점수를 내고, 두 축으로 판정합니다.



Q-Tier · 품질 등급

이 데이터는 얼마나 좋은가

Tier 0 Tier 1 Tier 2 Tier 3 Tier 4

데이터를 개선하면 오른다

DM · 진단 성숙도

그 판단을 얼마나 신뢰할 수 있는가

품질 점수와 합산하지 않고 별도 축으로 관리한다

측정 범위를 넓히면 오른다

진단 결과는 메타데이터에 남습니다
측정 결과는 메타데이터에 증강되어 **Quality-Ready**의 근거가 됩니다

(3) 정제: 형식이 비슷해도 같은 식별자로 묶지 않습니다

정제는 세 층위에서 진행되고, 원본과 정제 데이터는 이력으로 구분되어 추적될 수 있습니다.

컬럼명 정규화

제각각인 표기를 하나로

인스턴스 정제

참조데이터로 같은 것을 묶는다

메타데이터 증강

정제 근거를 문서에 기재

인스턴스 정제 예시

한국세라믹기술원_세라믹소재정보은행 데이터소스 정보.csv

특허출원번호 10-2024-0123456

표준 표기

대한석탄공사_지식재산권 특허 보유현황.csv

출원번호 1020240123456

하이픈 없음

한국석유공사_상표권 보유 현황.csv

등록번호 상표 제40-1234567호

텍스트 혼입

참조데이터

특허청 KIPRIS 특허 · 상표 출원/등록정보

10-2024-0123456

✓ 특허출원번호 확인

1020240123456 → 10-2024-0123456

✓ 구분자 표준화

상표 제40-1234567호 → 40-1234567

✓ 상표등록번호 분류

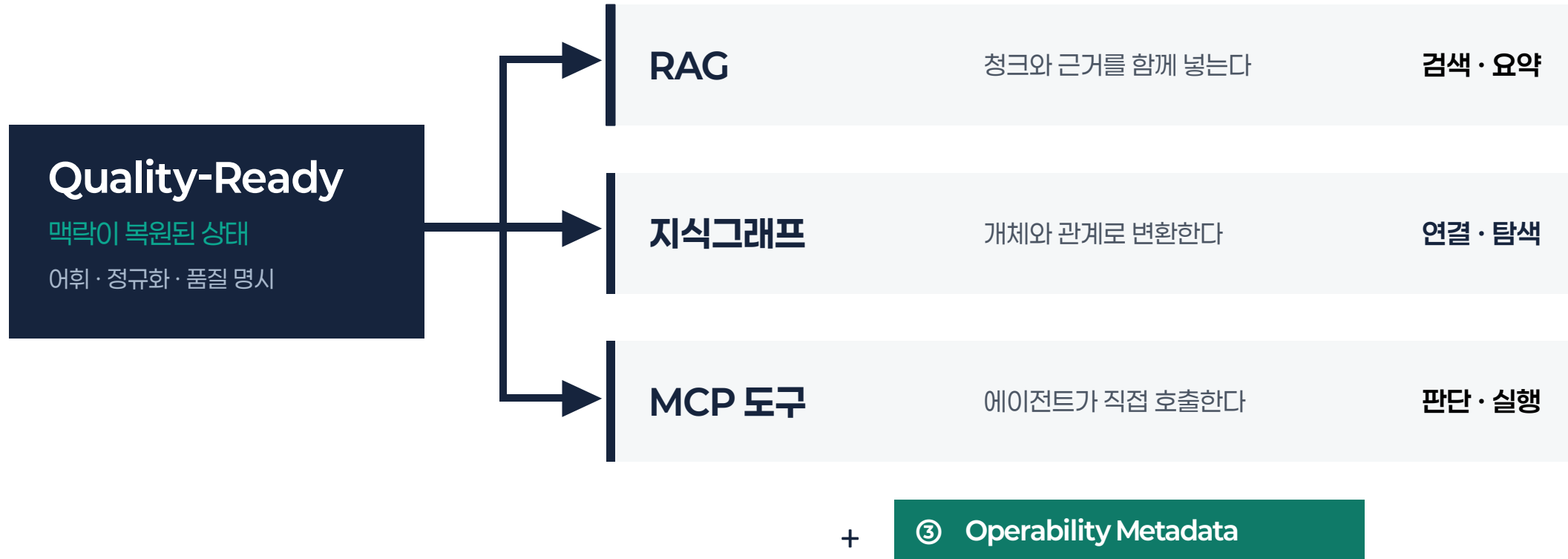
1234567 (종류 불명)

× 확정 불가 — 재확인

형식이 비슷해도 같은 식별자로 통합하지 않습니다

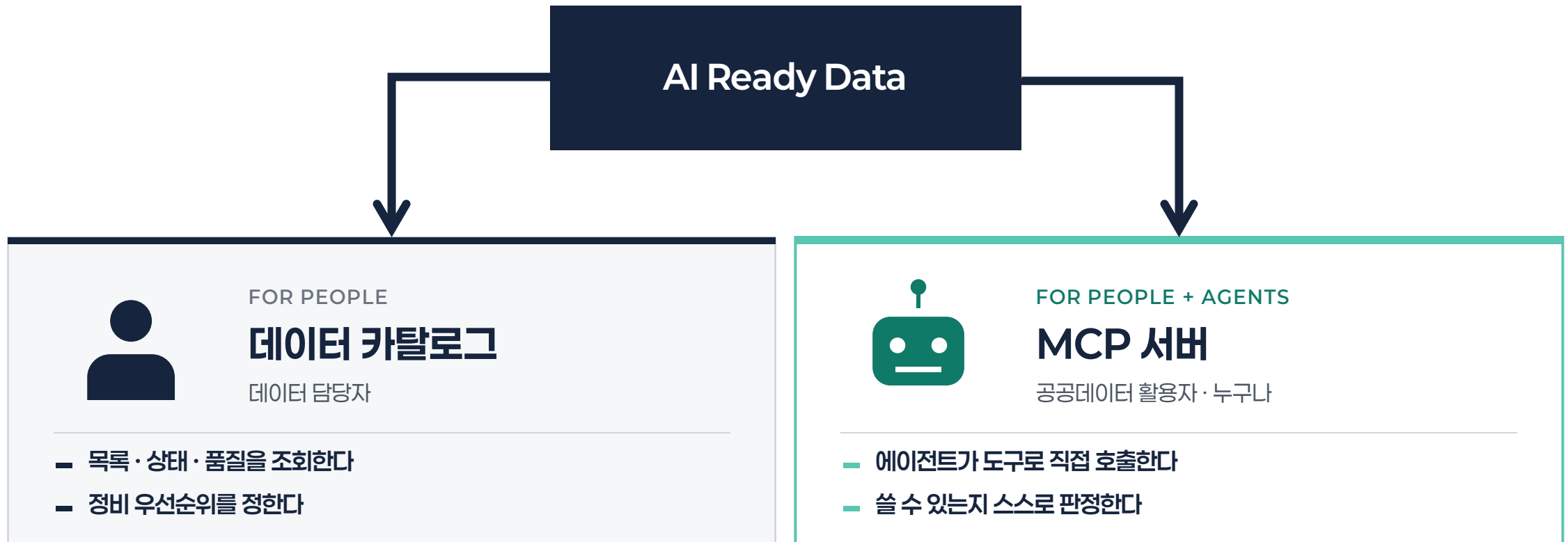
참조데이터를 기반으로 유형을 구분해 각각의 표준 형식으로 정제하고, 판정 근거를 함께 남깁니다

(4) 목적별 팩: 목적이 형태를 정합니다



미리 파생해 두면 에이전트는 매번 재구성하지 않아 **추론 비용을 줄일 수 있습니다**

같은 상태를 사람도, 에이전트도 그대로 읽습니다



사람이 쓰는 화면을 넘어, **에이전트가 접근하고 판단하는 과정에 그대로 들어갑니다**

PART 3 · DEMO

데모

– 에이전트가 읽는 데이터

준비된 공공데이터를 에이전트가 스스로 확인하고,
판정의 과정을 거치는 데모

온톨로지의 역할은 넓습니다

앞의 역할은 그대로 남고, 그 위에 새로운 역할이 더해졌습니다.

온톨로지의 역할

역할 1

지식을 넣어주는 것

그대로 유지

여전히 필요한 것

- 개념과 속성을 정의한다
- 같은 것을 같다고 선언한다

역할 2

판정하고 막는 것

쓸 만한지 판정하고, 해도 되는 일과 안되는 일을 제약

새로 필요해진 것

- 이 값이 쓸 만한지 판정한다
- 누가 무엇을 해도 되는지 정한다
- 어떻게 여기까지 왔는지 남긴다

지식을 넣어준 뒤 그 위에서 판정까지의 역할을 수행할 수 있습니다

AI Ready Data에서 온톨로지는 이렇게 활용됩니다

AI Ready Data를 구축하고 활용하는 과정에서 온톨로지는 기술 · 판정 · 허용과 금지 · 이력 기록에 활용됩니다

역할	질문	표준
① 기술한다	이것이 무엇인가	DCAT 3 · CSVW · SKOS
② 판정한다	이 값이 쓸 만한가	DQV(정도) · SHACL(통과 여부)
③ 허용 · 금지한다	누가 무엇을 해도 되는가	ODRL
④ 이력을 남긴다	어떻게 여기까지 왔는가	PROV-O · DCAT 3 버전

AI Ready Data는 이 네 역할이 데이터 쪽에 적혀 있는 상태입니다.

AI READY DATA 2026

감사합니다

이정운 · 중앙대학교 HIKE Lab.

2026. 9. 3.

