# 데이터 표준화 방법론

KANDLE Data Standardization Methodology



# Contents


들어가며 . . . . . 9

I 1부: 데이터 표준화 개요

1 데이터 표준화의 필요성 . . . . . 13
1.1 개요 . . . . . 13
1.2 데이터 표준화의 배경 . . . . . 14
1.3 표준화 부재의 비용 . . . . . 15
1.4 데이터 표준화의 가치 . . . . . 16
1.5 표준화 도입의 도전과 실패 원인 . . . . . 17
1.6 이 책의 구조와 활용 방법 . . . . . 18

2 핵심 개념과 용어 . . . . . 21
2.1 개요 . . . . . 21
2.2 데이터의 본질 . . . . . 22
2.3 표준화 관련 핵심 용어 . . . . . 23
2.4 데이터 관리 핵심 개념 . . . . . 24
2.5 데이터 품질과 거버넌스 용어 . . . . . 26
2.6 최신 개념 . . . . . 27
2.7 개념 지도와 관계 . . . . . 27
2.8 사례 연구 . . . . . 28

3 주요 표준화 프레임워크 . . . . . 31
3.1 개요 . . . . . 31

3.2 일반 관리 프레임워크 . . . 32
3.3 국제 표준(ISO/IEC) . . . 33
3.4 산업별 표준 . . . 35
3.5 국가·정부 프레임워크 . . . 37
3.6 프레임워크 선택 및 조합 . . . 37
3.7 사례 연구 . . . 39

II 2부: 표준화 방법론

4 데이터 모델링 표준 . . . 45
4.1 개요 . . . 45
4.2 핵심 개념 . . . 46
4.3 방법론 . . . 48
4.4 실무 적용 . . . 51
4.5 사례 연구 . . . 53

5 메타데이터 관리 . . . 57
5.1 개요 . . . 57
5.2 핵심 개념 . . . 58
5.3 방법론 . . . 60
5.4 실무 적용 . . . 63
5.5 사례 연구 . . . 64

6 데이터 품질 관리 . . . 69
6.1 개요 . . . 69
6.2 핵심 개념 . . . 70
6.3 방법론 . . . 72
6.4 실무 적용 . . . 74
6.5 사례 연구 . . . 76

7 마스터 데이터 관리(MDM) . . . 79
7.1 개요 . . . 79
7.2 핵심 개념 . . . 80
7.3 방법론 . . . 82
7.4 실무 적용 . . . 83
7.5 사례 연구 . . . 85

III 3부: 실무 적용

8 표준화 구현 전략 . . . . . 91
8.1 개요 . . . . . 91
8.2 핵심 개념 . . . . . 92
8.3 방법론 . . . . . 93
8.4 실무 적용 . . . . . 96
8.5 사례 연구 . . . . . 98

9 데이터 거버넌스 체계 . . . . . 101
9.1 개요 . . . . . 101
9.2 핵심 개념 . . . . . 102
9.3 방법론 . . . . . 104
9.4 실무 적용 . . . . . 106
9.5 실무 적용 . . . . . 108
9.6 사례 연구 . . . . . 108

10 표준화 도구와 기술 . . . . . 113
10.1 개요 . . . . . 113
10.2 핵심 개념 . . . . . 114
10.3 방법론 . . . . . 115
10.4 실무 적용 . . . . . 117
10.5 사례 연구 . . . . . 119

11 실제 적용 사례 . . . . . 123
11.1 개요 . . . . . 123
11.2 금융 산업 사례 . . . . . 124
11.3 공공 부문 사례 . . . . . 125
11.4 소매·서비스 산업 사례 . . . . . 127
11.5 헬스케어 산업 사례 . . . . . 129
11.6 제조 산업 사례 . . . . . 130
11.7 종합 교훈 . . . . . 132
11.8 여러분 조직에 적용하기 . . . . . 133

IV 4부: 고급 주제

12 AI·ML과 데이터 표준화 . . . . . 137
12.1 개요 . . . . . 137
12.2 핵심 개념 . . . . . 138
12.3 방법론 . . . . . 140

12.4 실무 적용 . . . 141
12.5 사례 연구 . . . 143

13 클라우드 환경의 표준화 . . . 147
13.1 개요 . . . 147
13.2 핵심 개념 . . . 148
13.3 방법론 . . . 150
13.4 실무 적용 . . . 151
13.5 사례 연구 . . . 153

14 국제 표준 심화 . . . 157
14.1 개요 . . . 157
14.2 ISO/IEC 11179: 메타데이터 등록소 심화 . . . 158
14.3 ISO 8000: 데이터 품질 파트별 심화 . . . 160
14.4 ISO/IEC 25012: 데이터 품질 모델 심화 . . . 161
14.5 ISO/IEC 42001: AI 관리 시스템 심화 . . . 162
14.6 표준 조합과 통합 활용 . . . 164
14.7 사례 연구 . . . 166

V 5부: 디플래닉스 방법론

15 디플래닉스 방법론 개요 . . . 171
15.1 개요 . . . 171
15.2 본 장의 학습 목표 . . . 171
15.3 KDSM의 배경 . . . 172
15.4 KDSM의 구성 . . . 172
15.5 KDSM의 특징 . . . 176
15.6 다른 프레임워크와의 관계 . . . 177
15.7 KDSM의 적용 . . . 177

16 KDSM 데이터 모델링 . . . 181
16.1 개요 . . . 181
16.2 본 장의 학습 목표 . . . 181
16.3 KDSM 모델링 원칙 . . . 182
16.4 KDSM 모델링 프로세스 . . . 183
16.5 KDSM 명명 규칙 . . . 186
16.6 산출물 표준 . . . 187
16.7 검토와 승인 . . . 189

16.8 지원 도구 . . . 190
16.9 사례 연구 . . . 192

17 KDSM 메타데이터 관리 . . . 195
17.1 개요 . . . 195
17.2 본 장의 학습 목표 . . . 195
17.3 KDSM 메타데이터 원칙 . . . 196
17.4 KDSM 메타데이터 등록소 . . . 196
17.5 KDSM 메타데이터 수집·관리 프로세스 . . . 196
17.6 KDSM 메타데이터 활용 . . . 196
17.7 도구와 인프라 . . . 196
17.8 사례 연구 . . . 196

18 KDSM 데이터 품질 관리 . . . 197
18.1 개요 . . . 197
18.2 본 장의 학습 목표 . . . 197
18.3 KDSM 품질 관리 원칙 . . . 198
18.4 KDSM 품질 차원 . . . 198
18.5 KDSM 품질 측정 체계 . . . 198
18.6 KDSM 품질 관리 프로세스 (PDCA) . . . 198
18.7 데이터 품질 이슈 대응 . . . 198
18.8 도구와 인프라 . . . 198
18.9 사례 연구 . . . 198

용어 사전 . . . 199

표준화 템플릿 모음 . . . 217

실무 체크리스트 . . . 221

참고문헌 . . . 223

# 들어가며

## 이 책에 대하여

데이터는 현대 조직의 핵심 자산입니다. 그러나 데이터의 가치는 그것이 **얼마나 잘 표준화되어 있는가**에 크게 좌우됩니다. 서로 다른 시스템, 부서, 조직 간에 데이터가 일관성 없이 관리될 때 발생하는 문제들 — 중복, 오류, 해석의 불일치 — 은 의사결정의 질을 떨어뜨리고 디지털 혁신을 가로막는 주요 장애물이 됩니다.

이 책은 데이터 표준화의 **이론적 기반**부터 **실무 적용**까지를 체계적으로 다룹니다.

## 이 책의 구성

이 책은 5부로 구성되어 있습니다.

**1부: 데이터 표준화 개요**는 데이터 표준화의 배경, 핵심 개념, 그리고 주요 국제 프레임워크를 소개합니다.

**2부: 표준화 방법론**은 데이터 모델링, 메타데이터 관리, 데이터 품질, 마스터 데이터 관리(MDM) 등 핵심 방법론을 심층적으로 다룹니다.

**3부: 실무 적용**은 조직에서 데이터 표준화를 실제로 구현하는 방법, 거버넌스 체계 수립, 도구 활용, 그리고 실제 사례를 살펴봅니다.

**4부: 고급 주제**는 AI/ML 시대의 데이터 표준화, 클라우드 환경, 국제 표준 등 최신 동향을 다룹니다.

**5부: 디플래닉스 방법론**은 디플래닉스의 데이터 표준화 방법론으로서의 데이터 모델링, 메타데이터 관리, 데이터 품질 방법론을 다룹니다.

## 독자 대상

> **i** 이 책은 다음 독자를 위해 쓰였습니다
>
> - 데이터 아키텍처 및 거버넌스 담당자
> - 데이터 엔지니어 및 분석가
> - IT 관리자 및 CIO
> - 데이터 표준화를 처음 접하는 실무자

## 감사의 말

(감사의 말을 작성하세요)

## 표기 규칙

이 책에서 사용하는 표기 규칙은 다음과 같습니다.

- `고정폭 글꼴`: 코드, 파일명, 명령어
- **굵은 글꼴**: 핵심 용어의 첫 등장
- *기울임꼴*: 외래어, 강조

> 팁
>
> 실무에서 바로 활용할 수 있는 유용한 정보를 제공합니다.

> ⚠ 주의
>
> 주의해야 할 사항이나 흔히 발생하는 실수를 알려줍니다.

> ! 중요
>
> 반드시 숙지해야 할 핵심 내용입니다.

> **i** 참고
>
> 주요한 참고 내용을 함축해서 알려줍니다.

# I

# 1부: 데이터 표준화 개요


1 데이터 표준화의 필요성 . . . . . 13
1.1 개요 . . . . . 13
1.2 데이터 표준화의 배경 . . . . . 14
1.3 표준화 부재의 비용 . . . . . 15
1.4 데이터 표준화의 가치 . . . . . 16
1.5 표준화 도입의 도전과 실패 원인 . . . . . 17
1.6 이 책의 구조와 활용 방법 . . . . . 18

2 핵심 개념과 용어 . . . . . 21
2.1 개요 . . . . . 21
2.2 데이터의 본질 . . . . . 22
2.3 표준화 관련 핵심 용어 . . . . . 23
2.4 데이터 관리 핵심 개념 . . . . . 24
2.5 데이터 품질과 거버넌스 용어 . . . . . 26
2.6 최신 개념 . . . . . 27
2.7 개념 지도와 관계 . . . . . 27
2.8 사례 연구 . . . . . 28

3 주요 표준화 프레임워크 . . . . . 31
3.1 개요 . . . . . 31
3.2 일반 관리 프레임워크 . . . . . 32
3.3 국제 표준(ISO/IEC) . . . . . 33
3.4 산업별 표준 . . . . . 35
3.5 국가·정부 프레임워크 . . . . . 37
3.6 프레임워크 선택 및 조합 . . . . . 37
3.7 사례 연구 . . . . . 39

CHAPTER 1

# 데이터 표준화의 필요성

## 1.1 개요

1854년, 런던의 콜레라 유행 당시 의사 John Snow는 감염자 위치와 우물의 위치를 지도 위에 체계적으로 기록하여 오염된 우물이 감염원임을 규명했다 (Snow 1855). 그의 조사가 성공한 이유는 단순히 데이터를 '많이' 수집해서가 아니라, **동일한 방식으로 기록·해석·비교** 할 수 있는 표준화된 형식을 사용했기 때문이다. 데이터 표준화의 역사는 통계학과 역학(Epidemiology)의 태동만큼이나 오래되었다.

170년이 지난 오늘, 조직은 매일 수 페타바이트의 데이터를 생성하지만 그 상당 부분이 표준화되지 않아 활용되지 못하고 있다. IBM의 2022년 조사에 따르면, **조직이 수집한 데이터의 약 80%는 다시 활용되지 못하는 '다크 데이터(Dark Data)'**이며, 이 다크 데이터의 주된 원인은 표준화 부재이다 (IBM 2022). 데이터의 양이 아니라 데이터의 '언어 통일'이 조직의 데이터 활용 능력을 결정한다.

Ackoff (1989) 는 이미 1989년 논문 'From Data to Wisdom' 에서 '우리는 정보에 익사하고 있으나 지식에 굶주려 있다(drowning in information but starving for knowledge)' 라고 진단했다. 30여 년이 지난 지금 이 진단은 오히려 더 절실해졌다. AI와 클라우드 시대에 데이터 표준화는 선택이 아닌 조직 생존의 필수 조건이 되었다.

### 1.1.1 데이터 표준화의 정의

본서는 데이터 표준화를 다음과 같이 정의한다.

> **i** 데이터 표준화(Data Standardization)의 정의
>
> 조직 내·외부에서 데이터를 공통의 정의·구조·형식·값으로 일관되게 관리하기 위한 정책·프로세스·기술의 총체. 데이터가 사람과 시스템 사이에서 오해 없이 이해·교환·활용될 수 있도록 하는 '데이터의 공통 언어'를 확립하는 활동이다.

이 정의는 DAMA International (2017) 의 관점을 반영한다. 표준화는 단순히 '이름을 통일하는 것'이 아니라, 데이터의 의미(semantics), 구조(syntax), 값(values), 프로세스(processes)에 걸친 다층적 활동이다. 이는 다음 장부터 상세히 다룬다.

### 1.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 데이터 표준화의 배경과 오늘날 특별히 중요한 이유를 설명할 수 있다
- 표준화 부재의 4대 비용(운영·의사결정·규제·전략적 기회 상실)을 이해한다
- 데이터 표준화가 창출하는 정량적·정성적 가치를 제시할 수 있다
- 표준화 도입의 5대 실패 원인과 대응 방향을 안다
- 조직에 데이터 표준화의 비즈니스 케이스를 설명할 수 있다

## 1.2 데이터 표준화의 배경

### 1.2.1 데이터 폭증의 시대

IDC (2023) 의 Global DataSphere 보고서에 따르면, **2025년 전 세계 데이터 총량은 181제타바이트(ZB)에 이를 것**으로 전망된다. 이 수치는 2015년의 15배이며, 2010년의 100배가 넘는다. 그러나 이 폭증하는 데이터 중 유의미하게 활용되는 비율은 오히려 감소하고 있다.

| 구분 | 2015 | 2020 | 2025 전망 | 2028 전망 |
|---|---|---|---|---|
| 글로벌 데이터 총량 (ZB) | 12 | 64 | 181 | 394 |
| 엔터프라이즈 데이터 비율 | 40% | 55% | 70% | 78% |
| 활용되는 데이터 비율 | ~30% | ~20% | ~15% | ~10% |

이러한 역설 — **'더 많은 데이터, 더 낮은 활용률'** — 의 원인은 명확하다. 데이터가 서로 다른 시스템, 형식, 정의로 흩어져 있어 통합·해석이 불가능한 것이다. 표준화 없이는 데이터는 그저 저장 비용의 근원일 뿐이다.

### 1.2.2 시스템 다양성의 폭발적 증가

20년 전 기업의 IT 아키텍처는 상대적으로 단순했다. ERP, CRM, HR, 회계 등 몇 개의 핵심 시스템이 온프레미스에 존재했다. 오늘날 조직은 평균 130~200개의 SaaS 애플리케이션을 사용하며, 여기에 클라우드 데이터 서비스, IoT 센서, 스트리밍 플랫폼, AI/ML 시스템이 얽혀 있다 (IDC 2024a).

각 시스템은 자체 데이터 모델과 형식을 가지고 있다. 시스템 수가 N개일 때 **가능한 데이터 통합 경로는 N(N-1)/2개**로 증가하는 '조합 폭발(combinatorial explosion)' 문제가 발생한다. 200개 시스템은 이론적으로 19,900개의 통합 경로가 가능하며, 표준화 없이는 관리가 불가능하다.

### 1.2.3 AI·ML의 부상

2020년대 인공지능의 폭발적 확산은 데이터 표준화를 새로운 차원으로 밀어올렸다. AI 모델은 데이터로부터 패턴을 학습하고 자동화된 의사결정을 내린다. 이때 학습 데이터의 품질과 표준화가 모델 성능을 결정한다. Google 연구팀의 Sambasivan et al. (2021) 은 **AI 프로젝트 실패의 92%가 데이터 문제**에서 기인한다고 보고했다.

Stanford의 Andrew Ng은 이를 계기로 Data-Centric AI 운동을 제창했다. 알고리즘 튜닝보다 데이터 품질과 표준화가 모델 성능 향상에 더 큰 기여를 한다는 것이다 (Ng 2021). AI 시대에 데이터 표준화는 '뒤에서 지원하는 활동'이 아니라 '가치 창출의 핵심'이다.

### 1.2.4 규제 환경의 강화

2010년대 후반부터 데이터 관련 규제가 급격히 확산되고 있다. GDPR(2018), CCPA(2020), 개인정보보호법 개정(2020), European Parliament (2024) (2024), ISO/IEC (2023a) (2023) 등이 조직에 명시적인 데이터 관리·문서화·표준화 의무를 부과하고 있다.

| 규제 | 발효 | 핵심 요구 | 위반 시 벌금 |
|---|---|---|---|
| GDPR | 2018 | 개인정보 처리 표준·기록 유지 | 전 세계 매출의 4% |
| CCPA | 2020 | 판매·삭제 권한, 표준 응답 | 위반당 7,500 USD |
| 개인정보보호법 (한국) | 2020 개정 | 가명·익명처리 표준, 정보주체 권리 | 매출의 3% |
| BCBS 239 | 2013 발효 | 리스크 데이터 집계 표준 | 감독 조치, 자본 규제 |
| EU AI Act | 2024 발효 | 학습 데이터 품질·문서화 | 매출의 7% |
| ISO/IEC 42001 | 2023 발효 | AI 관리 시스템 | 인증 상실 |

## 1.3 표준화 부재의 비용

데이터 표준화 부재는 조직에 다층적 비용을 부과한다. 이 비용은 종종 명시적으로 인식되지 않고 '그림자 비용(shadow cost)'으로 존재한다. Gartner (2021) 의 조사에 따르면 **품질이 낮은 데이터로 인한 조직 평균 손실은 연간 1,290만 달러**에 달한다. 이 비용은 네 가지 범주로 분해할 수 있다.

### 1.3.1 운영 비용 (Operational Cost)

표준화되지 않은 데이터는 매일의 업무 운영에서 지속적으로 비용을 발생시킨다.

- 데이터 준비에 소요되는 시간 — 데이터 사이언티스트가 업무 시간의 45~80%를 데이터 정제에 사용
- 반복 작업 — 부서마다 유사한 표준화 작업을 독립적으로 수행
- 통합 프로젝트 지연 — 시스템 간 데이터 매핑에 예상보다 3~5배 시간 소요
- 품질 이슈 대응 — 데이터 오류로 인한 재작업, 고객 문의 대응
- 교육 비용 — 부서별 상이한 용어·정의로 인한 신규 인력 교육 부담

> **i** 사례: 부서별 상이한 '고객' 정의의 대가
>
> 한 국내 대형 제조사는 마케팅·영업·고객서비스·재무 부서가 각기 다른 '고객' 정의를 사용하고 있었다. 그 결과 전사 CRM 통합 프로젝트가 원 계획의 3배(예산 45억, 기간 26개월) 소요되었으며, 이 중 60%가 데이터 표준화 작업이었다. 프로젝트 전에 표준화를 선행했다면 총 비용의 절반이 절감되었을 것으로 추정되었다.

### 1.3.2 의사결정 비용 (Decision Cost)

표준화되지 않은 데이터는 잘못된 의사결정을 야기하여 훨씬 큰 비용을 발생시킨다. 이는 종종 정량화하기 어렵지만 실제 영향은 운영 비용보다 훨씬 크다.

- 잘못된 KPI — 부서별 다른 정의의 KPI로 인한 잘못된 경영 판단
- 고객 뷰 왜곡 — 마케팅 캠페인 실패, 개인화 서비스 부실

- 재고·수요 예측 오류 — 재고 과다·부족으로 인한 손실
- 리스크 판단 오류 — 신용·사기 리스크 평가의 정확도 저하
- 규제 보고 오류 — 잘못된 규제 보고로 인한 감독 조치

### 1.3.3 규제·법적 비용 (Regulatory Cost)

규제 강화 시대에 데이터 표준화 부재는 직접적인 규제 위반 리스크로 전환된다. 이는 벌금뿐 아니라 브랜드 손상, 사업 라이선스 상실 등 광범위한 영향을 미친다 (European Parliament 2024).

- 규제 벌금 — GDPR 위반 시 매출의 최대 4%
- 규제 대응 인력 비용 — 감사 대비 데이터 준비의 반복 작업
- 소송 리스크 — 개인정보 유출, 잘못된 자동화 의사결정
- 보험료 상승 — 사이버·데이터 보험료의 급격한 상승
- 사업 기회 상실 — 규제 미준수로 인한 신규 시장 진입 제약

### 1.3.4 전략적 기회 상실 (Strategic Opportunity Cost)

가장 큰 비용은 종종 눈에 보이지 않는 '기회 상실'이다. 데이터 표준화 부재는 조직이 새로운 사업 기회, 혁신, 경쟁 우위를 만들지 못하게 한다.

- AI/ML 도입 지연 — 학습 데이터 준비 부담으로 인한 도입 실패
- 데이터 파트너십 불가 — 외부와의 데이터 교환 어려움
- M&A 통합 지연 — 인수합병 후 데이터 통합에 수 년 소요
- 디지털 트랜스포메이션 실패 — 근본 데이터가 준비되지 않아 실패
- 시장 대응 속도 저하 — 시장 변화에 따른 신속한 분석 불가

> ! 4대 비용의 상대적 크기
>
> 연구에 따르면 4대 비용의 상대적 비중은 다음과 같다 (조직별 편차 있음):
>
> - **운영 비용**: 20% (가장 가시적, 그러나 상대적으로 작음)
> - **의사결정 비용**: 30% (정량화 어렵지만 실질 영향 큼)
> - **규제·법적 비용**: 15% (산업·규제 노출도에 따라 다름)
> - **전략적 기회 상실**: 35% (가장 큰 그러나 가장 안 보이는 비용)
>
> 경영진 소통 시 이 4대 비용을 모두 언급해야 한다. 운영 비용만 강조하면 표준화 투자가 '비용 절감 프로젝트'로 축소되고, 전략적 기회 상실을 함께 언급하면 '미래 성장 투자'로 자리매김된다.

## 1.4 데이터 표준화의 가치

표준화가 창출하는 가치는 앞서 논의한 비용의 정반대이자, 그 이상이다. Sebastian-Coleman (2018)은 '표준화는 비용을 절감하는 것이 아니라 새로운 가능성을 여는 것'이라 표현한다. 표준화의 가치는 정량적 지표와 정성적 지표로 측정된다.

### 1.4.1 정량적 가치

| 영역 | 지표 | 일반적 개선 폭 |
|---|---|---|
| 운영 효율 | 데이터 준비 시간 | 60~75% 감소 |
| 운영 효율 | 통합 프로젝트 기간 | 40~60% 단축 |
| 의사결정 | 리포팅 오류율 | 70~85% 감소 |

| 영역 | 지표 | 일반적 개선 폭 |
|---|---|---|
| 의사결정 | 고객 데이터 정확도 | 20~30%p 향상 |
| 규제 대응 | 감사 준비 시간 | 80% 이상 단축 |
| 규제 대응 | 규제 위반 벌금 | 0에 근접 |
| AI/ML | 모델 개발 기간 | 50% 단축 |
| AI/ML | 모델 정확도 | 10~25%p 향상 |

### 1.4.2 정성적 가치

정량적 가치보다 정성적 가치가 더 크고 지속적인 경우가 많다.

#### 1.4.2.1 조직 차원

- 데이터 기반 문화 조성 — 데이터로 대화하는 조직
- 부서 간 협업 향상 — 공통 언어로 인한 커뮤니케이션 개선
- 직원 만족도 향상 — 데이터 씨름 시간 감소, 가치 있는 업무 집중
- 변화 대응력 강화 — 신규 시스템·규제·기회에 신속 대응

#### 1.4.2.2 고객 차원

- 일관된 고객 경험 — 모든 채널에서 동일한 고객 이해
- 개인화 서비스 향상 — 정확한 고객 데이터 기반 맞춤 서비스
- 빠른 문제 해결 — 정확한 정보로 신속한 응대
- 신뢰 강화 — 개인정보 보호와 투명성

#### 1.4.2.3 전략적 차원

- 새로운 사업 모델 — 데이터를 자산으로 활용한 신규 사업
- M&A 시너지 실현 — 신속한 데이터 통합
- 파트너십 확대 — 외부 데이터 교환 능력
- 규제 리더십 — 규제 강화 시 오히려 기회로 활용

### 1.4.3 ROI 사례

실증 연구에 따르면 데이터 표준화 프로젝트의 ROI는 매우 높다. Harvard Business Review (2023) 조사에 따르면, **성숙한 데이터 관리 조직은 그렇지 않은 조직 대비 매출 성장률이 30~40% 높고, 순이익률은 2~3배**에 이른다. MIT Sloan CISR (2021) 의 종단 연구는 이러한 상관관계가 인과관계임을 시사한다.

## 1.5 표준화 도입의 도전과 실패 원인

데이터 표준화의 가치가 명확함에도 불구하고, 많은 조직에서 표준화 프로젝트가 실패하거나 정착하지 못한다. Ladley (2019) 은 데이터 관리 프로젝트의 약 **60~70%가 최초 목표를 달성하지 못한다**고 지적한다. 그 원인은 대체로 다음 5가지로 분류된다.

### 1.5.1 실패 원인 5대 패턴

#### 1.5.1.1 원인 1: 경영진 후원 부족

데이터 표준화는 필연적으로 부서 간 권한 재조정, 프로세스 변경, 예산 투입을 수반한다. 경영진의 강력한 후원 없이는 어떤 부서도 자발적으로 협력하지 않는다. 'IT 프로젝트'로 인식되면 실패는 예정된 것이다. 데이터 표준화는 **'전사 비즈니스 프로세스 개선 프로젝트'**로 자리매김되어야 한다.

#### 1.5.1.2 원인 2: 완벽주의의 함정

'완벽한 표준 체계'를 처음부터 만들려는 시도는 거의 항상 실패한다. 완벽한 표준 정의에 6개월~1년을 소비하는 사이 조직은 표준화의 가치를 잃고 프로젝트 자체가 좌초한다. 최소 실용 표준(Minimum Viable Standard)으로 시작해 실전에서 진화시키는 접근이 성공률이 훨씬 높다.

#### 1.5.1.3 원인 3: 도구 중심 접근

'비싼 도구를 사면 표준화가 될 것'이라는 착각은 흔한 함정이다. 도구는 이미 정의된 표준화 활동을 자동화·확장하는 수단이지, 표준화 자체를 만들어주지 않는다. 조직·프로세스·문화 정비 없이 도구만 도입한 조직은 '비싼 데이터베이스만 가진 조직'이 된다.

#### 1.5.1.4 원인 4: 문화적 저항 무시

부서마다 '우리 데이터는 우리가 안다'는 자부심이 있다. 이는 정당한 것이지만, 전사 표준화 관점에서는 저항의 원천이 된다. 문화적 저항을 무시하고 '올바른 표준'을 강제하려 하면 반드시 저항에 부딪힌다. 변화 관리와 조직 정치에 대한 이해가 기술적 접근만큼이나 중요하다.

#### 1.5.1.5 원인 5: 조기 성과 부재

표준화 프로젝트는 '눈에 보이는 성과'가 나오기까지 최소 6개월 이상 걸린다. 그 사이 경영진의 관심이 식고 예산이 삭감되는 경우가 많다. 초기부터 작지만 가시적인 성과(Quick Win)를 만들어 지속적으로 프로젝트 정당성을 유지해야 한다.

### 1.5.2 성공 조건

반대로 성공한 조직들의 공통 요인을 분석한 MIT Sloan CISR (2021) 의 종단 연구는 다음 5개 요인을 지목한다.

1. **이사회·CEO의 명시적 후원** — CDO의 실질적 권한 부여
2. **비즈니스 가치 중심 접근** — IT 프로젝트가 아닌 비즈니스 프로젝트
3. **점진적 확산** — 작은 파일럿에서 확장
4. **변화 관리 투자** — 교육·소통·문화 조성에 상당한 자원 투입
5. **장기적 관점** — 3~5년의 여정으로 인식

## 1.6 이 책의 구조와 활용 방법

### 1.6.1 본서의 구성

본서는 데이터 표준화를 다음 4부 구성으로 다룬다.

| 부 | 주제 | 포함 챕터 |
|---|---|---|
| 1부 | 데이터 표준화 개요 | 1장 필요성, 2장 핵심 개념, 3장 주요 프레임워크 |
| 2부 | 표준화 방법론 | 4장 데이터 모델링, 5장 메타데이터, 6장 품질, 7장 마스터 데이터 |

| 부 | 주제 | 포함 챕터 |
|---|---|---|
| 3부 | 실무 적용 | 8장 구현 전략, 9장 거버넌스, 10장 도구, 11장 사례 |
| 4부 | 고급 주제 | 12장 AI·ML, 13장 클라우드, 14장 국제 표준 |
| 5부 | 디플래닉스 방법론 | 15장 디플래닉스 방법론 개요, 16장 KDSM 데이터 모델링, 17장 KDSM 메타데이터 관리, 18장 KDSM 데이터 품질 관리 |

부록으로 A. 용어집, B. 표준 문서 템플릿, C. 조직 진단 체크리스트를 제공한다.

### 1.6.2 독자 유형별 활용 방법

| 독자 유형 | 권장 읽기 순서 |
|---|---|
| CDO/CIO/경영진 | 1장 → 3장 → 8장 → 11장 → 15장 → (관심 도메인) |
| 데이터 아키텍트 | 1~3장 → 4~5장 → 7장 → 10장 → 13장 → 14~15장 |
| 데이터 스튜어드 | 1~2장 → 5~7장 → 9장 → 11장 → 15장 → 17장 |
| 데이터 엔지니어·과학자 | 1~2장 → 4~6장 → 10장 → 12~13장→ 15~17장 |
| 규제·컴플라이언스 담당자 | 1장 → 3장 → 9장 → 14장 → (해당 산업 사례) |
| 학생·자격증 준비자 | 1~14장 |

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **데이터 표준화는 조직의 언어 통일** — 데이터가 사람과 시스템 사이에서 오해 없이 이해·교환·활용될 수 있도록 하는 정책·프로세스·기술의 총체이다. 이는 단순한 IT 활동이 아닌 전사 비즈니스 프로세스 개선 활동이다.
2. **오늘날 특별히 중요한 4대 배경** — 데이터 폭증(2025년 181ZB), 시스템 다양성 폭발(평균 200+ SaaS), AI·ML의 부상(실패의 92%가 데이터 문제), 규제 강화(GDPR, EU AI Act 등)가 표준화의 시급성을 높이고 있다.
3. **표준화 부재의 4대 비용** — 운영 비용(20%), 의사결정 비용(30%), 규제·법적 비용(15%), 전략적 기회 상실(35%)의 총 손실은 조직 평균 연 1,290만 달러에 달한다. 가장 큰 비용은 눈에 보이지 않는 '기회 상실'이다.
4. **표준화의 정량·정성적 가치** — 데이터 준비 시간 60~75% 감소, 통합 프로젝트 40~60% 단축, AI 모델 정확도 10~25%p 향상 등 정량적 가치와 조직 문화·고객 경험·전략적 기회의 정성적 가치가 결합된다.
5. **표준화 성공의 조건** — 경영진 후원, 비즈니스 가치 중심 접근, 점진적 확산, 변화 관리 투자, 장기적 관점의 5요인이 성공의 핵심이다. 완벽주의·도구 중심·문화 무시가 흔한 실패 원인이다.
6. **ROI의 명확성** — 성숙한 데이터 관리 조직은 매출 성장률 30~40% 높고, 순이익률은 2~3배에 이른다. 이 상관관계는 인과관계로 해석할 수 있는 증거가 축적되고 있다.
7. **본서의 구성과 활용** — 4부 14장 + 부록 3개로 구성되어 CDO, 아키텍트, 스튜어드, 엔지니어, 규제 담당자 등 다양한 독자에게 맞춤 학습 경로를 제시한다.

# 다음 학습 가이드

## 본 장의 심화 학습

- Sebastian-Coleman (2018) — Navigating the Labyrinth — 경영진 관점의 데이터 관리
- Ladley (2019) — Data Governance — 표준화 도입의 조직·문화 접근
- Redman (2016) — Getting in Front of Data — 품질과 표준화의 비즈니스 케이스
- DAMA International (2017) — DAMA-DMBOK2 — 데이터 관리의 표준 지식 체계

## 다음 장 예고

제 2장 '핵심 개념과 용어'에서는 본서 전반에 걸쳐 사용되는 핵심 용어들 — DIKW 계층, 데이터의 4대 유형, 3계층 모델, 표준 vs 표준화 vs 정규화 vs 조화, 데이터 요소, 메타데이터, 라이프사이클, 계보, 마스터 데이터, 스튜어드십 등 — 을 정확한 정의와 함께 다룬다. 특히 조직에서 이러한 용어들이 어떻게 혼동되고, 어떻게 일관되게 사용되어야 하는지의 실무적 관점을 제공한다.

# 연습 문제

1. 여러분 조직에서 발생한 데이터 표준화 부재로 인한 사례를 1개 이상 찾아, 4대 비용(운영·의사결정·규제·전략적 기회)의 어느 범주에 해당하는지 분석하시오.
2. 여러분 조직의 데이터 관리 성숙도를 5단계 중 어디에 위치시킬 수 있는가? 그 근거는 무엇인가?
3. 데이터 표준화 프로젝트를 경영진에게 제안한다면, 어떤 3가지 비즈니스 케이스(운영·의사결정·전략)를 제시할 것인가?

CHAPTER 2

# 핵심 개념과 용어

## 2.1 개요

모든 학문과 실무 분야는 그 분야만의 **언어(vocabulary)**를 가진다. 의학이 의학용어를 갖고, 법학이 법률용어를 갖듯이 데이터 관리에도 고유한 어휘와 개념 체계가 있다. 이 언어를 공유하지 못하면 조직 내 의사소통이 혼란스러워지고, 표준화 노력이 애초에 시작될 수 없다. Sebastian-Coleman (2018)은 Navigating the Labyrinth 에서 '용어의 혼란이 데이터 프로젝트 실패의 첫 번째 원인'이라 지적한다.

본 장은 데이터 표준화를 이해하고 실행하는 데 필요한 핵심 개념과 용어를 정리한다. 이는 단순한 용어집이 아니라, 각 개념이 어떻게 서로 연결되어 데이터 관리의 지식 체계를 형성하는지를 보여주는 개념 지도(concept map)이다. DAMA International (2020) 은 이러한 개념 체계의 표준 참조 자료로 널리 활용된다.

### 2.1.1 용어 표준화의 어려움

데이터 관리 용어는 다음 세 가지 이유로 표준화가 어렵다 (Sebastian-Coleman 2018; DAMA International 2017).

1. **역사적 축적** — 여러 세대의 기술이 각자 용어를 남김 (예: '테이블'과 '엔티티'와 '데이터셋')
2. **영역 간 중첩** — 데이터 아키텍처, 모델링, 품질이 유사 용어를 다르게 사용
3. **벤더별 차이** — Oracle·SAP·AWS 등이 자체 용어 체계를 유지

### 2.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 데이터 관련 기본 용어를 정확히 정의하고 구분할 수 있다
- 표준·표준화·정규화·조화의 미묘한 차이를 설명할 수 있다
- 데이터의 4대 유형(정형·반정형·비정형·메타데이터)을 이해한다
- 데이터의 3계층 개념(개념·논리·물리)을 활용할 수 있다

- 데이터 관리의 주요 관계 개념(라이프사이클, 계보, 도메인)을 파악한다

## 2.2 데이터의 본질

### 2.2.1 데이터·정보·지식·지혜의 계층 (DIKW)

데이터(Data)라는 용어를 정확히 이해하려면 그 상위 개념들과의 관계를 파악해야 한다. 정보과학에서 오랫동안 사용되어온 **DIKW 피라미드(Data-Information-Knowledge-Wisdom)**는 이 관계를 시각화한다. Ackoff (1989) 가 1989년 논문에서 처음 명시적으로 제시했으며, 이후 정보학·지식경영의 기초 프레임워크가 되었다.

| 계층 | 정의 | 예시 |
|---|---|---|
| Data (데이터) | 가공되지 않은 사실·기호·측정치 | '27', '홍길동', '2024-11-15' |
| Information (정보) | 맥락과 의미가 부여된 데이터 | '홍길동의 나이는 27세이다' |
| Knowledge (지식) | 패턴·관계·경험이 축적된 정보 | '27세는 주 고객층이다' |
| Wisdom (지혜) | 지식을 바탕으로 한 판단·통찰 | '27세를 겨냥한 캠페인을 하자' |

> **i** DIKW의 한계와 재해석
>
> DIKW 피라미드는 유용한 개념 도구이지만, Rowley (2007) 의 논문은 이 계층이 지나치게 단순화된 모델이라고 비판한다. 특히 AI 시대에는 데이터에서 직접 지식이 추출되기도 하고, 정보와 지식의 경계가 흐려지기도 한다. 실무에서는 이 계층을 절대적 위계가 아닌 '가치 증가의 방향'으로 이해하는 것이 유용하다.

### 2.2.2 데이터의 4대 유형

데이터는 구조화 정도에 따라 네 가지 유형으로 분류된다. 각 유형은 다른 저장·처리·표준화 방식을 요구한다 (Reis and Housley 2022).

| 유형 | 특성 | 예시 | 비중 추정 |
|---|---|---|---|
| 정형(Structured) | 미리 정의된 스키마, 관계형 DB | 고객·주문·재고 테이블 | 약 20% |
| 반정형(Semi-structured) | 일관된 구조는 있으나 유연 | JSON, XML, 로그, IoT 이벤트 | 약 20% |
| 비정형(Unstructured) | 고정 구조 없음 | 문서, 이미지, 음성, 영상 | 약 55% |
| 메타데이터(Metadata) | 다른 데이터를 설명하는 데이터 | 스키마, 태그, 계보 정보 | 약 5% |

IDC (2023) 의 조사에 따르면 2025년까지 전 세계 데이터의 약 80%가 비정형 데이터가 될 것으로 예상된다. 이는 텍스트·이미지·영상 데이터를 다루는 표준화의 중요성이 급격히 증가하고 있음을 의미한다. 특히 대규모 언어 모델(LLM)의 등장으로 비정형 데이터의 표준화·품질 관리가 새로운 국면을 맞고 있다.

### 2.2.3 데이터의 3계층: 개념·논리·물리

데이터 모델링의 고전적 개념인 3계층 모델은 ANSI/X3/SPARC (1975) 가 제안한 3-Schema Architecture에서 유래한다. 이후 데이터 모델링뿐 아니라 데이터 표준화 전반에서 활용되는 핵심 프레임워크가 되었다.

| 계층 | 설명 | 관심사 | 산출물 |
|---|---|---|---|
| 개념(Conceptual) | 비즈니스 관점의 실체와 관계 | '무엇'을 관리하는가 | 개념 데이터 모델, 용어 사전 |
| 논리 (Logical) | 기술 독립적인 구조 | '어떻게' 구성되는가 | 논리 데이터 모델, ERD |
| 물리 (Physical) | 특정 DBMS·저장 기술 구현 | '어디서·어떻게' 저장되는가 | 물리 스키마, 인덱스 설계 |

> 3계층이 표준화에 주는 시사점
>
> 각 계층은 서로 다른 표준화 대상과 이해관계자를 가진다:
>
> - **개념 계층**: 비즈니스 용어 표준 → 비즈니스 부서 주도
> - **논리 계층**: 데이터 요소 표준 → 데이터 아키텍트 주도
> - **물리 계층**: 명명 규칙·형식 표준 → 데이터베이스 관리자 주도
>
> 계층별 표준이 서로 연결되어야 조직 차원의 표준화가 실현된다.

## 2.3 표준화 관련 핵심 용어

### 2.3.1 표준 vs 표준화 vs 정규화 vs 조화

'표준화'와 유사한 여러 용어가 종종 혼용된다. 각 용어는 미묘하지만 중요한 차이를 가진다 (ISO/IEC 2004b).

| 용어 | 영어 | 의미 |
|---|---|---|
| 표준 | Standard | 합의된 규칙, 규격, 기준 |
| 표준화 | Standardization | 표준을 개발·채택·적용하는 활동 |
| 정규화 | Normalization | 중복을 제거하고 데이터를 최적 구조로 배치 (특히 관계형 DB) |
| 조화 | Harmonization | 이질적 표준·시스템 간의 상호운용성 확보 |
| 통합 | Integration | 여러 데이터·시스템을 하나의 관점으로 결합 |

> 실무에서의 구분
>
> - **표준화(Standardization)**: 조직 내부에서 통일된 규칙 채택
> - **조화(Harmonization)**: 여러 조직·표준을 상호운용 가능하게 매핑
> - **예**: 국내 은행이 자체 계좌 코드를 SWIFT 표준과 매핑하는 것은 '조화'

### 2.3.2 표준의 5대 유형

ISO/IEC (2004b) 에 따르면 표준은 강제성과 개발 주체에 따라 다음 5가지 유형으로 분류된다.

| 유형 | 특성 | 예시 |
|---|---|---|
| 강제 표준(Regulatory) | 법률·규제에 의해 준수 의무 | 개인정보보호법, BCBS 239 |
| 국제 표준(International) | ISO/IEC 등 국제기구 표준 | ISO 8000, ISO/IEC 11179 |
| 국가 표준(National) | 국가 표준기관 제정 | KS(한국), ANSI(미국), DIN(독일) |
| 산업 표준(Industry) | 산업 협회·컨소시엄 제정 | HL7 FHIR, CDISC, SWIFT |
| 사실상 표준(De facto) | 관행적으로 널리 채택 | SQL, JSON, Parquet |

### 2.3.3 데이터 요소와 관련 개념

ISO/IEC (2023b) 이 정의하는 **데이터 요소(Data Element)** 개념은 표준화의 최소 단위이다. 이는 다음 개념들과 계층적 관계를 이룬다.

| 개념 | 정의 | 예시 |
|---|---|---|
| Concept (개념) | 관심 대상의 추상적 표현 | '사람의 나이' |
| Data Element (데이터 요소) | 특정 표현 방식이 부여된 개념 | '사람의 나이 (정수, 0~150)' |
| Data Element Value (값) | 데이터 요소의 실제 인스턴스 | 27 |
| Domain (도메인) | 값이 가질 수 있는 집합 | 정수 0~150 |
| Value Set (값 집합) | 미리 정의된 유효값 목록 | 성별: {M, F, X} |

## 2.4 데이터 관리 핵심 개념

### 2.4.1 데이터 라이프사이클

**데이터 라이프사이클(Data Lifecycle)**은 데이터가 생성되어 폐기될 때까지의 여정을 개념화한 모델이다. 표준화는 라이프사이클의 모든 단계에 적용되어야 하며, 각 단계마다 서로 다른 표준화 활동이 요구된다 (DAMA International 2017).

| 단계 | 주요 활동 | 표준화 초점 |
|---|---|---|
| 생성(Create) | 데이터 발생, 수집, 입력 | 형식·명명·유효성 표준 |
| 저장(Store) | 데이터 저장, 인덱싱 | 저장 형식·파티션 표준 |
| 사용(Use) | 조회, 분석, 보고 | 접근·쿼리 인터페이스 표준 |
| 공유(Share) | 내외부 배포, API 제공 | 교환 형식·API 표준 |
| 보관(Archive) | 저비용 장기 보관 | 보관 형식·정책 표준 |
| 폐기(Destroy) | 안전한 폐기 및 삭제 | 폐기 프로세스·감사 표준 |

### 2.4.2 데이터 계보(Data Lineage)

데이터 계보는 **'데이터가 어디서 왔고, 어떻게 변형되었으며, 어디로 흐르는가'**를 추적하는 개념이다. Kimball and Ross (2013) 은 이를 '데이터 감사의 기반'이라 표현한다. 규제 준수, 영향도 분석, 문제 해결에 필수적이다.

#### 2.4.2.1 계보의 3가지 유형

- Business Lineage: 비즈니스 의미 수준의 흐름
- Technical Lineage: 시스템·컬럼 수준의 상세 흐름
- Operational Lineage: 실행 로그·시점을 포함한 흐름

### 2.4.3 데이터 도메인과 서브젝트 영역

데이터 도메인(Data Domain)은 관련된 데이터 요소들의 논리적 그룹이다. 조직의 데이터를 도메인으로 나누는 것은 표준화 대상을 관리 가능한 단위로 조직화하는 기본 방법이다. Inmon (2005) 이 데이터 웨어하우스 이론에서 제시한 '서브젝트 영역(Subject Area)' 개념과 유사하다.

| 도메인 | 주요 실체 | 관련 프로세스 |
|---|---|---|
| 고객(Customer) | 고객, 세그먼트, 접점 | 마케팅, 서비스, 판매 |
| 제품(Product) | 제품, 서비스, 재고 | 제조, 유통, 판매 |
| 재무(Finance) | 계정, 거래, 예산 | 회계, 재무 보고 |
| 인적자원(HR) | 임직원, 직위, 급여 | 채용, 평가, 급여 |
| 운영(Operations) | 주문, 배송, 재고 | 물류, 공급망 |
| 규제(Regulatory) | 보고, 감사, 정책 | 컴플라이언스, 감사 |

### 2.4.4 마스터 데이터·참조 데이터·트랜잭션 데이터

데이터는 성격에 따라 세 가지 유형으로 구분되며, 각기 다른 관리 방식을 요구한다 (Loshin 2010a). 이 구분은 제 7장 마스터 데이터 관리에서 상세히 다루지만, 기본 개념 이해가 필수적이다.

| 유형 | 정의 | 예시 | 관리 특성 |
|---|---|---|---|
| 마스터(Master) | 핵심 비즈니스 실체 | 고객, 제품, 공급업체 | 상대적 안정, MDM 관리 |
| 참조(Reference) | 코드·표준값 집합 | 국가 코드, 통화 코드 | 매우 안정, 외부 표준 준수 |
| 트랜잭션(Transaction) | 비즈니스 이벤트 기록 | 주문, 결제, 로그 | 빈번한 생성, 시간 종속 |

### 2.4.5 메타데이터의 3계층

메타데이터는 '데이터에 관한 데이터'이며, 표준화의 핵심 관리 대상이다. 일반적으로 세 계층으로 분류된다 (Riley 2017).

| 계층 | 정의 | 예시 |
|---|---|---|
| Business Metadata | 비즈니스 의미·규칙·오너십 | 용어 정의, 담당자, 규제 태그 |
| Technical Metadata | 기술적 구조·형식·위치 | 스키마, 데이터 타입, 저장 위치 |

| 계층 | 정의 | 예시 |
|---|---|---|
| Operational Metadata | 실행·처리 이력·상태 | 실행 시간, 오류 로그, 사용 통계 |

## 2.5 데이터 품질과 거버넌스 용어

### 2.5.1 데이터 품질의 정의

데이터 품질은 '의도된 사용 목적에 적합한 정도(fitness for use)'로 정의된다. Juran의 품질 정의를 데이터에 적용한 것으로, Redman (2016) 이 데이터 품질 관리 분야에 확산시켰다. 이 정의의 중요한 함의는 '품질이 사용 맥락에 의존한다'는 점이다.

#### 2.5.1.1 품질 관련 핵심 용어

| 용어 | 정의 |
|---|---|
| Data Quality | 데이터의 사용 목적 적합도 |
| Data Integrity | 데이터의 정확성·완전성·일관성 유지 |
| Data Validation | 정의된 규칙에 부합하는지 검증 |
| Data Verification | 다른 신뢰 소스와 대조 확인 |
| Data Profiling | 데이터 특성을 통계적으로 조사 |
| Data Cleansing | 품질 문제를 식별하고 수정 |

### 2.5.2 거버넌스와 스튜어드십

거버넌스(Governance)와 스튜어드십(Stewardship)은 종종 혼용되지만 명확히 구분되어야 한다. Ladley (2019) 는 이를 '결정과 실행의 차이'로 설명한다.

| 구분 | 거버넌스 | 스튜어드십 |
|---|---|---|
| 역할 | 정책·표준 결정 | 정책·표준 실행 |
| 초점 | '무엇을' 할 것인가 | '어떻게' 실행할 것인가 |
| 주체 | 위원회, 리더십 | 실무 스튜어드 |
| 예시 | 품질 기준 승인 | 일상 품질 관리 |

### 2.5.3 데이터 오너·스튜어드·카스토디안

데이터 관리의 세 가지 핵심 역할은 다음과 같이 구분된다 (Ladley 2019; Plotkin 2020). 이 구분은 특히 마스터 데이터 관리와 데이터 거버넌스에서 중요하다.

- Data Owner (오너): '무엇을' — 비즈니스 정의·품질 기준 결정 (부서장 수준)
- Data Steward (스튜어드): '어떻게' — 일상 운영·품질 관리 (실무 책임자)
- Data Custodian (카스토디안): '어디에' — 기술적 저장·보안 관리 (IT)

## 2.6 최신 개념

### 2.6.1 데이터 메시(Data Mesh)

Dehghani (2022) 가 제안한 Data Mesh 는 중앙집중형 데이터 관리의 한계를 극복하기 위한 새로운 패러다임이다. 네 가지 핵심 원칙에 기반한다.

1. 도메인 중심 데이터 오너십 (Domain-Oriented Ownership)
2. Data as a Product — 데이터를 제품으로 취급
3. 셀프 서비스 데이터 인프라
4. 연합 컴퓨팅 거버넌스

### 2.6.2 데이터 계약(Data Contracts)

마이크로서비스의 API 계약처럼, 데이터에도 생산자·소비자 간 명시적 약속이 필요하다는 인식이 확산되고 있다. **데이터 계약(Data Contract)**은 스키마·SLA·품질·의미를 코드로 관리하는 접근이다 (Data Mesh Learning 2023).

### 2.6.3 데이터 관측성(Data Observability)

**데이터 관측성**은 데이터 시스템의 상태를 지속적으로 모니터링하고 이상을 조기 탐지하는 접근이다. 5개 기둥으로 구성된다: (1) 신선도, (2) 분포, (3) 볼륨, (4) 스키마, (5) 계보 (Moses et al. 2022).

### 2.6.4 시맨틱 레이어(Semantic Layer)

**시맨틱 레이어**는 물리적 데이터와 비즈니스 사용자 사이의 추상화 층이다. 사용자가 물리적 스키마를 알 필요 없이 비즈니스 용어로 데이터를 조회할 수 있게 한다. LookML, dbt Metrics Layer, Cube 등이 대표 도구이다 (Cube Dev 2024).

## 2.7 개념 지도와 관계

### 2.7.1 개념 간의 관계

본 장에서 다룬 개념들은 개별적으로 존재하지 않고 상호 연결된 체계를 이룬다. 다음 지도는 주요 개념 간의 관계를 요약한다.

| 관계 유형 | 예시 |
|---|---|
| 구성 관계 (Composition) | 데이터 요소 → 논리 모델 → 물리 스키마 |
| 지배 관계 (Governance) | 거버넌스 → 표준 → 스튜어드십 → 일상 운영 |
| 평가 관계 (Assessment) | 품질 차원 → KPI → 측정 → 개선 |
| 흐름 관계 (Flow) | 생성 → 저장 → 사용 → 공유 → 폐기 (라이프사이클) |
| 역할 관계 (Roles) | 오너(결정) → 스튜어드(관리) → 카스토디안(구현) |

### 2.7.2 용어 관리 실무

> 팁 1: 조직 용어 사전을 구축하라
>
> 본 장의 용어를 참고하되, 각 조직은 자체 용어 사전(Business Glossary)을 구축해야 한다. 조직 내부의 실제 사용 관행, 규제 요구, 시스템 명명 규칙을 반영한 사전만이 실용적이다. 데이터 카탈로그 도구(Collibra, Atlan, DataHub)가 이를 지원한다.

> 팁 2: 동의어와 다의어를 관리하라
>
> 동일 개념에 여러 이름(동의어), 동일 이름에 여러 의미(다의어)가 존재하는 것은 매우 흔한 문제다. 예: '고객'이 마케팅에서는 '잠재 고객'을, 재무에서는 '지급 이력이 있는 고객'을 의미할 수 있다. 조직 용어 사전은 이러한 관계를 명시적으로 관리해야 한다.

> 팁 3: 국제 표준 용어를 우선하라
>
> 가능한 한 조직 자체 용어보다 국제 표준(ISO (2022b), ISO/IEC (2023b), DAMA-DMBOK) 용어를 우선 채택하라. 이는 외부 시스템·규제·컨설팅과의 상호운용성을 확보한다. 특히 CDMP(Certified Data Management Professional) 등 인증도 이 용어 체계를 사용한다 (DAMA International 2017).

> 주의: 용어의 정치성
>
> 용어 정의는 종종 정치적 이슈가 된다. '고객'의 정의가 마케팅·영업·재무 부서에 따라 다르게 요구되는 것은 각 부서의 이해관계 반영이다. 용어 표준화는 순수 기술 활동이 아니라 부서 간 조정과 합의의 과정임을 인식해야 한다. 강제적 정의보다는 대화를 통한 합의가 지속 가능하다.

## 2.8 사례 연구

### 2.8.1 사례 1: 글로벌 은행의 용어 통합 프로젝트

#### 2.8.1.1 배경

유럽 대형 은행 X사는 40개국에서 사업을 운영하며 20,000개 이상의 데이터 요소를 관리한다. 국가별·부서별로 동일 개념이 다르게 정의되어 규제 보고, 리스크 관리, 마케팅 캠페인에 큰 혼란이 발생했다. 특히 Basel Committee on Banking Supervision (2013) 준수 요구가 명확한 용어 표준의 필요성을 부각시켰다.

#### 2.8.1.2 접근 방법

1. 전사 용어 사전(Enterprise Business Glossary) 구축
2. Collibra 도구 도입 — 워크플로우와 승인 체계
3. 도메인별 용어 위원회 (10개 도메인)
4. 동의어·다의어 명시적 관리
5. ISO/IEC 11179 기반 데이터 요소 정의
6. 각 시스템의 물리 컬럼을 표준 용어에 매핑

#### 2.8.1.3 성과

- 표준 용어 8,500개 확립 (핵심 용어)
- 규제 보고 준비 시간 60% 단축
- 부서 간 데이터 요청 처리 시간 70% 감소
- BCBS 239 규제 감사 우수 등급

#### 2.8.1.4 교훈

용어 표준화는 기술적 작업이 아니라 조직 차원의 합의 프로세스이다. 20,000개 데이터 요소를 동시에 표준화하려는 시도는 실패했을 것이다. X사는 규제 관련 핵심 용어부터 시작하여 점진적으로 확장하는 전략을 사용했다. 특히 도메인별 위원회 구조가 부서 간 이해관계 조정에 결정적이었다.

### 2.8.2 사례 2: 한국 공공기관의 표준 용어 사전

#### 2.8.2.1 배경

한국 정부는 2013년 이후 공공데이터 표준화를 위해 '공통표준용어' 체계를 구축·운영하고 있다. 행정안전부 주도로 개발되었으며, 5만 개 이상의 표준 용어를 정의하고 있다. 각 공공기관이 개별 시스템에서 이를 준수하도록 요구한다 (행정안전부 2023a).

#### 2.8.2.2 주요 특징

- 영역별 표준 용어 (행정, 통계, 지리 등)
- 공식 정의와 대체 용어(동의어) 관리
- 코드 표준화 (예: 표준행정코드)
- 정기 갱신 및 확장
- 공공데이터포털을 통한 공개

#### 2.8.2.3 교훈

국가 차원의 표준 용어 체계는 여러 조직 간 상호운용성 확보에 결정적이다. 특히 공공-민간 데이터 공유가 확산되는 시대에, 이러한 표준은 데이터 유통 인프라의 핵심 요소가 된다. 다만 표준 용어의 정기 갱신과 확장이 지속되지 않으면 실무와의 괴리가 발생할 수 있다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **용어의 근본적 중요성** — 용어의 혼란은 표준화 실패의 첫 번째 원인이다. 공통 어휘 없이는 부서 간 소통도, 시스템 간 통합도 불가능하다. 용어 표준화가 모든 데이터 표준화의 출발점이다.
2. **DIKW 계층의 이해** — 데이터·정보·지식·지혜의 계층은 데이터가 어떻게 가치로 전환되는지의 개념 모델이다. AI 시대에는 이 계층 간의 경계가 흐려지고 있다.
3. **데이터의 4대 유형과 3계층** — 정형·반정형·비정형·메타데이터의 4대 유형은 서로 다른 표준화 방식을 요구한다. 개념·논리·물리의 3계층은 데이터를 이해하고 관리하는 기본 프레임워크이다.
4. **표준화 관련 용어의 정확한 구분** — 표준·표준화·정규화·조화·통합은 유사해 보이지만 명확히 구분되는 개념이다. 실무에서 이 구분이 명확하지 않으면 프로젝트 목표와 접근이 흐려진다.
5. **데이터 라이프사이클과 계보** — 라이프사이클의 각 단계에 다른 표준화가 필요하다. 계보는 규제 준수와 문제 해결의 기반이며, 비즈니스·기술·운영의 세 계층에서 관리된다.
6. **역할의 명확한 구분** — Owner(무엇을), Steward(어떻게), Custodian(어디에)의 세 역할이 명확히 구분되어야 거버넌스가 실질적으로 작동한다.
7. **최신 개념의 등장** — Data Mesh, Data Contracts, Data Observability, Semantic Layer 등 새 개념이 지속적으로 등장한다. 실무자는 이러한 개념의 등장 배경과 실무 함의를 이해해야 한다.

## 다음 학습 가이드

### 2.8.3 필수 참고 문헌


- DAMA International (2020) — The DAMA Dictionary of Data Management
- DAMA International (2017) — DAMA-DMBOK2 (용어 부록)
- Sebastian-Coleman (2018) — Navigating the Labyrinth
- Redman (2016) — Getting in Front on Data


### 2.8.4 관련 국제 표준

- ISO/IEC (2023b) — 메타데이터 개념 프레임워크
- ISO (2022b) — 데이터 품질 용어

- ISO/IEC (2004b) — 표준화 일반 원칙

### 2.8.5 관련 본서 참조 장

- 제 3장 주요 표준화 프레임워크 — 프레임워크별 용어 체계
- 제 4장 데이터 모델링 표준 — 모델링 용어 심화
- 제 5장 메타데이터 관리 — 메타데이터 심화
- 제 7장 마스터 데이터 관리 — 마스터·참조 데이터 심화

## 연습 문제

1. 여러분 조직의 핵심 데이터 도메인 5개를 식별하고, 각 도메인의 주요 데이터 요소를 ISO/IEC 11179 방식(개념·특성·표현)으로 정의하시오.
2. 여러분 조직에서 자주 혼용되는 용어 3~5개를 발견하고, 이를 명확히 구분·정의하시오.
3. Data Mesh와 Data Contracts 같은 최신 개념이 여러분 조직에 적용된다면 어떤 변화가 있을지 논하시오.

CHAPTER 3

# 주요 표준화 프레임워크

## 3.1 개요

데이터 표준화를 조직 차원에서 체계적으로 추진하려면 **'무엇을 어떻게 관리할 것인가'**를 정의하는 뼈대가 필요하다. 이 뼈대가 바로 **표준화 프레임워크(Standardization Framework)**이다. 지난 수십 년간 국제 표준화 기구, 산업 협회, 학계, 정부 기관이 다양한 프레임워크를 개발해왔으며, 각 프레임워크는 서로 다른 관점과 목적을 반영한다. Henderson et al. (2017) 은 DAMA-DMBOK2의 서문에서 '프레임워크는 데이터 관리의 언어를 통일한다'고 표현한다.

프레임워크는 단순히 '체크리스트'가 아니다. 그것은 조직이 데이터에 대해 사고하고 의사결정하는 방식 자체를 형성한다. Sebastian-Coleman (2018) 은 Navigating the Labyrinth 에서 프레임워크의 가치를 세 가지로 정리한다:

1. 공통 어휘와 개념 제공
2. 활동 간 관계와 우선순위 명확화
3. 조직 성숙도의 평가·비교 가능한 기준 제시

### 3.1.1 프레임워크의 분류

데이터 표준화 관련 프레임워크는 목적과 범위에 따라 다음 네 가지 유형으로 분류된다 (DAMA International 2017; Sebastian-Coleman 2018).

| 분류 | 성격 | 대표 사례 |
|---|---|---|
| 일반 관리 프레임워크 | 데이터 관리 전반을 포괄 | DAMA-DMBOK2, DCAM |
| 국제 표준(ISO/IEC) | 국제 합의를 통한 공식 표준 | ISO 8000, ISO/IEC 11179, ISO/IEC 25012 |
| 산업별 표준 | 특정 산업의 데이터 규범 | BCBS 239(금융), HL7 FHIR(의료) |
| 국가·정부 표준 | 공공·정부 부문 표준 | 한국 공공데이터 표준, EU Data Strategy |

### 3.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 주요 프레임워크의 목적·범위·구조를 비교할 수 있다
- DAMA-DMBOK2의 11개 지식 영역과 그 상호관계를 설명할 수 있다
- ISO/IEC (2023b), ISO (2022b), ISO/IEC (2008) 의 핵심 내용을 이해한다
- 산업별 표준(Basel Committee on Banking Supervision (2013), HL7 FHIR, ISO (2022c) 등)의 적용 대상과 요구사항을 파악한다
- 조직 상황에 맞는 프레임워크 조합을 설계할 수 있다

## 3.2 일반 관리 프레임워크

### 3.2.1 DAMA-DMBOK2 (Data Management Body of Knowledge)

DAMA International이 발간한 DAMA International (2017) 은 데이터 관리 분야의 사실상 국제 표준(de facto standard)이다. 1985년 설립된 DAMA는 데이터 관리 전문가들의 국제 협회이며, 초판 DMBOK을 2009년, 2판을 2017년에 출간했다. CDMP(Certified Data Management Professional) 자격증도 이 체계에 기반한다.

#### 3.2.1.1 11개 지식 영역

DAMA-DMBOK2의 핵심은 데이터 관리를 **11개 지식 영역(Knowledge Areas)**으로 구성하고, 그 중심에 '데이터 거버넌스'를 배치한 다이어그램이다. 이는 'DMBOK Wheel' 또는 'Aiken Pyramid'로 알려져 있으며, 모든 데이터 관리 활동이 거버넌스의 지배를 받는다는 원칙을 시각화한다.

| 번호 | 지식 영역 | 핵심 활동 |
|---|---|---|
| 1 | 데이터 거버넌스 | 정책, 표준, 권한, 감사 (중심 영역) |
| 2 | 데이터 아키텍처 | 청사진, 데이터 흐름, 통합 구조 |
| 3 | 데이터 모델링·설계 | 개념·논리·물리 모델 |
| 4 | 데이터 저장·운영 | DBMS, 성능, 백업, 복구 |
| 5 | 데이터 보안 | 접근 통제, 암호화, 규제 준수 |
| 6 | 데이터 통합·상호운용성 | ETL, 실시간 통합, API |
| 7 | 문서·콘텐츠 관리 | 비정형 데이터, 문서 관리 |
| 8 | 참조·마스터 데이터 | MDM, 코드 표준 |
| 9 | 데이터 웨어하우징·BI | DW, 데이터 마트, 리포팅 |
| 10 | 메타데이터 관리 | 비즈니스·기술 메타데이터, 카탈로그 |
| 11 | 데이터 품질 | 품질 차원, 측정, 개선 |

> **i** DAMA-DMBOK2의 강점과 한계
>
> **강점**:
>
> - 포괄적 커버리지 — 데이터 관리의 모든 측면을 다룸
> - 전 세계 데이터 전문가 커뮤니티의 합의
> - CDMP 자격증을 통한 표준 지식 체계
>
> **한계**:
>
> - 방대함 — 초기 도입 시 부담
> - 규제·산업 특수성 부족 — 별도 표준으로 보완 필요
> - 실행 지침보다는 개념적 프레임워크

### 3.2.2 DCAM (Data Management Capability Assessment Model)

EDM Council이 개발한 DCAM은 데이터 관리 역량을 평가하고 개선하기 위한 프레임워크이다. 특히 금융 산업에서 널리 활용되며, Basel Committee on Banking Supervision (2013) 준수의 진단 도구로도 사용된다 (EDM Council 2020).

#### 3.2.2.1 DCAM의 8개 능력 영역

1. 데이터 관리 전략 및 비즈니스 케이스
2. 데이터 관리 프로그램 및 재원
3. 비즈니스·데이터 아키텍처
4. 데이터·기술 아키텍처
5. 데이터 품질 관리
6. 데이터 거버넌스
7. 데이터 오퍼레이션
8. 애널리틱스 관리

각 영역은 **6단계 성숙도(Non-Existent → Enhanced)**로 평가되며, 조직의 현 위치와 목표 수준을 명확히 정의할 수 있다. DAMA-DMBOK2가 '지식 체계'라면 DCAM은 '평가·개선 도구'라는 차이가 있다.

### 3.2.3 CMMI-DMM (Data Management Maturity Model)

CMMI Institute(현 ISACA)가 개발한 DMM은 소프트웨어 개발의 CMMI를 데이터 관리에 적용한 성숙도 모델이다 (CMMI Institute 2019). 25개 프로세스 영역을 5단계 성숙도로 평가한다.

| 단계 | 명칭 | 특징 |
|---|---|---|
| 1 | Performed (수행) | 임시적, 개인 의존 |
| 2 | Managed (관리) | 일부 정의, 반복 가능 |
| 3 | Defined (정의) | 표준 프로세스 확립 |
| 4 | Measured (측정) | 정량 측정, 통제 |
| 5 | Optimized (최적화) | 지속 개선, 혁신 |

## 3.3 국제 표준(ISO/IEC)

국제표준화기구(ISO)와 국제전기기술위원회(IEC)는 데이터 관련 여러 국제 표준을 발간하고 있다. 이들 표준은 국가 간 상호운용성과 규제 준수의 기반이 된다. 데이터 표준화 실무자가 특히 주목해야 할 표

준은 ISO/IEC (2023b), ISO (2022b), ISO/IEC (2008) 이며, 최근에는 AI 관리 시스템 표준인 ISO/IEC (2023a) 도 중요해지고 있다 (ISO 2024).

### 3.3.1 ISO/IEC 11179: 메타데이터 등록소

ISO/IEC (2023b) 시리즈는 **메타데이터 등록소(Metadata Registry, MDR)**의 국제 표준이다. 데이터 요소(Data Element)를 조직 간에 어떻게 정의·기록·공유할 것인가의 표준을 제시한다. 1994년 초판 이후 여러 차례 개정되었으며, 여러 국가 정부의 공공데이터 표준의 기반이 되고 있다.

#### 3.3.1.1 ISO/IEC 11179의 6개 파트

| 파트 | 주제 | 핵심 내용 |
|---|---|---|
| Part 1 | Framework | 전체 프레임워크와 개념 |
| Part 2 | Classification | 데이터 요소의 분류 체계 |
| Part 3 | Registry metamodel | 등록소 메타모델과 속성 |
| Part 4 | Formulation of definitions | 정의 작성 규칙 |
| Part 5 | Naming and identification | 명명 규칙과 식별자 |
| Part 6 | Registration | 등록 절차와 관리 |

#### 3.3.1.2 데이터 요소 3계층

ISO/IEC 11179의 핵심 개념은 데이터 요소를 세 계층으로 분해하는 것이다.

- Object Class: 관심 대상 (예: 사람, 조직)
- Property: 대상의 특성 (예: 이름, 나이, 주소)
- Representation: 표현 방식 (예: 문자열, 정수, 코드)

예를 들어 '사람의 나이(만 나이, 정수)'라는 데이터 요소는 'Object Class = 사람, Property = 나이, Representation = 정수(0~150)'로 정의된다. 이러한 분해는 서로 다른 시스템 간 의미적 상호운용성을 확보하는 핵심이다.

### 3.3.2 ISO 8000: 데이터 품질

ISO (2022b) 은 데이터 품질 관리를 위한 국제 표준 시리즈이다. 마스터 데이터와 트랜잭션 데이터의 품질에 대한 요구사항, 측정 방법, 관리 프로세스를 다룬다. 데이터 품질 관리를 다루는 대표적인 국제 표준으로, 조직의 데이터 품질 인증 기준으로도 활용된다.

#### 3.3.2.1 주요 파트

| 파트 그룹 | 주제 | 적용 범위 |
|---|---|---|
| 1~99: General | 개념, 용어, 정보 품질 | 전 영역 |
| 100~199: Master Data | 마스터 데이터 품질 | MDM |
| 200~299: Transaction Data | 트랜잭션 데이터 품질 | 일반 업무 데이터 |
| 300~399: Product Data | 제품 데이터 품질 | 제조, 유통 |

#### 3.3.2.2 ISO 8000-61: 데이터 품질 관리 프로세스

ISO 8000-61은 데이터 품질 관리를 위한 조직 차원의 프로세스 표준이다. Deming의 PDCA 사이클을 데이터 품질에 적용한 것으로, 다음 활동을 정의한다.

- Data quality planning
- Data quality control
- Data quality assurance
- Data quality improvement

### 3.3.3 ISO/IEC 25012: 데이터 품질 모델

ISO/IEC (2008) 는 ISO/IEC 25000 시리즈(SQuaRE)의 일부로, 데이터 품질을 15개 특성으로 정의한다. ISO (2022b) 이 '프로세스'에 초점을 둔다면, 이 표준은 '측정 모델'에 초점을 둔다.

#### 3.3.3.1 15개 품질 특성

| 분류 | 특성 |
|---|---|
| 고유 특성(Inherent) | 정확성, 완전성, 일관성, 신뢰성, 시기 적절성 |
| 고유+시스템 종속 | 접근성, 준수성, 기밀성, 효율성, 정밀성, 추적성, 이해성 |
| 시스템 종속 (System-dependent) | 가용성, 이식성, 복구성 |

> ISO 8000 vs ISO/IEC 25012
>
> 두 표준은 상호 보완적이며, 실무에서는 함께 활용하는 것이 효과적이다:
>
> - ISO (2022b) = 관리 프로세스 (Deming의 PDCA)
> - ISO/IEC (2008) = 측정 모델 (품질 특성 정의)
> - **실무 접근**: ISO/IEC (2008) 로 '무엇을 측정할지'를 정하고, ISO (2022b) 으로 '어떻게 관리할지'를 정의

### 3.3.4 ISO/IEC 42001: AI 관리 시스템

2023년 12월 발효된 ISO/IEC (2023a) 은 인공지능 관리 시스템의 국제 표준이다. AI 시스템의 학습·평가·배포·모니터링에 대한 조직 차원의 관리 체계를 요구한다. 특히 학습 데이터의 품질·거버넌스·문서화 요구사항이 포함되어 있어, AI 시대의 데이터 표준화와 밀접한 관련이 있다.

- AI 시스템 요구사항 문서화
- 학습·검증 데이터의 표준
- 데이터 준비·라벨링 프로세스
- AI 리스크 평가와 완화
- 정기 감사와 지속 개선

## 3.4 산업별 표준

각 산업은 고유한 데이터 요구사항과 규제 환경을 가진다. 이에 대응하여 산업별 전문 표준이 개발되었으며, 이들은 국제 표준을 산업 맥락에 맞게 구체화한다.

### 3.4.1 금융: BCBS 239

바젤 은행감독위원회(BCBS)가 2013년 발표한 '효과적인 리스크 데이터 집계 및 리스크 보고 원칙' Basel Committee on Banking Supervision (2013) 은 글로벌 시스템적 중요 은행(G-SIB)이 준수해야 할 14개 원칙을 정의한다. 2008년 금융위기의 교훈으로, 은행이 위기 상황에서 정확하고 신속한 리스크 정보를 집계할 능력을 요구한다.

#### 3.4.1.1 BCBS 239의 14개 원칙 (4개 영역)

| 영역 | 원칙 수 | 핵심 내용 |
|---|---|---|
| I. 전반적 거버넌스·인프라 | 2 | 거버넌스, IT 아키텍처 |
| II. 리스크 데이터 집계 능력 | 4 | 정확성, 완전성, 적시성, 적응성 |
| III. 리스크 보고 실무 | 5 | 정확성, 포괄성, 명확성, 빈도, 배포 |
| IV. 감독 검토·조치 | 3 | 감독, 시정, 협력 |

### 3.4.2 금융: ISO 20022

ISO (2022c) 는 금융 서비스 메시지의 국제 표준이다. SWIFT, 지급 시스템, 증권 결제, 카드 등 금융 인프라의 표준 메시지 형식으로 자리잡고 있으며, 2025년까지 전 세계 주요 지급 시스템의 ISO 20022 마이그레이션이 완료될 예정이다.

- XML/JSON 기반 구조화된 메시지
- 풍부한 정보(remittance data) 포함 가능
- SWIFT MT 메시지의 후속 표준
- 한국 오픈뱅킹·마이데이터에도 적용

### 3.4.3 의료: HL7 FHIR

Health Level 7의 FHIR (Fast Healthcare Interoperability Resources)는 의료 정보 교환을 위한 최신 국제 표준이다 (HL7 International 2023). REST API와 JSON을 기반으로 설계되어 웹 개발자에게 친숙하며, 현대적 의료 시스템 통합의 표준이 되고 있다. 미국 21st Century Cures Act, EU EHDS(European Health Data Space) 등이 FHIR을 공식 표준으로 채택했다.

#### 3.4.3.1 FHIR의 특징

- Resource: 환자·관찰·처방 등 표준 데이터 단위
- RESTful API로 접근
- JSON, XML, RDF 다중 형식 지원
- Profile로 국가·기관별 확장
- SMART on FHIR: 앱 통합 표준

### 3.4.4 의료: CDISC

CDISC (2023b) (Clinical Data Interchange Standards Consortium)는 임상시험 데이터의 표준화 기구이다. FDA가 신약 승인 신청 시 CDISC 표준(SDTM, ADaM) 준수를 요구하고 있다.

- CDASH: 데이터 수집 표준
- SDTM: 표준화된 임상시험 데이터 모델
- ADaM: 분석 데이터셋 모델
- SEND: 비임상 데이터 표준

### 3.4.5 제조: ISA-95

ISA (2018b) 는 기업 정보 시스템과 제어 시스템(공장 자동화) 간의 통합을 위한 국제 표준이다. ANSI/ISA-95 및 IEC 62264로 발간되었으며, 스마트 팩토리, IIoT, MES 표준의 기반이다.

#### 3.4.5.1 ISA-95 5계층 모델

| 레벨 | 명칭 | 예시 |
|---|---|---|
| Level 4 | Business Planning & Logistics | ERP, 계획 시스템 |
| Level 3 | Manufacturing Operations | MES, 생산 실행 |
| Level 2 | Monitoring & Control | SCADA, HMI |
| Level 1 | Sensing & Manipulation | PLC, 센서·액추에이터 |
| Level 0 | Physical Process | 물리적 생산 공정 |

## 3.5 국가·정부 프레임워크

### 3.5.1 대한민국 공공데이터 표준

한국은 세계에서 가장 체계적인 공공데이터 표준화 정책을 운영하는 국가 중 하나이다. 행정안전부의 '공공데이터의 제공 및 이용 활성화에 관한 법률'(2013)에 근거하여 공공데이터 표준 관리 체계를 운영한다 (행정안전부 2023a). 한국지능정보사회진흥원(NIA)과 한국데이터산업진흥원(K-DATA)이 표준 개발과 확산을 담당한다.

| 구분 | 표준명 | 적용 범위 |
|---|---|---|
| 공통 | 공공데이터 공통표준용어 | 행정 용어 표준 |
| 개방 | 공공데이터 개방 표준 | 51개 분야 표준 데이터셋 |
| 품질 | 공공데이터 품질관리 지침 | 품질 진단·개선 |
| 형식 | 공공데이터 표준 파일 형식 | CSV, JSON, XML |
| 보안 | 공공데이터 보안 지침 | 개인정보·비식별 처리 |

### 3.5.2 EU Data Strategy 및 GAIA-X

유럽연합은 2020년 'A European Strategy for Data' 를 발표하여 유럽 단일 데이터 시장 구축을 추진하고 있다 (European Commission 2020). 주요 구성 요소는:

- Data Governance Act (2022): 공공 데이터 재사용 촉진
- Data Act (2024): IoT 데이터 접근 권한
- European Data Spaces: 9개 산업별 데이터 공간
- GAIA-X: 데이터 주권 클라우드 인프라

### 3.5.3 미국 Federal Data Strategy

미국 연방정부의 데이터 전략은 2020년 발표된 'Federal Data Strategy 2020 Action Plan' 을 기반으로 한다 (U.S. Federal Government 2020). OMB(Office of Management and Budget)가 주도하며, 각 부처의 CDO를 통해 시행된다. 10대 원칙과 40개 실행 과제를 정의한다.

## 3.6 프레임워크 선택 및 조합

### 3.6.1 조직 상황별 선택 지침

어떤 프레임워크도 그 자체로 완벽하지 않다. 조직의 산업·규모·성숙도·규제 환경에 따라 여러 프레임워크를 조합하는 것이 실무적 접근이다 (Sebastian-Coleman 2018).

| 조직 유형 | 권장 조합 |
|---|---|
| 대형 은행 | DAMA-DMBOK2 + BCBS 239 + ISO 20022 + DCAM |
| 의료 기관 | DAMA-DMBOK2 + HL7 FHIR + ISO 8000 + HIPAA |
| 제약 회사 | DAMA-DMBOK2 + CDISC + FDA 21 CFR Part 11 |
| 제조 기업 | DAMA-DMBOK2 + ISA-95 + ISO 9001 |
| 국내 공공기관 | 공공데이터 표준 + ISO/IEC 11179 + ISO 8000 |
| AI 스타트업 | DAMA-DMBOK2 + ISO/IEC 42001 + NIST AI RMF |
| 중견 일반 기업 | DAMA-DMBOK2 + ISO 8000 (핵심만) |

### 3.6.2 프레임워크 도입 절차

1. **조직 진단** — 산업, 규모, 규제 요구, 현재 성숙도
2. **프레임워크 후보 검토** — 후보 3개 이내로 축소
3. **Gap 분석** — 프레임워크 요구사항 vs 현재 상태
4. **우선순위 결정** — 비즈니스 임팩트와 규제 마감 기준
5. **파일럿 적용** — 1~2개 도메인에서 검증
6. **전사 확산** — 성공 사례 기반 확대
7. **성숙도 재평가** — 연 1회 정기 진단

### 3.6.3 실무 팁

> 팁 1: '표준의 표준'을 만들지 말라
>
> 여러 프레임워크를 참고하여 조직만의 완벽한 표준을 처음부터 만들려는 시도는 거의 항상 실패한다. 이미 검증된 프레임워크를 그대로 채택하고, 조직 특수성이 있는 부분만 최소한으로 커스터마이징하라. '표준의 표준'을 만들려다 오히려 표준화가 지연되는 아이러니를 피해야 한다.

> 팁 2: 인증을 활용하라
>
> ISO 인증, DCAM 평가, CMMI-DMM 진단은 외부의 객관적 평가로서 조직 내 표준화 노력의 정당성을 강화한다. 특히 경영진의 관심을 확보하고 예산을 정당화하는 데 유용하다. 다만 인증 자체가 목적이 되면 '체크박스 컴플라이언스'로 전락하므로 주의해야 한다 (Sebastian-Coleman 2018).

> 팁 3: 산업 표준을 우선하라
>
> 일반 프레임워크(DAMA-DMBOK2)와 산업 표준(BCBS 239, FHIR 등)이 충돌하면 산업 표준을 우선하라. 산업 표준은 규제 준수와 상호운용성에 직결되며, 위반 시 구체적 페널티가 따른다. 일반 프레임워크는 그 위에 조직 관리 체계를 얹는 '상위 지침'으로 활용하라.

> **팁 4: 성숙도 모델로 여정을 관리하라**
>
> DCAM(6단계), CMMI-DMM(5단계)와 같은 성숙도 모델은 조직의 현 위치를 객관화하고 목표 수준을 설정하는 데 유용하다. 연 1회 성숙도 평가를 통해 진전을 측정하고, 각 지식 영역별 성숙도의 균형을 관리하라. 특정 영역만 지나치게 성숙하고 다른 영역이 낙후된 '불균형 성숙(uneven maturity)'은 데이터 관리의 흔한 실패 패턴이다.

> **⚠ 주의: 프레임워크는 도구, 목적이 아니다**
>
> 프레임워크의 방대한 요구사항을 모두 충족하려는 시도는 '표준화를 위한 표준화'의 함정에 빠지기 쉽다. 프레임워크는 조직의 비즈니스 목표를 달성하기 위한 도구이지 그 자체가 목적이 아니다. 항상 '이 표준화 활동이 어떤 비즈니스 가치를 창출하는가?'를 묻고, 우선순위를 명확히 하라.

> **⚠ 주의: 프레임워크 업데이트 추적**
>
> 국제 표준과 산업 프레임워크는 지속적으로 개정된다. ISO 표준은 통상 5년 주기로 재검토되며, 산업 표준(FHIR, ISO 20022 등)은 더 자주 업데이트된다. 조직이 채택한 프레임워크의 최신 버전을 추적하고 정기적으로(연 1회) 업데이트하는 프로세스를 갖추어야 한다. 오래된 버전을 고수하면 규제 위반이나 상호운용성 문제를 야기할 수 있다.

## 3.7 사례 연구

### 3.7.1 사례 1: 글로벌 은행의 프레임워크 조합 전략

#### 3.7.1.1 배경

유럽 소재 글로벌 은행 X사는 2016년 Basel Committee on Banking Supervision (2013) 준수 마감을 앞두고 데이터 표준화 프로젝트를 시작했다. 30개국 지사, 200개 이상의 원천 시스템, 규제 보고 요구가 복잡하게 얽혀 있어 단일 프레임워크로는 대응이 불가능했다.

#### 3.7.1.2 접근 방법

1. Basel Committee on Banking Supervision (2013): 규제 준수의 최상위 요구사항
2. DCAM: 조직 역량 진단 및 개선 로드맵
3. DAMA-DMBOK2: 데이터 관리 활동의 공통 어휘
4. ISO (2022c): 지급 시스템 메시징 표준
5. ISO (2022b): 데이터 품질 프로세스

#### 3.7.1.3 성과

- BCBS 239 준수 인정 (2년 앞당김)
- DCAM 성숙도 Level 2 → Level 4
- SWIFT MT → ISO 20022 마이그레이션 완료
- 규제 벌금 리스크 제거 (연 5억 유로 리스크)

#### 3.7.1.4 교훈

복잡한 규제 환경에서는 단일 프레임워크로 대응할 수 없다. 각 프레임워크의 강점을 이해하고 계층적으로 조합하는 전략이 필요하다. 특히 BCBS 239와 같은 구체적 규제가 최상위 요구를 정의하고, 그 아래 DAMA-DMBOK2 같은 일반 프레임워크가 관리 체계를 제공하는 계층 구조가 효과적이다.

### 3.7.2 사례 2: 한국 대형 병원의 FHIR 도입

#### 3.7.2.1 배경

한국의 한 대형 대학병원은 2022년 EMR 시스템의 데이터 상호운용성 확보를 위해 HL7 FHIR을 도입했다. 정부의 EHR(전자건강기록) 인증제 요구와 국제 임상 연구 협력을 위해 국제 표준 채택이 필수적이었다 (HL7 International 2023).

#### 3.7.2.2 도전 과제

- 기존 EMR의 독자적 데이터 모델
- 의료진의 새 표준에 대한 저항
- 한국 의료 코드(KCD, KDC) 매핑 필요
- 개인정보 보호와 데이터 공유의 균형

#### 3.7.2.3 접근 방법

1. Korean FHIR Profile 개발 (KCD 등 국내 표준 매핑)
2. 점진적 마이그레이션 — 신규 데이터부터 FHIR 형식
3. 의료진 교육 및 임상 워크플로우 검증
4. API 게이트웨이로 기존 시스템과 병행
5. 보건복지부 EMR 인증 획득

#### 3.7.2.4 성과

- EMR 인증 3등급 획득
- 국제 임상 연구 3건 데이터 공유
- 환자 App 개발 시간 60% 단축
- 의료 AI 스타트업 3개사와 데이터 협력

#### 3.7.2.5 교훈

국제 표준의 국내 적용은 단순한 채택이 아니라 '국내화(localization)' 작업이 필요하다. 특히 의료 분야는 국가별 코드 체계·규제·워크플로우가 달라 Profile을 통한 확장이 필수이다. FHIR과 같은 열린 표준은 이러한 확장이 용이하도록 설계되어 있어, 다양한 국가·기관 요구에 유연하게 대응할 수 있다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **프레임워크의 필요성** — 프레임워크는 데이터 관리의 공통 어휘, 활동 관계, 성숙도 평가 기준을 제공한다. 조직이 데이터에 대해 사고하고 의사결정하는 방식 자체를 형성하는 역할을 한다.
2. **4대 프레임워크 유형** — 일반 관리 프레임워크(DAMA-DMBOK2, DCAM), 국제 표준(ISO/IEC 시리즈), 산업별 표준(BCBS 239, FHIR 등), 국가·정부 표준의 네 유형이 있으며, 각기 다른 목적과 범위를 가진다.
3. **DAMA-DMBOK2의 중심성** — 11개 지식 영역과 데이터 거버넌스 중심 구조는 데이터 관리의 사실상 표준으로 자리잡았다. CDMP 자격증도 이 체계에 기반한다.
4. **ISO 표준의 국제성** — ISO/IEC (2023b) (메타데이터), ISO (2022b) (품질 관리), ISO/IEC (2008) (품질 모델), ISO/IEC (2023a) (AI)는 국가 간 상호운용성과 규제 준수의 기반이다.
5. **산업별 표준의 구체성** — Basel Committee on Banking Supervision (2013) (금융), HL7 FHIR(의료), CDISC (2023b) (제약), ISA (2018b) (제조) 등 산업 특화 표준은 규제 준수와 실무 상호운용성에 직결된다.
6. **프레임워크의 조합** — 단일 프레임워크로는 조직의 모든 요구를 충족할 수 없다. 산업 표준 + 일반 프레임워크 + 국제 표준을 계층적으로 조합하는 것이 실무적 접근이다.

7. **도구로서의 프레임워크** — 프레임워크는 도구이지 목적이 아니다. 항상 비즈니스 가치를 먼저 정의하고, 프레임워크의 요구사항을 선별적으로 적용해야 한다.

## 다음 학습 가이드

### 3.7.3 필수 참고 문헌

- DAMA International (2017) — DAMA-DMBOK2
- EDM Council (2020) — EDM Council DCAM
- Sebastian-Coleman (2018) — Navigating the Labyrinth

### 3.7.4 주요 국제 표준

- ISO/IEC (2023b) — 메타데이터 등록소
- ISO (2022b) — 데이터 품질
- ISO/IEC (2008) — 데이터 품질 모델
- ISO/IEC (2023a) — AI 관리 시스템

### 3.7.5 주요 산업 표준

- Basel Committee on Banking Supervision (2013) — 금융 리스크 데이터
- ISO (2022c) — 금융 메시징
- HL7 International (2023) — 의료 정보 교환
- CDISC (2023b) — 임상시험 데이터
- ISA (2018b) — 제조 자동화

### 3.7.6 관련 본서 참조 장

- 제 5장 메타데이터 관리 — ISO/IEC 11179 심화
- 제 6장 데이터 품질 관리 — ISO 8000, ISO/IEC 25012 심화
- 제 9장 데이터 거버넌스 체계 — DAMA-DMBOK2 거버넌스
- 제 12장 AI·ML과 데이터 표준화 — ISO/IEC 42001
- 제 14장 국제 표준 심화 — 표준별 상세 분석

## 연습 문제

1. 여러분 조직에 적합한 프레임워크 조합을 제안하시오. 각 프레임워크가 다루는 요구와 조합의 근거를 명시하시오.
2. DAMA-DMBOK2의 11개 지식 영역 중 여러분 조직에서 가장 취약한 영역을 3개 식별하고, 개선 방향을 제시하시오.
3. 산업 표준(예: BCBS 239, FHIR, ISA-95 중 하나)의 요구사항을 조사하고, 여러분 조직 또는 관심 산업에 적용 시 예상되는 도전과 대응 방안을 논하시오.

# II 2부: 표준화 방법론


4 데이터 모델링 표준 . . . . . 45
4.1 개요 . . . . . 45
4.2 핵심 개념 . . . . . 46
4.3 방법론 . . . . . 48
4.4 실무 적용 . . . . . 51
4.5 사례 연구 . . . . . 53

5 메타데이터 관리 . . . . . 57
5.1 개요 . . . . . 57
5.2 핵심 개념 . . . . . 58
5.3 방법론 . . . . . 60
5.4 실무 적용 . . . . . 63
5.5 사례 연구 . . . . . 64

6 데이터 품질 관리 . . . . . 69
6.1 개요 . . . . . 69
6.2 핵심 개념 . . . . . 70
6.3 방법론 . . . . . 72
6.4 실무 적용 . . . . . 74
6.5 사례 연구 . . . . . 76

7 마스터 데이터 관리(MDM) . . . . . 79
7.1 개요 . . . . . 79
7.2 핵심 개념 . . . . . 80
7.3 방법론 . . . . . 82
7.4 실무 적용 . . . . . 83
7.5 사례 연구 . . . . . 85

CHAPTER 4

# 데이터 모델링 표준

## 4.1 개요

데이터 모델링(Data Modeling)은 **현실 세계의 비즈니스를 데이터 구조로 추상화하여 표현하는 활동**이다. Hoberman (2014) 은 데이터 모델을 '데이터에 대한 정확한 시각적 의사소통 도구'로 정의하며, 잘 만들어진 데이터 모델은 비즈니스 요구사항을 정확히 반영하면서 기술적 구현의 청사진 역할을 한다고 강조한다. 데이터 모델링 표준은 이러한 모델링 활동에 일관성을 부여하여 조직 전반에서 통합 가능하고 유지보수 가능한 데이터 자산을 구축하는 기반이 된다.

DAMA International (2017) 은 데이터 모델링을 '데이터 관리의 가장 기본적이고 본질적인 활동'으로 규정하며, 다른 모든 데이터 관리 영역(품질, 보안, 통합, 분석)이 데이터 모델 위에서 작동한다고 본다. 그럼에도 불구하고 많은 조직에서 데이터 모델링은 일관된 표준 없이 개별 프로젝트나 개인의 스타일에 의존하여 수행되는 경우가 많아, 시스템 간 통합과 유지보수에 심각한 비용을 초래한다.

### 4.1.1 데이터 모델링 표준의 필요성

Inmon (2005) 은 그의 저서 Building the Data Warehouse에서 '표준 없는 데이터 모델링은 같은 언어로 다른 말을 하는 것과 같다'고 지적한다. 데이터 모델링 표준이 부재할 때 조직은 다음과 같은 문제를 겪는다.

1. 같은 개체(entity)가 시스템마다 다른 이름으로 정의되어 통합이 어려움
2. 동일한 속성(attribute)에 대해 데이터 타입·길이가 시스템별로 달라 데이터 변환 비용 증가
3. 관계(relationship) 표기 방식이 모델러마다 달라 ERD 해석에 혼선
4. 정규화 수준이 일관되지 않아 데이터 중복과 무결성 문제 발생
5. 신규 모델러가 기존 모델을 이해하는 데 과도한 시간 소요

TDWI (2023b) 의 2023년 조사에 따르면, 데이터 모델링 표준을 체계적으로 운영하는 기업은 그렇지 않은 기업 대비 데이터 통합 프로젝트 기간이 평균 **42% 단축**되었고, 시스템 유지보수 비용이 **35% 절감**되었다[4].

### 4.1.2 본 장의 학습 목표

- 데이터 모델의 3대 추상화 레벨을 구분하고 각 레벨의 표준화 요소를 설명할 수 있다
- ERD, 차원 모델링 등 주요 표기법의 표준 규칙을 적용할 수 있다
- 엔티티·속성·관계 명명 규칙을 ISO 표준에 따라 설계할 수 있다
- 정규화 원칙과 의도적 비정규화의 판단 기준을 이해한다
- OLTP와 OLAP 환경에 맞는 모델링 접근법을 선택할 수 있다

## 4.2 핵심 개념

### 4.2.1 데이터 모델의 3대 추상화 레벨

ANSI/X3/SPARC (1975) 이 제시한 **3-스키마 아키텍처**는 오늘날 데이터 모델링의 기본 프레임워크다. 이 프레임워크는 데이터 모델을 **개념·논리·물리** 세 가지 추상화 레벨로 구분한다.

#### 4.2.1.1 개념 데이터 모델(Conceptual Data Model)

비즈니스 관점에서 핵심 개체와 관계를 표현하는 가장 추상적인 수준의 모델이다. 기술적 세부사항을 배제하고 비즈니스 이해관계자가 검토할 수 있는 수준으로 작성된다. 주요 산출물은 핵심 엔티티(Core Entity)와 주요 관계만을 표현한 단순한 ERD다.

- 대상: 임원, 비즈니스 분석가, 도메인 전문가
- 표현 요소: 주요 엔티티(10~30개), 주요 관계, 비즈니스 규칙
- 제외 요소: 속성 상세, 데이터 타입, 정규화

#### 4.2.1.2 논리 데이터 모델(Logical Data Model)

개념 모델을 상세화하여 모든 엔티티, 속성, 관계, 제약조건을 표현하는 모델이다. 데이터베이스 기술과는 독립적이며, 정규화가 적용된다. 실제 비즈니스 요구사항을 가장 충실하게 반영하는 모델이다.

- 대상: 데이터 아키텍트, 시스템 분석가, 데이터 모델러
- 표현 요소: 모든 엔티티·속성·관계, 기본키·외래키, 정규화, 도메인
- 제외 요소: 인덱스, 파티션, 물리 저장 구조

#### 4.2.1.3 물리 데이터 모델(Physical Data Model)

특정 데이터베이스 시스템에 맞춰 논리 모델을 구체화한 모델이다. 성능 최적화를 위한 비정규화, 인덱스, 파티셔닝 등이 포함된다.

- 대상: DBA, 데이터 엔지니어
- 표현 요소: 테이블·컬럼명, 데이터 타입, 인덱스, 파티션, 저장 옵션
- DBMS 종속적: Oracle, PostgreSQL, MySQL 등 각각 다른 표현

| 특성 | 개념 모델 | 논리 모델 | 물리 모델 |
|---|---|---|---|
| 추상화 수준 | 매우 높음 | 중간 | 낮음 |
| 기술 종속성 | 없음 | 없음 | DBMS별 종속 |
| 엔티티 수 | 10~30개 | 수십~수백 개 | 논리 + $\alpha$ |
| 속성 표현 | 없거나 핵심만 | 전체 | 전체 + 물리속성 |
| 주요 사용자 | 비즈니스 | 분석가/모델러 | DBA/엔지니어 |
| 변경 빈도 | 낮음 | 중간 | 높음 |

### 4.2.2 ER 모델링: 데이터 모델링의 표준 표기법

Chen (1976) 이 발표한 Entity-Relationship Model은 이후 50년 가까이 데이터 모델링의 사실상 표준 표기법이 되어왔다. ERD(Entity-Relationship Diagram)는 다음 세 가지 핵심 요소로 구성된다.

#### 4.2.2.1 엔티티(Entity)

비즈니스에서 관리하고자 하는 사람·사물·개념·사건 등의 객체를 의미한다. 예: 고객, 상품, 주문, 직원. 일반적으로 사각형으로 표현하며, 단수 명사로 명명한다.

#### 4.2.2.2 속성(Attribute)

엔티티가 가진 특성이나 정보를 의미한다. 속성은 다음과 같이 분류된다.

- 단순 속성(Simple): 더 이상 분해되지 않는 속성 (예: 이름, 나이)
- 복합 속성(Composite): 여러 속성으로 구성된 속성 (예: 주소 = 시 + 구 + 동)
- 다중값 속성(Multi-valued): 여러 값을 가질 수 있는 속성 (예: 전화번호)
- 유도 속성(Derived): 다른 속성으로부터 계산되는 속성 (예: 나이 = 현재일자 - 생년월일)
- 식별자(Identifier): 엔티티를 유일하게 식별하는 속성 (예: 고객 ID)

#### 4.2.2.3 관계(Relationship)

엔티티 간의 연관성을 의미한다. 관계는 다음 두 가지 차원으로 분석된다.

- 카디널리티(Cardinality)
  - 일대일(1:1): 한 엔티티가 다른 엔티티의 단 하나의 인스턴스와 관련
  - 일대다(1:N): 한 엔티티가 다른 엔티티의 여러 인스턴스와 관련
  - 다대다(M:N): 양쪽 엔티티 모두 여러 인스턴스와 관련 (연관 엔티티로 해소)
- 선택성(Optionality)
  - 필수(Mandatory): 반드시 관계가 존재해야 함
  - 선택(Optional): 관계가 존재하지 않을 수도 있음

> **i** ER 표기법 변형
>
> - **Chen 표기법**: 원본, 다이아몬드로 관계 표현
> - **Crow's Foot (까마귀발)**: 가장 널리 사용, 카디널리티를 시각적으로 표현
> - **IDEF1X**: 미국 국방부 표준, 정부·국방 분야 사용
> - **Barker (Oracle)**: Oracle Designer에서 사용된 표기법

### 4.2.3 정규화(Normalization)

Codd (1970) 가 제안한 **정규화**는 **데이터 중복을 최소화하고 무결성을 보장하기 위한 체계적 절차**이다. 1NF부터 6NF까지 존재하지만 실무에서는 3NF 또는 BCNF까지 적용하는 것이 일반적이다.

| 정규형 | 정의 | 해결 문제 |
|---|---|---|
| 1NF | 모든 속성이 원자값 | 반복 속성 제거 |
| 2NF | 1NF + 부분 함수 종속 제거 | 복합키 부분 종속 |
| 3NF | 2NF + 이행 함수 종속 제거 | 키 외 속성 간 종속 |
| BCNF | 3NF + 모든 결정자가 후보키 | 3NF 예외 처리 |
| 4NF | BCNF + 다중값 종속 제거 | 독립 다중값 속성 |

| 정규형 | 정의 | 해결 문제 |
|---|---|---|
| 5NF | 4NF + 조인 종속 제거 | 복잡 조인 의존성 |

### 4.2.4 차원 모델링(Dimensional Modeling)

Kimball and Ross (2013) 이 정립한 **차원 모델링**은 **분석 목적의 데이터 웨어하우스를 위한 모델링 기법**이다. 정규화된 OLTP 모델과 달리 의도적으로 비정규화하여 조회 성능과 사용 편의성을 최적화한다.

#### 4.2.4.1 차원 모델의 핵심 구성 요소

**팩트 테이블(Fact Table)**
측정 가능한 비즈니스 이벤트(매출, 주문 등)를 저장하는 테이블이다. 외래키(차원 참조)와 측정값(measure)으로 구성된다.

**차원 테이블(Dimension Table)**
팩트를 분석하는 관점(언제, 어디서, 누가, 무엇을)을 제공하는 테이블이다. 고객 차원, 상품 차원, 시간 차원, 지역 차원 등이 대표적이다.

#### 4.2.4.2 스타 스키마 vs 스노우플레이크 스키마

| 구분 | 스타 스키마 | 스노우플레이크 |
|---|---|---|
| 차원 정규화 | 비정규화 | 정규화 |
| 조인 수 | 적음 | 많음 |
| 조회 성능 | 빠름 | 느림 |
| 저장 공간 | 많음 | 적음 |
| 유지보수 | 쉬움 | 복잡 |

## 4.3 방법론

### 4.3.1 데이터 모델링 표준 수립 절차

데이터 모델링 표준 수립은 단순히 명명 규칙 몇 개를 정하는 작업이 아니라, 조직의 데이터 거버넌스 체계 전반과 맞물려 진행되어야 하는 체계적 활동이다. DAMA International (2017) 의 권고에 따른 6단계 절차는 다음과 같다.

#### 4.3.1.1 Phase 1: 표준 범위 정의

- 적용 대상 시스템·도메인·프로젝트 식별
- 이해관계자 분석 및 영향도 평가
- 기존 모델링 관행 조사 및 갭 분석

#### 4.3.1.2 Phase 2: 표준 문서 작성

- 명명 규칙(엔티티·속성·관계·테이블·컬럼)
- 데이터 타입 및 도메인 표준
- 관계 표기 규칙 (ERD 표기법 선택)
- 정규화 적용 기준
- 모델 문서화 양식

#### 4.3.1.3 Phase 3: 도구 및 환경 표준화

- 표준 모델링 도구 선정 (ERwin, ER/Studio, DA#, MySQL Workbench 등)
- 모델 저장소(Repository) 구축
- 버전 관리 체계 수립

#### 4.3.1.4 Phase 4: 검토 및 승인

- 데이터 거버넌스 위원회 검토
- 주요 이해관계자 의견 수렴
- 표준 공식 승인 및 공표

#### 4.3.1.5 Phase 5: 교육 및 적용

- 데이터 모델러 교육 프로그램 운영
- 표준 적용 가이드 및 체크리스트 보급
- 초기 적용 지원 및 멘토링

#### 4.3.1.6 Phase 6: 운영 및 개선

- 표준 준수도 모니터링
- 정기적 표준 검토 및 개정
- 우수 사례 공유 및 확산

### 4.3.2 엔티티 명명 표준

엔티티 명명은 데이터 모델의 가독성과 일관성에 결정적인 영향을 미친다. ISO/IEC (2015a) 의 명명 원칙과 업계 베스트 프랙티스를 결합한 표준 규칙은 다음과 같다.

#### 4.3.2.1 엔티티 명명 7대 원칙

1. 단수 명사 사용 — '고객들'이 아닌 '고객'
2. 비즈니스 용어 사용 — '테이블1'이 아닌 '주문'
3. 모호한 약어 회피 — 'CUST'가 아닌 'CUSTOMER' (다만 산업 표준 약어는 허용)
4. 일관된 표기법 — 한글/영문, 대소문자 규칙 통일
5. 동음이의 회피 — 같은 이름이 다른 의미로 사용되지 않도록
6. 이음동의 회피 — 같은 의미가 다른 이름으로 사용되지 않도록
7. 최대 길이 제한 — 일반적으로 30자 이내

> **i** 엔티티 명명 예시
>
> - 좋은 명명:
>  - Customer, Product, Order, Employee
>  - 고객, 상품, 주문, 직원
> - 나쁜 명명:
>  - Tbl_Cust, T_001, Customers (복수형)
>  - 고객정보 (이미 엔티티는 정보이므로 불필요)

### 4.3.3 속성 명명 표준

속성 명명은 ISO/IEC (2015a) 가 제시하는 [수식어 + 한정자 + 객체 클래스 용어 + 특성 용어 + 표현어] 구조를 따른다. 다만 실무에서는 [객체 클래스 + 특성 + 표현어]의 단순화된 3단 구조를 가장 많이 사용한다.

> **i 속성 명명 패턴 예시**
>
> 패턴: `[Object Class] + [Property] + [Representation Term]`
>
> 예시:
>
> - Customer + Name + Text → `Customer_Name_Text`
> - Order + Total + Amount → `Order_Total_Amount`
> - Product + Launch + Date → `Product_Launch_Date`
> - Employee + Active + Flag → Employee_Active_Flag

#### 4.3.3.1 표준 표현어(Representation Term)

| 표현어 | 한글 | 의미 | 예시 |
|---|---|---|---|
| ID | 식별자 | 유일 식별값 | Customer_ID |
| Name | 명칭 | 고유 이름 | Product_Name |
| Code | 코드 | 분류 코드 | Country_Code |
| Amount | 금액 | 금전적 값 | Tax_Amount |
| Quantity | 수량 | 수량값 | Stock_Quantity |
| Date | 일자 | 날짜 | Birth_Date |
| DateTime | 일시 | 날짜+시간 | Created_DateTime |
| Number | 번호 | 일련번호 | Phone_Number |
| Text | 텍스트 | 자유 텍스트 | Description_Text |
| Flag | 플래그 | Y/N 또는 1/0 | Deleted_Flag |
| Rate | 비율 | 백분율 또는 분수 | Discount_Rate |
| Percent | 백분율 | 백분율값 | Tax Percent |

### 4.3.4 정규화 적용 기준

정규화는 무조건 높은 수준이 좋은 것이 아니라 **데이터의 사용 패턴과 시스템 목적에 맞춰 선택**해야 한다. Hoberman (2014) 은 '정규화는 종교가 아니라 도구'라고 표현한다.

#### 4.3.4.1 환경별 정규화 권장 수준

| 환경 | 권장 수준 | 근거 |
|---|---|---|
| OLTP (트랜잭션) | 3NF 또는 BCNF | 무결성, 빠른 입력/수정 |
| OLAP (분석) | 스타/스노우플레이크 | 조회 성능, 사용 편의성 |
| 운영 데이터 저장소(ODS) | 2NF~3NF | 원천·분석 사이의 균형 |
| 데이터 레이크 | 원시 형태 유지 | 다양한 분석 요구 대응 |
| 마스터 데이터 | BCNF 이상 | 최고 수준의 무결성 |

**의도적 비정규화(Denormalization)**
성능 향상을 위해 정규화 규칙을 의도적으로 어기는 것을 비정규화라고 한다. 다음 경우에 비정규화를 고려한다.

- 조회 빈도가 매우 높고 데이터 변경이 드문 경우
- 조인 비용이 시스템 성능의 병목인 경우
- 집계 결과를 자주 조회하는 경우 → 집계 컬럼 추가
- 이력 추적이 필요한 경우 → 스냅샷 컬럼 추가

## 4.4 실무 적용

### 4.4.1 데이터 모델링 도구 선정

데이터 모델링 도구는 모델의 작성·관리·공유를 지원하는 핵심 인프라다. Gartner (2024e) 의 시장 가이드에 따른 2024년 기준 주요 주요 도구와 특징은 다음과 같다.

| 구분 | 도구 | 특징 |
|---|---|---|
| 엔터프라이즈 상용 | erwin, ER/Studio, PowerDesigner | 포괄적 기능, 거버넌스 통합 |
| DB 벤더 도구 | Oracle SQL Developer, SQL Server Mgmt Studio | 특정 DBMS 최적화 |
| 오픈소스 | MySQL Workbench, pgModeler, DBeaver | 낮은 비용, 기본 기능 충실 |
| 클라우드 기반 | Lucidchart, dbdiagram.io, SqlDBM | 협업 강점, 웹 기반 |
| 코드 기반 | DBML, PlantUML, Mermaid | Git 통합, 버전 관리 친화적 |
| 국산 | DA#, ERMaster, DBmaestro | 한국어 지원, 국내 표준 친화적 |

#### 4.4.1.1 도구 선정 체크리스트

- 표준 표기법(Crow's Foot, IDEF1X 등) 지원
- 개념·논리·물리 모델 통합 관리
- Forward Engineering (모델 → DDL)
- Reverse Engineering (DDL → 모델)
- 모델 비교(diff) 및 동기화
- 협업 및 버전 관리
- 문서 자동 생성
- 메타데이터 카탈로그 연동

### 4.4.2 모델 검토(Model Review) 체크리스트

데이터 모델은 일단 구축되면 변경 비용이 크므로 사전 검토가 매우 중요하다. Silverston (2008) 은 모델 검토를 '모델의 품질을 결정짓는 가장 중요한 단계'라고 강조한다. 다음 체크리스트를 활용하라.

#### 4.4.2.1 비즈니스 적합성 검토

- 모든 비즈니스 요구사항이 모델에 반영되었는가?
- 주요 비즈니스 규칙이 제약조건으로 표현되었는가?
- 미래 변화 가능성이 고려되었는가?

#### 4.4.2.2 구조적 정합성 검토

- 모든 엔티티에 기본키가 정의되어 있는가?
- 외래키 관계가 누락 없이 표현되었는가?
- M:N 관계가 연관 엔티티로 해소되었는가?
- 정규화 수준이 일관되게 적용되었는가?

#### 4.4.2.3 표준 준수 검토

- 엔티티·속성 명명이 표준을 따르는가?
- 데이터 타입과 도메인이 표준과 일치하는가?
- 표준 표기법이 일관되게 적용되었는가?

#### 4.4.2.4 성능 고려사항 검토

- 주요 조회 패턴이 인덱스로 지원되는가?
- 데이터 증가량을 고려한 파티셔닝 전략이 있는가?
- 의도적 비정규화가 적절히 적용되었는가?

### 4.4.3 실무 팁

> 실무 팁 1: 모델은 비즈니스의 거울이다
>
> Silverston (2008) 은 '데이터 모델을 보면 그 조직의 비즈니스가 보인다'고 말한다. 모델링을 시작하기 전에 **비즈니스 프로세스와 용어를 충분히 이해**하라. 기술적 우아함보다 비즈니스 적합성이 우선이다. 현업 인터뷰, 업무 매뉴얼 검토, 기존 보고서 분석 등에 충분한 시간을 투자해야 한다.

> 실무 팁 2: 단순함을 추구하라
>
> 복잡한 모델은 거의 항상 잘못된 모델이다. 엔티티가 100개를 넘어가면 비즈니스 요구사항을 잘못 분해했을 가능성이 높다. 모델을 단순화하는 것은 모델러의 핵심 역량이다. '더 추가할 것이 없을 때'가 아니라 '더 뺄 것이 없을 때' 모델이 완성된다.

> 실무 팁 3: 표준은 도구로 강제하라
>
> 사람의 의지에만 의존하는 표준은 시간이 지나면 무너진다. **모델링 도구의 검증 기능, CI/CD 파이프라인의 자동 검사, 코드 리뷰** 등을 통해 표준을 자동으로 강제하라. DataOps 관점에서 데이터 모델도 코드처럼 다루어야 한다 (Cataldo and Foote 2022).

> ⚠ 주의: 과도한 정규화의 함정
>
> 정규화는 데이터 무결성을 위한 도구이지 목적이 아니다. 4NF, 5NF까지 강박적으로 적용하다 보면 단순한 조회에도 10개 이상의 조인이 필요한 사용 불가능한 모델이 되기 쉽다. 일반적인 OLTP 시스템에서는 3NF가 충분하다.

> ⚠ 주의: 미래에 대한 과도한 일반화
>
> '언젠가 필요할지 모른다'는 이유로 과도하게 일반화된 모델은 거의 항상 후회를 남긴다. EAV(Entity-Attribute-Value) 모델, 메타 스키마, 동적 컬럼 추가 등은 유연성을 제공하는 듯 보이지만 데이터 무결성, 성능, 가독성 모두를 희생시킨다. 현재의 요구사항에 충실하고, 변경이 필요할 때 리팩토링하라.

## 4.5 사례 연구

### 4.5.1 사례 1: 글로벌 은행 G사의 통합 데이터 모델

#### 4.5.1.1 배경

글로벌 은행 G사는 인수합병으로 인해 12개의 서로 다른 핵심 뱅킹 시스템을 운영하게 되었다. 각 시스템마다 '고객', '계좌', '거래' 등 핵심 개체의 정의가 달라 그룹 차원의 리스크 관리와 규제 보고에 심각한 어려움을 겪고 있었다. Basel Committee on Banking Supervision (2013) 준수를 위해 통합 데이터 모델 수립이 필수였다.

#### 4.5.1.2 도전 과제

- 12개 시스템의 데이터 모델이 모두 상이
- 같은 '고객 ID'가 시스템마다 다른 형식과 의미
- BCBS 239(바젤 리스크 데이터 집계 원칙) 준수 요구
- 기존 시스템을 중단할 수 없는 운영 제약

#### 4.5.1.3 접근 방법

1. ISO (2022c) 금융 메시징 표준을 기반으로 한 표준 데이터 모델 수립
2. 개념 모델 → 논리 모델 → 시스템별 매핑의 3단계 접근
3. 통합 데이터 모델(EDM) 거버넌스 위원회 구성
4. 신규 시스템은 표준 모델 의무 적용, 기존 시스템은 점진적 매핑

#### 4.5.1.4 성과

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 규제 보고서 작성 | 8주 | 2주 |
| 고객 통합 뷰 생성 | 수일 | 실시간 |
| 데이터 변환 매핑 수 | 약 450개 | 약 60개 |
| 신규 시스템 통합 비용 | 기준 100 | 35 |

#### 4.5.1.5 교훈

기존 시스템을 한 번에 표준으로 전환하려 하지 말고, 신규 개발은 표준 모델로 시작하고 기존 시스템은 매핑 레이어로 연결한 후 점진적으로 마이그레이션하는 전략이 효과적이다. 표준화는 단거리 경주가 아닌 마라톤이다.

### 4.5.2 사례 2: 한국 공공기관의 행정정보 표준화

#### 4.5.2.1 배경

행정안전부 and 한국지능정보사회진흥원 (2023) 는 중앙부처와 광역자치단체의 행정정보시스템 간 데이터 호환성을 확보하기 위해 행정정보 데이터 모델 표준화 사업을 추진해왔다. 주요 행정 업무(주민, 부동산, 차량, 사업자 등)에 대한 표준 데이터 모델을 수립하고 보급하는 것이 목적이었다.

#### 4.5.2.2 주요 산출물

- 행정정보 공통표준용어집 (약 6,000여 개 용어)
- 핵심 행정 도메인별 표준 데이터 모델 (약 50개)
- 표준 코드체계 정의서
- 행정정보 데이터 모델링 가이드라인

#### 4.5.2.3 교훈

정부 주도 표준화는 강력한 추진력을 갖지만, 표준 적용에 따른 기존 시스템 개편 비용이 각 기관의 부담으로 남아 표준 준수율이 기대보다 낮게 나타나는 경우가 많다. 표준 수립과 함께 적용 인센티브(예: 정보화 평가 가점, 우수 사례 시상, 전환 비용 지원)를 병행하는 것이 표준의 실효성을 높인다.

### 4.5.3 사례 3: 글로벌 제조사 T사의 IoT 데이터 모델

#### 4.5.3.1 배경

글로벌 제조사 T사는 전 세계 50개 공장에서 수십만 개의 IoT 센서로부터 실시간 데이터를 수집한다. 기존의 관계형 데이터 모델은 이러한 시계열·이벤트 중심 데이터를 효율적으로 처리하지 못해 새로운 데이터 모델 표준이 필요했다 (International Society of Automation 2018).

#### 4.5.3.2 접근 방법

1. 계층적 시계열 모델 도입 (공장 〉 라인 〉 설비 〉 센서)
2. ISA-95(제조운영관리 국제표준) 메타모델 채택
3. 관계형 DB(마스터·구성정보) + 시계열 DB(센서 측정값)의 하이브리드 모델
4. 센서 메타데이터 표준화 (단위, 정밀도, 측정 주기)

#### 4.5.3.3 교훈

데이터 모델 표준은 환경에 따라 진화해야 한다. 전통적 OLTP 모델만으로는 현대의 다양한 데이터 워크로드(스트리밍, 그래프, 시계열, 문서)를 모두 다룰 수 없다. 표준은 폐쇄적인 규칙이 아니라 다양한 모델링 접근을 통합하는 메타 프레임워크여야 한다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. 데이터 모델링의 본질
   - 데이터 모델링은 현실 세계의 비즈니스를 데이터 구조로 추상화하는 활동이다. 표준화는 이 추상화에 일관성을 부여하여 조직 전반에서 통합 가능한 데이터 자산을 구축하는 기반이 된다.
2. 3대 추상화 레벨
   - ANSI/SPARC의 3-스키마 아키텍처에 따라 개념·논리·물리 모델로 구분된다. 각 레벨은 서로 다른 이해관계자와 목적을 가지며, 표준화 요소도 다르다.
3. ER 모델링의 사실상 표준
   - Peter Chen의 ER 모델은 50년간 데이터 모델링의 표준 표기법이 되어왔다. 엔티티·속성·관계의 명확한 구분과 일관된 표기가 핵심이다.
4. 정규화는 도구, 목적이 아니다

- 정규화는 데이터 무결성을 위한 수단이며, 시스템 목적(OLTP/OLAP)에 맞춰 적용 수준을 결정해야 한다. 의도적 비정규화도 충분히 정당한 선택이다.

5. 명명 규칙의 중요성
   - ISO/IEC 11179 Part 5의 명명 원칙은 데이터 모델의 가독성과 일관성을 결정하는 핵심 요소다. 표준 표현어와 명명 패턴을 일관되게 적용해야 한다.
6. 표준은 도구로 강제하라
   - 사람의 의지에만 의존하는 표준은 시간이 지나면 무너진다. 모델링 도구의 검증 기능, CI/CD 자동 검사, 모델 리뷰 프로세스로 표준을 자동으로 강제해야 한다.

## 다음 학습 가이드

### 4.5.4 국제 표준 및 산업 표준

- ISO/IEC 11179 — Metadata Registries (명명 규칙 포함)
- ISO 20022 — 금융 메시징 데이터 모델
- ISA-95 — 제조운영관리 데이터 모델
- HL7 FHIR — 의료 정보 데이터 모델
- UN/CEFACT Core Component Library — 무역 데이터 표준

### 4.5.5 핵심 학술 문헌

- Chen, P. (1976). The Entity-Relationship Model — 모든 ER 모델링의 시작
- Codd, E. F. (1970). A Relational Model of Data — 정규화 이론의 기원
- Kimball, R. (1996). The Data Warehouse Toolkit — 차원 모델링의 정전(canon)

### 4.5.6 관련 본서 참조 장

- 제 3장 주요 표준화 프레임워크 — 표준화 기구 및 표준
- 제 5장 메타데이터 관리 — 모델 메타데이터의 관리
- 제 6장 데이터 품질 관리 — 모델 기반 품질 규칙
- 제 8장 표준화 구현 전략 — 모델 표준의 조직 내 정착

## 연습 문제

1. 여러분 조직의 업무 영역 하나를 선택하여, 핵심 엔티티 5개와 그 관계를 ER 다이어그램으로 표현하시오. 각 엔티티의 주요 속성과 식별자(기본 키)를 함께 정의하시오.
2. 다음 상황에서 정규화 수준을 어떻게 결정할지 논하시오: (a) 초당 수천 건의 주문을 처리하는 전자상거래 트랜잭션 시스템, (b) 월 단위로 매출을 집계·분석하는 데이터 웨어하우스. 각 경우에 적합한 정규화 수준과 그 이유를 설명하시오.
3. ISO/IEC 11179 Part 5의 명명 원칙(개체 클래스 용어 + 속성 용어 + 표현 용어)에 따라, 다음 데이터 요소의 표준 이름을 작성하시오: (a) 고객의 생년월일, (b) 주문의 총 금액, (c) 상품의 재고 수량.
4. 여러분 조직에서 데이터 모델링 표준이 지켜지지 않는다면 그 원인은 무엇인지 진단하고, 표준을 '도구로 강제'하기 위한 구체적 방안 3가지를 제시하시오.

CHAPTER 5

# 메타데이터 관리

## 5.1 개요

메타데이터(Metadata)는 흔히 '데이터에 관한 데이터'로 정의되지만, 이 짧은 표현은 본질을 충분히 담아내지 못한다. DAMA International (2017) 은 메타데이터를 '데이터의 맥락을 제공하고 데이터 자산의 효과적인 사용을 가능하게 하는 정보'로 정의한다. 즉 메타데이터는 데이터를 단순히 '설명'하는 것을 넘어, 데이터가 비즈니스 가치를 창출할 수 있도록 **맥락(context)을 부여하는 정보 체계**이다.

조직이 보유한 데이터가 폭증하는 빅데이터 시대에 메타데이터의 중요성은 더욱 부각된다. Gartner는 '데이터 카탈로그와 메타데이터 관리가 향후 5년간 데이터·분석 분야에서 가장 영향력 있는 기술 트렌드 중 하나'로 평가한다 (Gartner 2023b). 메타데이터가 부재한 데이터는 '다크 데이터(Dark Data)'로 전락하며, IBM의 추정에 따르면 전체 기업 데이터의 약 80%가 검색·이해되지 못한 채 방치되고 있다 (IBM 2022).

### 5.1.1 메타데이터의 비즈니스 가치

메타데이터는 다음 다섯 가지 측면에서 조직의 비즈니스 가치를 직접적으로 창출한다.

1. 데이터 검색성(Findability) 향상 — 데이터 카탈로그를 통해 분석가가 필요한 데이터를 빠르게 발견
2. 데이터 신뢰성(Trust) 강화 — 출처, 품질, 변환 이력의 투명한 공개
3. 규제 준수(Compliance) — GDPR, 개인정보보호법 등이 요구하는 데이터 처리 기록의 자동화
4. 의사결정 속도 향상 — 데이터 의미의 모호성 제거로 분석 시간 단축
5. 리스크 관리 — 잘못된 데이터 사용으로 인한 의사결정 오류 방지

Forrester Research (2023c) 의 2023년 조사에 따르면, 체계적인 메타데이터 관리를 도입한 기업은 그렇지 않은 기업 대비 데이터 분석가의 생산성이 평균 **37% 향상**되었으며, 데이터 관련 의사결정 시간이 45% 단축되었다[4].

### 5.1.2 본 장의 학습 목표

- 메타데이터의 3대 유형(기술적·비즈니스·운영)을 구분하고 각각의 관리 방법을 설명할 수 있다.
- ISO/IEC 11179 기반의 메타데이터 레지스트리 구조를 이해한다.
- 데이터 계보, 데이터 카탈로그, 비즈니스 용어집을 설계할 수 있다.
- 메타데이터 관리 도구를 평가·선정할 수 있다.
- 조직 내 메타데이터 관리 체계의 단계적 구축 로드맵을 수립할 수 있다.

## 5.2 핵심 개념

### 5.2.1 메타데이터의 3대 유형

DAMA International (2017) 은 메타데이터를 **기술적(Technical)**, **비즈니스(Business)**, **운영(Operational)** 세 가지 유형으로 분류한다. 이는 메타데이터 관리 체계를 설계할 때 가장 기본이 되는 분류 체계이다.

#### 5.2.1.1 기술적 메타데이터(Technical Metadata)

데이터의 물리적·구조적 특성을 기술하는 정보다. 데이터베이스 관리자(DBA)와 데이터 엔지니어가 주요 사용자이며, 대부분 시스템에서 자동으로 추출 가능하다.

- 테이블·컬럼명, 데이터 타입, 길이, 제약 조건
- 기본키·외래키 관계, 인덱스 구조
- 저장 위치(스키마, 데이터베이스, 파일 경로)
- API 엔드포인트, 메시지 큐 토픽

#### 5.2.1.2 비즈니스 메타데이터(Business Metadata)

데이터의 비즈니스적 의미와 맥락을 설명하는 정보다. 데이터 분석가, 현업 사용자, 데이터 스튜어드가 주요 사용자이며, 사람이 직접 작성하고 관리해야 하는 가장 도전적인 영역이다.

- 비즈니스 용어 정의 및 동의어
- 데이터 소유자(Data Owner)와 스튜어드
- 비즈니스 규칙과 계산 로직
- 데이터 분류(공개/내부/기밀/극비)와 보존 정책
- 규제 준수 분류(개인정보, 금융정보 등)

#### 5.2.1.3 운영 메타데이터(Operational Metadata)

데이터의 처리 이력과 운영 상태를 기록하는 정보다. 시스템 운영자와 데이터 엔지니어가 활용하며, ETL/ELT 도구와 데이터 파이프라인에서 자동 생성된다.

- ETL 작업 실행 시간, 처리 레코드 수, 성공/실패 상태
- 데이터 갱신 빈도와 마지막 갱신 시점
- 데이터 품질 측정 결과와 추세
- 접근 로그와 사용 통계

#### 5.2.1.4 메타데이터의 3대 유형 요약

| 유형 | 설명 | 예시 | 주요 사용자 |
|---|---|---|---|
| 기술적 (Technical) | 물리적·구조적 특성 | 테이블·컬럼명, 데이터 타입, 인덱스 | DBA, 데이터 엔지니어 |
| 비즈니스 (Business) | 비즈니스 의미와 맥락 | 용어 정의, 소유자, 규제 분류 | 데이터 분석가, 스튜어드 |

| 유형 | 설명 | 예시 | 주요 사용자 |
|---|---|---|---|
| 운영 (Operational) | 처리 이력과 운영 상태 | ETL 실행 이력, 갱신 시점, 사용 통계 | 시스템 운영자 |

### 5.2.2 ISO/IEC 11179: 메타데이터 레지스트리 국제 표준

ISO/IEC (2023b) 는 **메타데이터 레지스트리(Metadata Registry, MDR)**의 구조와 운영에 관한 국제 표준으로, 1990년대 후반 첫 발간 이후 여러 차례 개정되었다. 이 표준은 전 세계 통계청, 의료기관, 정부기관이 데이터 요소를 표준화하는 기반이 되어왔다. 현재 6개의 파트로 구성되어 있으며, 각각의 역할은 다음과 같다.

| 파트 | 제목 | 주요 내용 |
|---|---|---|
| Part 1 | Framework | MDR의 개념적 프레임워크와 용어 |
| Part 3 | Registry metamodel | 메타모델 구조와 기본 속성 |
| Part 4 | Formulation of data definitions | 데이터 정의 작성 규칙 |
| Part 5 | Naming principles | 명명규칙과 식별자 부여 규칙 |
| Part 6 | Registration | 메타데이터 등록 절차 |
| Part 7 | Metadata for data sets | 데이터셋 단위 메타데이터 |

#### 5.2.2.1 ISO/IEC 11179의 핵심 개념: 데이터 요소(Data Element)

ISO/IEC 11179는 메타데이터 관리의 최소 단위로 **데이터 요소(Data Element)**라는 개념을 제시한다. 데이터 요소는 다음 세 가지 구성 요소의 결합으로 정의된다(ISO/IEC 2023b).

- 개념(Concept): 표현하고자 하는 추상적 개념 (예: '사람의 연령')
- 객체 클래스(Object Class): 개념이 속한 실세계 객체 (예: '사람')
- 표현(Representation): 값의 도메인과 표현 방식 (예: 0~150 범위의 정수, 단위는 년)

> **i** 데이터 요소 구성 예시
>
> '고객의 만 나이'라는 데이터 요소는 다음과 같이 구조화된다:
>
> - 개념: 사람의 연령 (Age)
> - 객체 클래스: 고객 (Customer)
> - 특성(Property): 만 나이 (Full Age)
> - 표현 클래스: 양의 정수 (Positive Integer)
> - 값 도메인: 0 ~ 150

### 5.2.3 데이터 카탈로그(Data Catalog)

데이터 카탈로그는 **조직 내 모든 데이터 자산의 메타데이터를 중앙에서 통합 관리하는 시스템**이다. Gartner는 데이터 카탈로그를 '데이터·분석 책임자의 필수 도구'로 분류하며, '데이터 자산의 위치, 의

미, 신뢰도, 사용법을 한 곳에서 제공하여 데이터 민주화(Data Democratization)를 실현하는 핵심 인프라'라고 정의한다(Gartner 2023b).

#### 5.2.3.1 데이터 카탈로그의 핵심 기능

| 기능 영역 | 주요 기능 |
|---|---|
| 데이터 자산 등록 | 테이블, 컬럼, 파일, API, 보고서, 대시보드 등 모든 데이터 자산 자동 수집 |
| 검색 및 탐색 | 키워드 검색, 패싯 필터, AI 추천 |
| 비즈니스 컨텍스트 | 용어집 연결, 비즈니스 정의, 태그·카테고리 |
| 데이터 계보 | 원천부터 소비까지의 흐름 시각화 |
| 품질 정보 | 품질 점수, 신선도, 사용 빈도 |
| 거버넌스 | 데이터 분류, 접근 권한, 정책 적용 |
| 협업 | 코멘트, 평가, 인증 마크, 질문/답변 |

### 5.2.4 데이터 계보 (Data Lineage)

데이터 계보는 데이터가 원천에서 최종 소비 지점까지 이동하는 경로와 변환 과정을 추적하는 메타데이터이다. Singh and Sharma (2022) 는 이를 '데이터의 족보'에 비유한다. 계보 추적은 다음 두 방향으로 수행된다.

- Upstream Lineage(상류 추적): 특정 데이터가 어디서 왔는지 거꾸로 추적
- Downstream Lineage(하류 추적): 특정 데이터가 어디로 흘러가는지 영향 분석

> 데이터 계보의 실무 가치
>
> - **영향도 분석**: 원천 시스템 변경 시 영향받는 다운스트림 자산 파악
> - **근본 원인 분석**: 보고서 오류 발견 시 어느 단계에서 문제가 발생했는지 추적
> - **규제 대응**: GDPR의 '잊혀질 권리' 요청 시 해당 개인정보의 모든 사본 식별
> - **신뢰성 평가**: 분석 결과의 신뢰도를 데이터 출처로부터 평가

### 5.2.5 비즈니스 용어집

비즈니스 용어집은 **조직 내에서 사용되는 비즈니스 용어의 공식적 정의 모음**이다. 같은 단어가 부서마다 다른 의미로 사용되는 용어의 바벨탑 문제를 해결하기 위한 핵심 도구다. Brackett (2012) 는 비즈니스 용어집을 '데이터 거버넌스의 출발점이자 모든 표준화 활동의 기반'이라고 강조한다.

## 5.3 방법론

### 5.3.1 메타데이터 레지스트리 구축 5단계

메타데이터 레지스트리(MDR) 구축은 ISO/IEC (2023b) 와 DAMA International (2017) 의 권고사항을 결합한 체계적 방법론을 따른다. 본 절에서는 실무에서 검증된 5단계 구축 방법론을 제시한다.

#### 5.3.1.1 Phase 1: 범위 정의(Scope Definition)

- 대상 데이터 영역 선정(도메인, 시스템, 부서)
- 이해관계자 식별 및 R&R 정의
- 기대 효과 및 성공 지표 설정

- 프로젝트 조직 구성(스폰서, PM, 데이터 스튜어드, 기술 팀)

#### 5.3.1.2 Phase 2: 메타모델 설계(Metamodel Design)

- ISO/IEC 11179 메타모델 기반의 조직 메타모델 정의
- 핵심 엔티티 식별: 데이터 요소, 값 도메인, 개념, 객체 클래스
- 관계 정의: 동의어, 관련어, 상위/하위 개념
- 확장 속성 정의: 비즈니스 규칙, 품질 규칙, 사용 제약

#### 5.3.1.3 Phase 3: 메타데이터 수집(Metadata Acquisition)

- 기술적 메타데이터: 데이터베이스 스키마 자동 스캐닝
- 운영 메타데이터: ETL 도구·로그 분석을 통한 자동 추출
- 비즈니스 메타데이터: 데이터 스튜어드와 SME 인터뷰
- 기존 문서·시스템(DDL, 데이터 사전, ERD)으로부터 마이그레이션

#### 5.3.1.4 Phase 4: 검증 및 등록(Validation & Registration)

- ISO/IEC 11179 명명 규칙 준수 검증
- 중복·충돌 데이터 요소 식별 및 통합
- 승인 워크플로우를 통한 공식 등록
- 버전 관리 및 변경 이력 기록

#### 5.3.1.5 Phase 5: 운영 및 진화(Operation & Evolution)

- 사용자 교육 및 변경 관리
- 정기적 메타데이터 품질 측정
- 이해관계자 피드백 기반 지속 개선
- 신규 시스템·도메인으로의 확장

### 5.3.2 ISO/IEC 11179 명명 규칙

ISO/IEC (2015a) 는 데이터 요소의 명명에 관한 상세한 원칙을 제공한다. 명명 규칙의 일관성은 메타데이터 레지스트리의 사용성을 결정하는 핵심 요소다.

#### 5.3.2.1 표준 명명 구조

ISO/IEC 11179가 권장하는 데이터 요소 명칭 구조는 다음과 같다:

> **i 표준 명명 패턴**
>
> `[Object Class] + [Property] + [Representation Term]`
>
> 예시:
>
> - Customer_Name_Text (고객명_텍스트)
> - Product_Price_Amount (상품가격_금액)
> - Order_Date_Date (주문일자_일자)
> - Employee_HireDate_Date (직원입사일_일자)

#### 5.3.2.2 표현어(Representation Term) 표준

ISO/IEC 11179가 정의하는 주요 표현어와 그 의미는 다음과 같다.

| 표현어 | 의미 | 사용 예 |
|---|---|---|
| Identifier (ID) | 유일 식별자 | Customer_ID, Order_ID |
| Name | 고유 명칭 | Customer_Name, Product_Name |
| Code | 분류 코드 | Country_Code, Status_Code |
| Amount | 금액 | Total_Amount, Tax_Amount |
| Quantity | 수량 | Order_Quantity, Stock_Quantity |
| Date | 날짜 | Birth_Date, Hire_Date |
| DateTime | 일시 | Created_DateTime, Updated_DateTime |
| Text | 자유 텍스트 | Description_Text, Note_Text |
| Flag | 참/거짓 표시 | Active_Flag, Deleted_Flag |
| Rate | 비율 | Interest_Rate, Discount_Rate |

### 5.3.3 데이터 계보 구축 기법

데이터 계보 추적 방법은 정밀도(granularity)와 자동화 수준에 따라 여러 단계로 구분된다(Singh and Sharma 2022). 조직의 성숙도와 필요에 따라 적절한 수준을 선택해야 한다.

| 수준 | 정밀도 | 구현 방법 | 자동화 |
|---|---|---|---|
| 테이블 수준 | 낮음 | ETL 도구 메타데이터 추출 | 높음 |
| 컬럼 수준 | 중간 | SQL 파싱, ETL 변환 매핑 | 중간 |
| 변환 로직 수준 | 높음 | 코드 정적 분석, 데이터 흐름 추적 | 낮음 |
| 값 수준(Field-level) | 매우 높음 | 런타임 추적, 데이터 태깅 | 매우 낮음 |

### 5.3.4 비즈니스 용어집 작성 방법론

비즈니스 용어집은 단순히 단어의 사전이 아니라 조직의 비즈니스 개념 체계를 정의하는 작업이다. Inmon (2005) 은 '용어집 없는 데이터 모델링은 모래 위의 성과 같다'고 표현했다. 효과적인 용어집 작성을 위한 표준 절차는 다음과 같다.

#### 5.3.4.1 용어집 표준 작성 절차

1. 후보 용어 수집: 현행 문서, 보고서, 시스템 화면, 이해관계자 인터뷰
2. 우선순위 정의: 비즈니스 임팩트와 사용 빈도 기반
3. 정의 초안 작성: 단순·명확·완전한 정의 원칙
4. 이해관계자 검토: 도메인 전문가 검토 및 합의
5. 공식 등록: 데이터 거버넌스 위원회 승인
6. 운영 및 갱신: 변경 요청 처리, 정기 검토

#### 5.3.4.2 좋은 정의의 5가지 조건

Brackett(2012)은 좋은 비즈니스 용어 정의의 조건으로 다음 다섯 가지를 제시한다[9].

- 명확성(Clarity): 모호한 표현 없이 명확하게 진술
- 정확성(Accuracy): 비즈니스 개념을 정확히 반영
- 완전성(Completeness): 개념의 핵심 속성을 모두 포함

- 간결성(Conciseness): 불필요한 수식어 없이 간결
- 일관성(Consistency): 다른 용어 정의와 모순 없음

## 5.4 실무 적용

### 5.4.1 메타데이터 관리 도구 선정

메타데이터 관리 도구 시장은 빠르게 발전하고 있으며, 2024년 기준 주요 도구들은 다음과 같이 분류된다(Gartner 2023b).

| 구분 | 도구 | 특징 |
|---|---|---|
| 엔터프라이즈 상용 | Collibra, Informatica EDC, IBM IGC, Alation | 포괄적 기능, 고비용 |
| 클라우드 네이티브 | AWS Glue Catalog, Azure Purview, GCP Data Catalog | 클라우드 통합 최적화 |
| 오픈소스 | Apache Atlas, DataHub, OpenMetadata, Amundsen | 낮은 비용, 커스터마이징 |
| AI 기반 신규 | Atlan, Castor, Select Star | AI 자동화, 모던 UX |

#### 5.4.1.1 도구 선정 체크리스트

- 기존 데이터 인프라와의 통합성(데이터베이스, ETL, BI 도구)
- 메타데이터 자동 수집 범위와 정확도
- 데이터 계보 추적 정밀도
- 비즈니스 사용자 인터페이스의 직관성
- API 및 확장성
- 데이터 거버넌스 워크플로우 지원
- 총 소유 비용(TCO) — 라이선스, 구축, 운영

### 5.4.2 단계적 도입 로드맵

#### 5.4.2.1 Stage 1: 파일럿 (0~3개월)

- 1개 핵심 도메인 선정(예: 고객, 상품)
- 기본 데이터 사전 구축
- 주요 데이터 자산 50~100개 등록
- 데이터 스튜어드 2~3명 지정

#### 5.4.2.2 Stage 2: 확장 (3~9개월)

- 3~5개 도메인으로 확장
- 비즈니스 용어집 100~300개 정의
- 자동 메타데이터 수집 파이프라인 구축
- 데이터 계보 추적 도입(테이블 수준)

#### 5.4.2.3 Stage 3: 통합 (9~18개월)

- 전사 모든 핵심 데이터 자산 등록
- 컬럼 수준 데이터 계보 추적
- 데이터 품질 메타데이터 통합
- 데이터 거버넌스 위원회 정착

#### 5.4.2.4 Stage 4: 최적화 (18개월~)

- AI 기반 메타데이터 자동 분류·태깅
- 실시간 메타데이터 동기화
- 자동화된 데이터 영향도 분석
- 외부 시스템과의 메타데이터 연동(파트너, 규제 기관)

### 5.4.3 실무 팁

> 팁 1: 완벽함보다 일관성을 추구하라
>
> 메타데이터 관리는 처음부터 완벽할 수 없다. 모든 데이터 자산에 100% 완전한 메타데이터를 부여하려는 시도는 거의 실패한다. 대신 '**핵심 자산에 일관된 표준을 적용**'하는 데 집중하라. Pareto 법칙처럼 상위 20%의 데이터 자산이 가치의 80%를 차지한다 (Loshin 2010b).

> 팁 2: 자동화를 최우선으로 설계하라
>
> 사람이 수동으로 입력하는 메타데이터는 결국 시스템 변경을 따라가지 못해 신뢰성을 잃는다. 기술적·운영 메타데이터는 가능한 모두 자동 수집되도록 파이프라인을 설계하고, 비즈니스 메타데이터만 사람의 손길이 필요하도록 분리하라.

> 팁 3: 데이터 스튜어드십을 제도화하라
>
> 메타데이터 관리는 기술이 아닌 사람의 문제다. 각 데이터 도메인에 명확한 책임자를 지정하고, 이들의 역할을 공식적인 직무로 인정해야 한다. Sambasivan et al. (2021) 등의 연구에 따르면, 메타데이터 관리 실패의 60% 이상은 기술 문제가 아닌 거버넌스와 책임 부재 때문이다.

> ⚠ 주의: 메타데이터 자체의 품질을 관리하라
>
> 메타데이터도 데이터다. 즉, 품질 문제가 발생할 수 있다. 오래된 정보, 부정확한 정의, 누락된 관계, 일관성 없는 명명은 모두 메타데이터 품질 저하의 사례다. 메타데이터에 대해서도 정기적인 품질 점검(완전성, 정확성, 적시성)을 수행해야 한다.

## 5.5 사례 연구

### 5.5.1 사례 1: 미국 인구조사국의 메타데이터 표준화

#### 5.5.1.1 배경

미국 인구조사국(U.S. Census Bureau)은 1990년대부터 ISO/IEC 11179 표준을 기반으로 메타데이터 레지스트리를 구축한 대표적 사례다(@ U.S. Census Bureau 2020). 통계 데이터의 국제적 비교 가능성과 정부 기관 간 데이터 공유의 필요성이 동기였다.

#### 5.5.1.2 도전 과제

- 수십 년에 걸친 다양한 조사 데이터의 통합
- 각 부처가 사용하는 상이한 데이터 정의의 표준화
- 국제 통계 기관(UN, OECD)과의 비교 가능성 확보
- 일반 시민의 데이터 접근성 향상

#### 5.5.1.3 접근 방법

1. ISO/IEC 11179 메타모델 채택 및 미국 정부 맞춤 확장
2. USGS(US Geological Survey) 데이터 표준과의 통합
3. 개방형 메타데이터 레지스트리(MDR) 구축
4. Data.gov 플랫폼과의 연계

#### 5.5.1.4 성과

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 표준화 데이터 요소 | 약 500개 | 약 12,000개 |
| 기관 간 데이터 공유 시간 | 수개월 | 수일 |
| 국제 통계 비교 가능 데이터 | 30% | 85% |
| 일반 시민 데이터 접근 건수(연간) | 수천 건 | 수백만 건 |

### 5.5.2 사례 2: 한국 행정안전부 공공데이터 표준화

#### 5.5.2.1 배경

한국 행정안전부는 2013년 공공데이터법 시행 이후 공공데이터 표준화를 본격 추진해왔다(@ 행정안전부 2023a). 중앙부처와 지방자치단체가 보유한 데이터를 표준화하여 공공데이터포털(data.go.kr)을 통해 민간에 개방하는 것이 목표였다.

#### 5.5.2.2 주요 추진 내용

- 공공데이터 공통표준용어 정립(2014년 시작, 현재 6,000여 개)
- 공공데이터 표준 데이터셋(약 1,500종) 제공
- 자치단체 행정정보 표준 데이터(64종) 구축
- 공공데이터 품질관리 매뉴얼 보급

#### 5.5.2.3 교훈

정부 주도의 표준화는 강력한 추진력을 갖지만, 현장의 실무자가 표준의 가치를 체감하지 못하면 형식적 준수에 그치기 쉽다. 한국 사례는 표준 수립과 더불어 표준 적용 시 받는 인센티브 (예: 평가 가점, 우수 사례 시상)를 함께 제공할 때 효과가 극대화됨을 보여준다.

### 5.5.3 사례 3: 글로벌 제약사 P사의 임상시험 메타데이터

#### 5.5.3.1 배경

글로벌 제약사 P사는 임상시험 데이터의 메타데이터를 CDISC(Clinical Data Interchange Standards Consortium) 표준에 맞춰 관리하는 체계를 구축했다(CDISC 2023b). FDA, EMA 등 규제 기관 제출 요건을 충족하면서 내부 R&D 효율성을 높이는 것이 목표였다.

#### 5.5.3.2 접근 방법

1. CDISC SDTM(Study Data Tabulation Model) 메타모델 채택
2. ISO/IEC 11179 호환 메타데이터 레지스트리 구축
3. 모든 임상시험 변수의 표준 정의 사전 구축(약 8,000개)
4. 자동화된 메타데이터 추출 및 검증 파이프라인

#### 5.5.3.3 성과

- FDA 제출 자료 준비 기간 50% 단축

- 규제 기관 요청 응답 시간 70% 단축
- 기존 임상시험 데이터의 재사용률 3배 증가
- 연구자 간 데이터 이해 불일치 90% 감소

#### 5.5.3.4 교훈

산업별 표준(CDISC, HL7, ISO 20022 등)과 일반 표준(ISO/IEC 11179)을 결합하는 것이 효과적이다. 산업 표준은 도메인 특화 의미를 제공하고, 일반 표준은 메타데이터 관리의 구조적 일관성을 제공한다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **메타데이터의 본질** — 메타데이터는 단순한 '데이터에 관한 데이터'가 아니라 데이터에 비즈니스 맥락을 부여하여 가치를 창출하게 하는 정보 체계다. 기술적·비즈니스·운영 세 유형이 균형있게 관리되어야 한다.
2. **국제 표준의 활용** — ISO/IEC 11179는 메타데이터 레지스트리의 사실상 국제 표준이며, 데이터 요소의 구조화와 명명 규칙을 위한 체계적 프레임워크를 제공한다.
3. **5대 메타데이터 자산** — 데이터 카탈로그, 비즈니스 용어집, 데이터 계보, 메타데이터 레지스트리, 품질 메타데이터는 현대 메타데이터 관리의 핵심 자산이다.
4. **단계적 구축의 중요성** — 완벽한 메타데이터 체계를 한 번에 구축하려는 시도는 거의 실패한다. 핵심 도메인부터 시작하여 점진적으로 확장하고, 자동화 가능한 부분부터 우선 처리하는 접근이 효과적이다.
5. **도구와 거버넌스의 균형** — 메타데이터 관리는 기술적 도구만으로 성공할 수 없다. 데이터 스튜어드십, 거버넌스 위원회, 변경 관리 프로세스 같은 인적·조직적 요소가 함께 작동해야 한다.
6. **지속적 진화** — 메타데이터는 한 번 만들고 끝나는 것이 아니라 비즈니스 변화에 따라 지속적으로 진화해야 하는 살아있는 자산이다. 메타데이터의 품질 자체에 대한 관리도 필요하다.

## 다음 학습 가이드

### 5.5.4 국제 표준 및 산업 표준

- ISO/IEC 11179 — Metadata Registries (Parts 1-7)
- ISO/IEC 19763 — Metamodel Framework for Interoperability
- Dublin Core Metadata Element Set — 문헌·디지털 자원 메타데이터
- CDISC — 임상시험 데이터 표준
- DCAT (W3C) — 데이터 카탈로그 어휘

### 5.5.5 관련 본서 참조 장

- 제 3장 주요 표준화 프레임워크 — ISO/IEC 11179의 위치
- 제 6장 데이터 품질 관리 — 메타데이터를 통한 품질 관리
- 제 9장 데이터 거버넌스 체계 — 메타데이터 스튜어드십
- 제 10장 표준화 도구와 기술 — 메타데이터 관리 도구
- 제 12장 AI·ML과 데이터 표준화 — AI 시대의 메타데이터 확장

## 연습 문제

1. 여러분 조직의 핵심 데이터 요소 하나를 선택하여, 기술적·비즈니스·운영 메타데이터를 각각 3가지 이상 나열하시오. 세 유형이 어떻게 상호 보완적으로 데이터의 가치를 높이는지 설명하시오.

2. ISO/IEC 11179 메타데이터 레지스트리를 조직에 도입한다고 가정하고, 5단계 구축 로드맵을 여러분 조직 상황에 맞게 구체화하시오. 각 단계의 주요 활동과 산출물을 제시하시오.
3. 데이터 계보(Data Lineage)가 다음 상황에서 어떤 실무 가치를 제공하는지 설명하시오: (a) 규제 대응을 위한 영향도 분석, (b) 데이터 품질 이슈의 근본 원인 추적, (c) 시스템 변경 시 파급 효과 평가.
4. "완벽함보다 일관성, 사람보다 자동화, 기술보다 거버넌스"라는 원칙 중 여러분 조직에서 가장 취약한 부분은 무엇인지 진단하고, 이를 개선하기 위한 구체적 방안을 제시하시오.

CHAPTER 6

# 데이터 품질 관리

## 6.1 개요

데이터 품질(Data Quality)은 데이터가 **의도된 사용 목적에 적합한 정도(fitness for use)**를 의미한다. Juran이 품질 일반에 대해 제시한 이 정의를 데이터에 적용한 것으로, Redman (1996) 은 이를 데이터 품질 관리의 출발점으로 삼는다. 동일한 데이터라도 사용 맥락에 따라 품질의 의미가 달라진다는 점이 데이터 품질을 다른 IT 영역과 구별짓는 본질적 특성이다. 예컨대 마케팅 분석에는 충분한 고객 데이터가 청구서 발송에는 부족할 수 있다.

데이터 품질 문제의 비용은 실로 막대하다. Gartner의 추정에 따르면 **조직은 데이터 품질 문제로 인해 연평균 1,290만 달러의 손실**을 입고 있으며 (Gartner 2021), IBM은 미국 경제 전체가 데이터 품질 문제로 연간 약 **3.1조 달러**의 손실을 본다고 추정한다 (IBM 2016). 이러한 손실의 80%는 잘못된 의사결정, 운영 비효율, 규제 벌금에서 발생하며, 나머지 20%만이 데이터 정제 비용으로 가시화된다.

### 6.1.1 데이터 품질의 비즈니스 가치

체계적인 데이터 품질 관리는 다음과 같은 직접적인 비즈니스 가치를 창출한다.

1. **의사결정의 신뢰성 향상** — 분석 결과의 정확도와 일관성 확보
2. **운영 비용 절감** — 데이터 오류로 인한 재작업과 고객 불만 감소
3. **규제 준수 강화** — GDPR, BCBS 239 (Basel Committee on Banking Supervision 2013) 등 데이터 정확성 요구 충족
4. **고객 경험 개선** — 정확한 고객 정보 기반 개인화 서비스
5. **디지털 전환 가속** — AI/ML 모델의 학습 데이터 품질 확보 (Sambasivan et al. 2021)

### 6.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 데이터 품질의 6대 차원과 그 측정 방법을 이해한다
- ISO (2022b) 과 ISO/IEC (2008) 국제 표준의 핵심 내용을 적용할 수 있다

- DQM 사이클(정의·측정·분석·개선·통제)을 실무에 적용할 수 있다
- 데이터 품질 KPI를 설계하고 측정 도구를 선정할 수 있다
- 데이터 품질 거버넌스 체계를 구축할 수 있다

## 6.2 핵심 개념

### 6.2.1 데이터 품질의 6대 차원

데이터 품질을 객관적으로 측정하려면 품질을 구성하는 **측정 가능한 차원(dimension)**들로 분해해야 한다. DAMA International (2017) 이 제시한 6대 품질 차원은 업계에서 가장 널리 채택되는 분류 체계다. ISO/IEC (2008) 는 이를 15개로 확장하지만, 실무에서는 6대 차원으로 시작하는 것이 효과적이다.

| 차원 | 정의 | 측정 예시 |
|---|---|---|
| 완전성(Completeness) | 필요한 데이터가 모두 존재하는 정도 | 필수 필드 충족률 = 채워진 레코드 ÷ 전체 |
| 정확성(Accuracy) | 데이터가 실제 사실과 일치하는 정도 | 외부 신뢰 소스와 비교, 형식 검증 |
| 일관성(Consistency) | 동일 데이터가 시스템 간에 일치하는 정도 | 시스템 간 동일 키 데이터 일치율 |
| 유효성(Validity) | 정의된 규칙·형식·범위에 부합하는 정도 | 도메인 위반 건수, 형식 위반율 |
| 유일성(Uniqueness) | 동일 개체가 중복 없이 표현되는 정도 | 1 - (중복 레코드 ÷ 전체) |
| 적시성(Timeliness) | 데이터가 필요한 시점에 사용 가능한 정도 | 데이터 신선도, 갱신 지연 시간 |

#### 6.2.1.1 완전성(Completeness)

완전성은 세 가지 하위 수준으로 분해된다 (DAMA International 2017). 컬럼 수준(필수 필드가 채워졌는가), 레코드 수준(개체에 필요한 모든 정보가 있는가), 데이터셋 수준(필요한 모든 개체가 데이터셋에 존재하는가)이다.

- **컬럼 완전성**: 단일 컬럼의 NULL 비율
- **레코드 완전성**: 한 레코드의 필수 필드 충족도
- **데이터셋 완전성**: 모집단 대비 데이터 커버리지

#### 6.2.1.2 정확성(Accuracy)

정확성은 가장 측정하기 어려운 차원이다. 데이터가 '실제 사실'과 일치하는지 알려면 비교 기준이 되는 '진실(ground truth)'이 있어야 하는데, 많은 경우 이 진실 자체를 알기 어렵다. 따라서 정확성은 다음과 같이 측정된다.

- **형식 정확성(Syntactic Accuracy)**: 정의된 형식에 맞는가 (예: 이메일 형식)
- **의미 정확성(Semantic Accuracy)**: 실제 의미와 일치하는가 (예: 실제 거주 주소)
- **참조 무결성(Referential)**: 참조하는 데이터가 실제 존재하는가

#### 6.2.1.3 일관성(Consistency)

일관성은 한 시스템 내부의 일관성과 시스템 간 일관성으로 구분된다. 예를 들어 한 고객의 전화번호가 CRM과 ERP에서 다르면 시스템 간 일관성 위반이고, 한 시스템 내에서 동일 고객의 주소가 두 레코드에 다르게 기록되면 시스템 내 일관성 위반이다.

#### 6.2.1.4 유효성(Validity)

유효성은 데이터가 정의된 비즈니스 규칙, 데이터 타입, 값 도메인을 따르는지를 의미한다. 예: 나이는 0~150 사이 정수, 우편번호는 5자리 숫자, 주문 상태는 정의된 코드값.

#### 6.2.1.5 유일성(Uniqueness)

유일성은 동일 실세계 개체가 데이터에서 중복으로 표현되지 않는 정도이다. 단순 PK 중복이 아니라 의미적 중복(예: 같은 사람이 다른 형식으로 두 번 등록)을 포함한다. 엔티티 해소(Entity Resolution)와 깊은 관련이 있다.

#### 6.2.1.6 적시성(Timeliness)

적시성은 두 가지 측면을 포함한다. 데이터의 신선도(현재 시점 데이터를 반영하는가)와 가용성(필요한 순간에 데이터를 얻을 수 있는가)이다. 실시간 거래 데이터와 월간 보고용 데이터의 적시성 요구는 매우 다르다.

### 6.2.2 ISO 8000과 ISO/IEC 25012

데이터 품질에 관한 두 개의 핵심 국제 표준이 있다. ISO (2022b) 은 데이터 품질의 **관리 프로세스**에 초점을 두고, ISO/IEC (2008) 는 데이터 품질의 **측정 모델**에 초점을 둔다.

#### 6.2.2.1 ISO 8000: 데이터 품질 관리 표준

ISO (2022b) 은 데이터 품질을 체계적으로 관리하기 위한 일련의 표준으로, 특히 마스터 데이터와 트랜잭션 데이터의 품질에 대한 명시적 요구사항을 정의한다. 주요 파트는 다음과 같다.

| 파트 | 주제 | 내용 |
|---|---|---|
| Part 1 | Overview | 용어 및 개념 정의 |
| Part 2 | Vocabulary | 표준 용어 사전 |
| Part 8 | Information & Data Quality | 정보·데이터 품질의 일반 개념 |
| Part 61 | Data Quality Management | 조직 차원의 DQM 프로세스 |
| Part 100~150 | Master Data Quality | 마스터 데이터 품질 측정 |
| Part 110 | Syntax Quality | 구문 품질 요구사항 |
| Part 120~140 | Provenance / Accuracy | 출처 및 정확성 |

#### 6.2.2.2 ISO/IEC 25012: 데이터 품질 모델

ISO/IEC (2008) 는 ISO/IEC 25000 시리즈(SQuaRE - Software Quality Requirements and Evaluation)의 일부로, 데이터 품질을 **15개 특성**으로 정의한다. 이 15개 특성은 **고유 특성(Inherent)**과 **시스템 종속 특성(System-dependent)**으로 분류된다.

| 분류 | 특성 |
|---|---|
| 고유 특성 | 정확성, 완전성, 일관성, 신뢰성, 시기 적절성 |
| 고유 + 시스템 종속 | 접근성, 준수성, 기밀성, 효율성, 정밀성, 추적성, 이해성 |
| 시스템 종속 특성 | 가용성, 이식성, 복구성 |

### 6.2.3 데이터 품질 관리 사이클 (DQM Cycle)

데이터 품질 관리는 일회성 정제 작업이 아니라 지속적인 사이클로 운영되어야 한다. Redman (1996) 이 제시하고 DAMA International (2017) 이 발전시킨 **5단계 DQM 사이클**이 업계 표준 프레임워크다. 이 사이클은 Deming의 PDCA 사이클을 데이터 품질에 적용한 것이다.

1. **정의(Define)**: 품질 차원과 측정 규칙 정의
2. **측정(Measure)**: 실제 데이터의 품질 상태 측정
3. **분석(Analyze)**: 품질 문제의 원인 분석
4. **개선(Improve)**: 데이터 정제 및 프로세스 개선
5. **통제(Control)**: 지속적 모니터링 및 예방

> **i** 핵심 원칙: '근본 원인'에 집중하라
>
> 데이터 정제는 증상 치료에 불과하다. 진정한 데이터 품질 향상은 데이터가 잘못 생성되는 프로세스 자체를 개선하는 데서 시작된다. Redman (2016) 은 '데이터 품질 문제의 80%는 데이터 입력·생성 단계에서 발생한다'고 강조한다.

### 6.2.4 데이터 프로파일링(Data Profiling)

데이터 프로파일링은 **데이터의 구조·내용·관계를 체계적으로 조사하여 품질 상태를 파악하는 활동**이다. Olson (2003) 은 프로파일링을 '데이터 품질 관리의 가장 중요한 진단 도구'로 묘사한다. 프로파일링은 세 가지 수준에서 수행된다.

#### 6.2.4.1 컬럼 프로파일링(Column Profiling)

- 기본 통계: 최소·최대·평균·표준편차
- 값 분포: 고유값 수, NULL 비율, 빈도 분포
- 패턴: 데이터 형식 패턴, 길이 분포
- 이상치: 비정상 값 탐지

#### 6.2.4.2 구조 프로파일링(Structural Profiling)

- 기본키 후보 식별: 유일한 컬럼 조합 발견
- 종속성 분석: 컬럼 간 함수 종속 탐지
- 정규화 위반: 부분 종속, 이행 종속 식별

#### 6.2.4.3 관계 프로파일링(Relationship Profiling)

- 외래키 후보 식별: 테이블 간 참조 관계 탐지
- 참조 무결성 검증: 고아 레코드 식별
- 중복 탐지: 동일 개체의 중복 표현 발견

## 6.3 방법론

### 6.3.1 데이터 품질 KPI 설계

Drucker (1954) 의 격언처럼 '측정할 수 없으면 관리할 수 없다'. 이는 데이터 품질에 그대로 적용된다. 효과적인 KPI 설계는 다음 5단계 절차를 따른다.

1. 핵심 데이터 자산(CDE: Critical Data Element) 식별
2. 각 CDE에 대한 품질 차원 적용
3. 측정 가능한 지표와 임계값 정의
4. 측정 빈도와 보고 체계 수립

5. 개선 목표와 책임자 지정

#### 6.3.1.1 표준 KPI 카탈로그

| KPI | 계산식 | 임계값 예시 |
|---|---|---|
| 완전성 점수 | (필수 필드 충족 레코드 / 전체) × 100 | ≥ 95% (양호), < 80% (위험) |
| 정확성 점수 | (형식·범위 준수 레코드 / 전체) × 100 | ≥ 98% (양호), < 90% (위험) |
| 일관성 점수 | (시스템 간 일치 레코드 / 전체) × 100 | ≥ 95% (양호), < 85% (위험) |
| 중복률 | (중복 레코드 / 전체) × 100 | ≤ 2% (양호), > 5% (위험) |
| 적시성(지연) | 평균 데이터 갱신 지연 시간 | ≤ SLA 기준 (도메인별) |
| 오류율 추세 | 월별 오류 발견 건수 변화 | 지속적 감소 추세 |

### 6.3.2 데이터 품질 평가 방법

데이터 품질 평가는 정량적 평가와 정성적 평가를 결합한다. English (1999) 가 제시한 TIQM(Total Information Quality Management)은 비용/효과 측면에서 품질을 평가하는 종합적 접근법이다.

#### 6.3.2.1 정량적 평가(Quantitative Assessment)

- 규칙 기반 검증: SQL/스크립트로 자동 검증
- 통계적 분석: 분포·이상치·추세 분석
- 샘플링 기반 평가: 대용량 데이터의 표본 검증

#### 6.3.2.2 정성적 평가(Qualitative Assessment)

- 사용자 만족도 조사: 데이터 사용자의 인식 평가
- 도메인 전문가 검토: 비즈니스 적합성 평가
- 프로세스 평가: 데이터 생성·관리 프로세스 성숙도

### 6.3.3 데이터 정제 방법론

데이터 정제는 식별된 품질 문제를 해결하여 데이터를 사용 가능한 상태로 만드는 활동이다. Loshin (2010b) 은 정제를 발견적(discovery), 표준화(standardization), 검증(verification), 강화(enhancement) 네 단계로 구분한다.

#### 6.3.3.1 1단계: 발견 및 진단

- 데이터 프로파일링 수행
- 이상치, 결측값, 중복 식별
- 형식·규칙 위반 탐지

#### 6.3.3.2 2단계: 표준화

- 표기 형식 통일 (대소문자, 공백, 특수문자)
- 주소·전화번호·이름 등 표준 형식 적용
- 코드값 표준화 (국가 코드, 성별 코드 등)

#### 6.3.3.3 3단계: 검증 및 보정

- 외부 신뢰 소스와 대조 검증
- 규칙 기반 자동 보정
- 의심 레코드의 수동 검토

#### 6.3.3.4 4단계: 강화 및 보완

- 결측값 보완 (통계적·룰 기반·외부 데이터)
- 파생 속성 추가
- 관련 데이터 연결

> ⚠ 자동 정제의 한계와 위험
>
> 자동 정제는 효율적이지만 잘못 사용하면 데이터의 의미를 손상시킬 수 있다. 예를 들어 '주소가 비어있는 고객의 우편번호를 00000으로 일괄 보정'하면 통계 분석 시 심각한 왜곡이 발생할 수 있다. 정제 규칙은 반드시 도메인 전문가의 검토를 거치고, 변경 이력은 추적 가능해야 한다.

### 6.3.4 데이터 품질 거버넌스

기술적 도구만으로는 데이터 품질을 유지할 수 없다. **조직적·프로세스적 거버넌스 체계**가 함께 작동해야 한다. Ladley (2019) 은 거버넌스 없는 품질 관리를 '구멍 난 양동이에 물을 붓는 것'에 비유한다.

| 역할 | 책임 | 권한 수준 |
|---|---|---|
| 데이터 오너 | 데이터의 비즈니스 정의·품질 기준 결정 | 임원/부서장 |
| 데이터 스튜어드 | 일상적 품질 관리·이슈 처리 | 실무 책임자 |
| 데이터 품질 분석가 | 품질 측정·분석·보고 | 전문가 |
| 데이터 사용자 | 이슈 보고·피드백 | 전 임직원 |
| DQ 위원회 | 표준 승인·우선순위 결정 | 교차 부서 |

## 6.4 실무 적용

### 6.4.1 데이터 품질 도구 선정

Gartner (2024b) 에 따른 2024년 기준 주요 데이터 품질 도구는 다음과 같이 분류된다.

| 구분 | 도구 | 특징 |
|---|---|---|
| 엔터프라이즈 상용 | Informatica DQ, IBM InfoSphere, SAS DQ, Talend | 포괄적 기능, 거버넌스 통합 |
| 오픈소스 | Great Expectations, Soda, Deequ, dbt tests | 코드 기반, 낮은 비용 |
| 클라우드 네이티브 | AWS Glue Data Quality, Azure Purview, GCP Dataplex | 클라우드 인프라 통합 |
| AI 기반 신규 | Monte Carlo, Anomalo, Lightup, Datafold | 이상 탐지 자동화 |
| 프로파일링 특화 | Talend Open Studio, Apache Griffin | 데이터 탐색 강점 |
| 관측성(Observability) | Bigeye, Acceldata, Validio | 파이프라인 모니터링 |

#### 6.4.1.1 도구 선정 체크리스트

- 데이터 소스 지원 범위(RDBMS, 클라우드, 빅데이터)
- 규칙 정의 방식(GUI, SQL, Python, YAML)
- 자동 이상 탐지(통계·ML 기반)

- 워크플로우 및 알람 기능
- 리포팅 및 대시보드
- 기존 데이터 카탈로그·메타데이터 도구와의 통합
- CI/CD 파이프라인 통합 가능성

### 6.4.2 단계적 도입 로드맵

| 단계 | 기간 | 주요 활동 |
|---|---|---|
| Stage 1: 파일럿 | 0~3개월 | 핵심 자산 1~2개 도메인, 6대 차원 KPI 정의 |
| Stage 2: 확장 | 3~9개월 | 5~10개 도메인 확장, 품질 대시보드 |
| Stage 3: 통합 | 9~18개월 | ETL/ELT 파이프라인에 품질 검증 통합 |
| Stage 4: 최적화 | 18개월~ | ML 기반 이상 탐지, 예측적 품질 관리 |

### 6.4.3 실무 팁

**팁 1: 핵심 데이터에 집중하라**

조직의 모든 데이터를 동일한 강도로 관리할 수는 없다. Critical Data Elements(CDE)를 식별하고 그 데이터에 80%의 자원을 집중하라. Pareto 법칙처럼 보통 핵심 데이터 자산 20%가 비즈니스 가치의 80%를 만든다 (Loshin 2010b). 마스터 데이터, 재무 데이터, 규제 보고 데이터가 대표적 CDE다.

**팁 2: 측정만 하지 말고 행동하라**

대시보드만 만들어두고 아무도 보지 않는 데이터 품질 프로젝트가 매우 많다. 측정 결과에 따른 명확한 책임자와 액션 아이템이 정의되어야 한다. '완전성이 90% 미만이면 누가 며칠 안에 어떤 조치를 한다'는 식의 운영 규칙이 필요하다.

**팁 3: 데이터 품질을 코드로 다루라**

현대의 DataOps 접근법은 데이터 품질 규칙을 코드(test as code)로 관리한다. Great Expectations, dbt tests, Soda 같은 도구는 품질 검증을 Git으로 버전 관리하고 CI/CD 파이프라인에서 자동 실행하게 해준다 (Cataldo and Foote 2022). 이는 품질 규칙의 추적성·재사용성을 크게 높인다.

**주의: 완벽주의의 함정**

100% 완벽한 데이터 품질은 비현실적인 목표일 뿐 아니라, 종종 비효율적이다. 데이터 정확도를 99.9%에서 99.99%로 올리는 비용이 종종 비즈니스 가치를 초과한다. 사용 목적에 적합한 수준(fit for purpose)이 진정한 목표이며, 완벽함이 아니다.

## 6.5 사례 연구

### 6.5.1 사례 1: 글로벌 통신사 V사의 고객 데이터 품질 개선

#### 6.5.1.1 배경

글로벌 통신사 V사는 1억 명 이상의 고객 데이터를 보유하고 있었지만, 기존에 인수한 다양한 자회사로부터 통합된 고객 데이터에 심각한 품질 문제가 있었다. 중복 고객, 잘못된 연락처, 결측된 인구통계 정보로 인해 마케팅 캠페인의 성과가 저조했다.

#### 6.5.1.2 도전 과제

- 중복 고객 추정 비율: 18%
- 유효 연락처 보유율: 71%
- 개인정보 보호 규제(GDPR) 준수 요구
- 실시간 마케팅 자동화 추진

#### 6.5.1.3 접근 방법

1. 6대 차원 기반 품질 진단 (3개월)
2. Informatica DQ 도입 및 정제 규칙 자동화
3. 엔티티 해소(Entity Resolution) 알고리즘 적용
4. 실시간 데이터 검증 API 구축
5. 데이터 스튜어드 50명 양성

#### 6.5.1.4 성과

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 중복 고객률 | 18% | 1.2% |
| 유효 연락처 비율 | 71% | 94% |
| 캠페인 응답률 | 기준 100 | 187 |
| GDPR 응대 시간 | 수주 | 수일 |

#### 6.5.1.5 교훈

데이터 품질은 단순히 IT 프로젝트가 아닌 비즈니스 변화 관리 프로젝트다. 마케팅 ROI 향상이라는 명확한 비즈니스 가치를 제시하고, 데이터 스튜어드 양성을 통해 조직 차원의 책임 의식을 확보한 것이 성공의 핵심이었다.

### 6.5.2 사례 2: 금융기관의 BCBS 239 준수 사례

#### 6.5.2.1 배경

글로벌 시스템적 중요 은행(G-SIB)들은 2013년 발표된 Basel Committee on Banking Supervision (2013) '효과적인 리스크 데이터 집계 및 리스크 보고 원칙'을 준수해야 한다. 이 원칙은 14개의 데이터 거버넌스 원칙을 정의하며, 데이터 정확성·완전성·적시성에 대해 엄격한 요구사항을 부과한다.

#### 6.5.2.2 핵심 요구사항

- 리스크 데이터의 정확성·무결성·완전성 보장
- 적시 위기 대응을 위한 신속한 집계
- 리스크 보고서의 명확성·유용성
- 이러한 활동에 대한 거버넌스 책임

#### 6.5.2.3 주요 구현 사항

1. 리스크 데이터에 대한 명시적 SLA 정의 (예: 일 마감 후 4시간 내 보고서)
2. 데이터 계보를 통한 모든 리스크 데이터의 출처 추적
3. 자동화된 데이터 조정(Reconciliation) 시스템
4. 실시간 데이터 품질 대시보드
5. 이사회 수준의 데이터 거버넌스 위원회

#### 6.5.2.4 교훈

규제 준수는 데이터 품질 투자의 강력한 동기가 된다. BCBS 239처럼 명시적이고 검증 가능한 외부 요구사항은 조직 내 데이터 품질 문화 정착을 가속한다. 단, 단순히 규제 준수만 목표로 하면 '체크박스 컴플라이언스'에 그치므로 비즈니스 가치 창출과 연계해야 한다.

### 6.5.3 사례 3: 헬스케어 데이터 품질의 환자 안전

#### 6.5.3.1 배경

미국의 한 대형 의료 시스템은 환자 식별 오류와 약물 처방 데이터 오류가 환자 안전 사고로 이어진다는 사실에 주목하여 전사적 데이터 품질 개선 프로그램을 시작했다. Makary and Daniel (2016) 의 연구에 따르면 미국에서 의료 오류로 인한 사망은 연간 약 25만 명에 달한다.

#### 6.5.3.2 접근 방법

1. 환자 마스터 인덱스(EMPI) 구축 — 환자 식별 일관성 확보
2. HL7 FHIR 표준 채택 — 시스템 간 데이터 일관성
3. 약물 데이터베이스 표준화 (RxNorm)
4. 실시간 데이터 검증 (처방 시점 약물 상호작용 확인)
5. 의료진을 위한 데이터 품질 교육

#### 6.5.3.3 성과

| 지표 | 변화 |
|---|---|
| 환자 식별 오류 | 87% 감소 |
| 처방 오류 | 64% 감소 |
| 환자 만족도 | 31% 향상 |
| 의료 분쟁 | 45% 감소 |

#### 6.5.3.4 교훈

데이터 품질이 직접적으로 인간의 생명·안전과 연결되는 영역에서는 기술적 통제만으로는 충분하지 않다. 데이터를 입력하는 사용자(의사·간호사)의 인식 변화와 실시간 검증의 결합이 필수적이다. 데이터 품질은 결국 '사람의 문제'이기도 하다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **데이터 품질의 본질** — 데이터 품질은 '의도된 사용 목적에 적합한 정도(fitness for use)'이며, 사용 맥락에 따라 의미가 달라지는 상대적 개념이다. 단일 절대 기준이 아닌 다차원적 평가가 필요하다.
2. **6대 품질 차원** — 완전성·정확성·일관성·유효성·유일성·적시성의 6대 차원은 데이터 품질 측정의 표준 프레임워크다. 각 차원은 측정 가능한 KPI로 구체화되어야 한다.

3. **국제 표준 활용** — ISO (2022b) (관리 프로세스)과 ISO/IEC (2008) (측정 모델)는 데이터 품질의 두 축을 이루는 국제 표준이며, 조직 차원의 품질 체계 수립 시 활용할 수 있다.
4. **지속 가능한 사이클** — 데이터 품질 관리는 일회성 정제 작업이 아니라 정의·측정·분석·개선·통제의 지속적 사이클이다. 근본 원인 해결에 집중해야 진정한 개선이 가능하다.
5. **거버넌스와 도구의 결합** — 기술적 도구만으로는 품질을 유지할 수 없다. 데이터 오너, 스튜어드, 사용자가 명확한 역할로 참여하는 거버넌스 체계가 함께 작동해야 한다.
6. **비즈니스 가치 중심 접근** — 완벽한 품질이 아닌 비즈니스 가치 기준의 적정 품질을 추구해야 한다. 핵심 데이터(CDE) 식별과 ROI 기반 우선순위 설정이 핵심이다.

## 다음 학습 가이드

### 6.5.4 국제 표준

- ISO (2022b) — 데이터 품질 관리 표준 시리즈
- ISO/IEC (2008) — 데이터 품질 모델
- Basel Committee on Banking Supervision (2013) — 효과적 리스크 데이터 집계 원칙

### 6.5.5 관련 본서 참조 장

- 제 5장 메타데이터 관리 — 품질 메타데이터
- 제 7장 마스터 데이터 관리 — 마스터 데이터 품질
- 제 9장 데이터 거버넌스 — 품질 거버넌스 체계
- 제 12장 AI·ML과 데이터 표준화 — AI 시대 품질 차원

## 연습 문제

1. 여러분 조직의 핵심 데이터 자산(CDE) 3~5개를 식별하고, 각각에 대해 6대 품질 차원별 KPI를 설계하시오.
2. 데이터 품질 문제로 인해 발생한 비즈니스 영향 사례를 분석하고, 근본 원인(root cause)을 추적하시오.
3. 자동 정제와 수동 검토의 적절한 균형점은 어디인가? 구체적인 데이터 유형별로 논하시오.

CHAPTER 7

# 마스터 데이터 관리(MDM)

## 7.1 개요

마스터 데이터(Master Data)는 조직의 비즈니스를 구성하는 **핵심 실체(entity)에 대한 데이터**이다. 고객, 제품, 조직, 위치, 계정과 같이 여러 비즈니스 프로세스와 시스템에서 반복적으로 참조되는 '명사적' 정보가 마스터 데이터의 전형이다. Loshin (2010a) 은 Master Data Management 에서 마스터 데이터를 '조직의 핵심 비즈니스 개념에 대한 공유된 진실'이라 정의한다.

트랜잭션 데이터(주문·거래·이벤트 등)는 매일 무수히 생성되지만, 그 트랜잭션이 참조하는 **'누가(who), 무엇을(what), 어디서(where)'**에 대한 정보는 상대적으로 안정적이다. 이 안정적이면서도 근본적인 정보가 잘못되면 그 위에 쌓이는 모든 트랜잭션과 분석이 왜곡된다. 이것이 마스터 데이터 관리(MDM)의 존재 이유이다.

Gartner (2024f) 의 조사에 따르면 **마스터 데이터 품질 문제로 인한 조직의 연간 손실은 평균 매출의 12~15%**에 달한다. 중복 고객 레코드는 마케팅 낭비를, 제품 정보 불일치는 재고 부정확을, 공급업체 마스터 중복은 사기 리스크를 야기한다. Berson and Dubov (2011) 은 'MDM은 IT 프로젝트가 아니라 비즈니스 프로세스 개선 프로젝트'라고 강조한다.

### 7.1.1 트랜잭션 데이터 vs 마스터 데이터 vs 참조 데이터

데이터 아키텍처에서 세 가지 데이터 유형은 명확히 구별되며, 각기 다른 관리 방식을 요구한다 (Loshin 2010a; Cervo and Allen 2011).

| 구분 | 정의 | 예시 | 관리 특성 |
|---|---|---|---|
| 마스터 데이터 | 핵심 비즈니스 실체 정보 | 고객, 제품, 공급업체, 조직 | 상대적 안정, 다양한 시스템 참조 |
| 트랜잭션 데이터 | 비즈니스 이벤트·활동 기록 | 주문, 결제, 재고 이동, 로그 | 빈번한 생성, 시간 종속적 |

| 구분 | 정의 | 예시 | 관리 특성 |
|---|---|---|---|
| 참조 데이터 | 코드·분류·표준값 집합 | 국가 코드, 통화, 성별, 우편번호 | 매우 안정, 외부 표준 준수 |

### 7.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 마스터 데이터의 특성과 다른 데이터 유형과의 차이를 설명할 수 있다
- MDM 4대 아키텍처 스타일의 강약점을 비교하고 조직에 맞게 선택할 수 있다
- 골든 레코드(Golden Record) 생성 프로세스를 설계할 수 있다
- 엔티티 해소(Entity Resolution)의 기법과 도구를 활용할 수 있다
- 데이터 스튜어드십 체계를 구축하고 운영할 수 있다
- MDM 도입 로드맵을 수립하고 성공 요인을 관리할 수 있다

## 7.2 핵심 개념

### 7.2.1 마스터 데이터의 5대 도메인

조직에서 MDM의 대상이 되는 마스터 데이터는 다섯 가지 주요 도메인으로 분류된다. Cervo and Allen (2011) 은 각 도메인이 서로 다른 비즈니스 목적과 통합 요구를 가진다고 지적한다.

| 도메인 | 주요 실체 | 비즈니스 임팩트 | 주요 이슈 |
|---|---|---|---|
| Party (당사자) | 고객, 임직원, 공급업체 | 마케팅, HR, 조달 | 중복, 개인정보 |
| Product (제품) | 제품, 서비스, SKU | 판매, 재고, 마케팅 | 속성 관리, SKU 분류 |
| Location (위치) | 주소, 매장, 국가, 지역 | 물류, 서비스 지역 | 표준 주소, 지리 코드 |
| Financial (재무) | 계정, 원가 센터, 화폐 | 재무 보고, 결산 | 계정과목 통합 |
| Reference (참조) | 코드, 분류, 계층 | 모든 시스템의 기준값 | 외부 표준 준수 |

### 7.2.2 MDM 4대 아키텍처 스타일

MDM 시스템은 데이터를 어떻게 저장·통합·배포하느냐에 따라 4가지 아키텍처 스타일로 분류된다. Loshin (2010a) 이 제시한 이 분류는 업계 표준이 되었으며, 조직의 요구·성숙도·기존 시스템에 따라 선택해야 한다.

#### 7.2.2.1 Style 1: Registry (등록소) 방식

각 원천 시스템은 그대로 두고, MDM 시스템은 **'인덱스와 참조 정보만 관리'**한다. 각 엔티티에 대해 시스템 간 매칭·연결 정보를 유지하되 실제 데이터는 이동하지 않는다. 가장 가벼운 접근이며, 원천 시스템 변경을 최소화한다.

- **장점**: 원천 시스템 최소 개입, 빠른 구축, 낮은 리스크
- **단점**: 실시간 조회 필요, 데이터 품질 개선 제한적
- **적합**: 대규모 이질 시스템 초기 통합

#### 7.2.2.2 Style 2: Consolidation (통합) 방식

원천 시스템의 데이터를 **MDM 허브에 복사·통합**하여 골든 레코드를 만든다. 그러나 원천으로의 피드백은 없이 분석·보고용으로만 활용한다. 데이터 웨어하우스와 유사한 접근이다.

- **장점**: 통합된 뷰 제공, 원천 최소 영향, 분석에 최적

- **단점**: 원천 시스템 개선 없음, 실시간성 부족
- **적합**: 분석·보고 중심 MDM, 규제 준수 대응

#### 7.2.2.3 Style 3: Coexistence (공존) 방식

MDM 허브가 골든 레코드를 관리하되, **원천 시스템으로 골든 레코드를 되돌려** 시스템 간 동기화를 유지한다. 원천 시스템은 여전히 데이터를 생성·수정할 수 있다.

- **장점**: 원천-허브 양방향 동기화, 점진적 개선
- **단점**: 복잡한 동기화 로직, 충돌 처리 필요
- **적합**: 단계적 MDM 성숙 조직, 다양한 이해관계자

#### 7.2.2.4 Style 4: Centralized (중앙집중) 방식

모든 마스터 데이터의 **생성·수정·삭제(CUD)를 MDM 허브에서만 수행**하고, 원천 시스템은 이를 참조만 한다. 가장 엄격한 통제이지만 구축 비용도 가장 크다.

- **장점**: 완전한 데이터 일관성, 최고 수준 품질
- **단점**: 높은 구축 비용, 조직 변화 부담
- **적합**: 규제 산업, 성숙한 조직, 신규 시스템 구축

> **i** 아키텍처 스타일 선택 지침
>
> 어느 스타일이 절대적으로 우수하지는 않다. 다음 기준으로 결정하라:
>
> - **원천 시스템 개입 허용도**: 낮음 → Registry, 높음 → Centralized
> - **데이터 품질 목표**: 분석용 → Consolidation, 운영 개선 → Coexistence/Centralized
> - **조직 성숙도**: 초기 → Registry, 성숙 → 상위 단계로 진화
> - **실무 팁**: Registry로 시작해 Coexistence로 진화하는 조직이 많다

### 7.2.3 골든 레코드(Golden Record)

골든 레코드는 특정 실체(예: 고객 홍길동)에 대해 조직이 참조하는 **유일하고 신뢰할 수 있는 표준 표현**이다. 여러 시스템에 중복·상이하게 존재하는 데이터로부터 '가장 정확하고 완전한' 하나의 레코드를 도출하는 것이 MDM의 핵심 산출물이다 (Loshin 2010a). 골든 레코드 생성은 다음 4단계로 진행된다.

1. **식별(Identification)** — 동일 실체에 해당하는 레코드 후보들의 매칭
2. **통합(Consolidation)** — 매칭된 레코드들의 속성 병합
3. **정제(Cleansing)** — 형식 표준화, 이상치 보정
4. **풍부화(Enrichment)** — 외부 데이터로 정보 보완

### 7.2.4 엔티티 해소(Entity Resolution)

엔티티 해소는 **여러 원천의 레코드 중 동일 실체를 가리키는 것을 식별**하는 문제이다. 예: 'John Smith', 'J. Smith', 'Smith, John'이 같은 사람인가? Christen (2012) 의 저서 Data Matching 은 이 문제의 기술적 접근을 포괄적으로 다룬다.

| 접근법 | 설명 | 장단점 |
|---|---|---|
| 결정적 매칭 | 완전 일치 규칙 기반 (예: SSN 동일) | 정확하나 커버리지 낮음 |
| 확률적 매칭 | Fellegi-Sunter 모델 등 통계적 가중치 | 커버리지 넓지만 임계값 튜닝 필요 |

| 접근법 | 설명 | 장단점 |
|---|---|---|
| ML 기반 | 지도 학습으로 매칭 여부 예측 | 정교하지만 학습 데이터 필요 |

Fellegi-Sunter 모델

1969년 Fellegi and Sunter (1969) 가 제안한 확률적 레코드 매칭의 이론적 기반이다. 각 속성의 일치·불일치에 확률 가중치를 부여하고, 전체 점수로 매칭 여부를 결정한다. 특정 속성이 우연히 일치할 확률(m-probability, u-probability)을 활용하여 정보량이 많은 속성에 더 높은 가중치를 준다. 현대의 확률적 매칭 엔진(Splink, Zingg 등)이 이 이론을 기반으로 구현되어 있다.

### 7.2.5 매치·머지·서바이버십

MDM의 핵심 프로세스는 매치(Match), 머지(Merge), 서바이버십(Survivorship)의 세 단계로 진행된다 (Loshin 2010a; Berson and Dubov 2011).

#### 7.2.5.1 Match: 매칭 규칙

동일 실체 후보 식별을 위한 규칙 설계. 이름·주소·연락처 유사도, 정확 일치, 표준화 후 비교 등. 성능(recall, precision) 균형이 관건이다.

#### 7.2.5.2 Merge: 병합 로직

매칭된 여러 레코드를 하나로 병합. 각 속성별로 어떤 원천의 값을 채택할지의 규칙이 필요하다.

#### 7.2.5.3 Survivorship: 서바이버십 규칙

속성별로 '어느 원천의 값이 우선하는가'를 결정하는 규칙. 예: '주소는 CRM 최신, 연락처는 콜센터 시스템 우선'. 이 규칙은 도메인 전문가와 협의하여 설계해야 한다.

| 규칙 유형 | 설명 | 예시 |
|---|---|---|
| Source-based (원천 기준) | 특정 원천을 우선 | CRM 〉 ERP 〉 마케팅 |
| Recency (최신성) | 가장 최근 데이터 선택 | last_updated 기준 |
| Completeness (완전성) | 값이 있는 데이터 우선 | NULL 아닌 것 우선 |
| Frequency (빈도) | 가장 자주 등장하는 값 | 다수결 |
| Trust score | 속성별 신뢰도 점수 기반 | 종합 점수 계산 |

## 7.3 방법론

### 7.3.1 MDM 도입 절차

MDM 도입은 다음 7단계 절차를 따른다. 각 단계는 명확한 산출물을 가져야 다음 단계로 진행할 수 있다 (Loshin 2010a; Berson and Dubov 2011; Plotkin 2020).

1. **도메인 선정 및 우선순위** — 고객·제품·공급업체 중 어디서 시작할지
2. **현행 진단** — 원천 시스템 매핑, 데이터 프로파일링, 중복률 분석
3. **표준 정의** — 마스터 데이터 모델, 속성 정의, 코드 표준
4. **아키텍처 결정** — 4대 스타일 중 선택
5. **도구 선정 및 POC** — 상용/오픈소스 MDM 도구
6. **파일럿 구현** — 1개 도메인, 3~5개 시스템 통합

7. **확장 및 운영** — 전사 확대, 스튜어드십 정착

### 7.3.2 데이터 스튜어드십

MDM의 성공은 기술이 아니라 사람에 달렸다. Plotkin (2020) 은 Data Stewardship 에서 스튜어드십을 '조직이 데이터를 자산으로 관리하기 위해 데이터 사용자들이 수행하는 활동'이라 정의한다. 마스터 데이터에는 다음 세 유형의 스튜어드가 필요하다.

| 스튜어드 유형 | 역할 | 예시 담당 |
|---|---|---|
| Business Steward | 비즈니스 정의·규칙 관리 | 마케팅 부서장 (고객 도메인) |
| Technical Steward | 기술적 구현·통합 관리 | 데이터 아키텍트, DBA |
| Operational Steward | 일상 운영·이슈 처리 | CRM 시스템 관리자 |

### 7.3.3 마스터 데이터 품질 측정

마스터 데이터의 품질은 특별한 방식으로 측정된다. 전통적 6대 품질 차원에 더해 '유일성(중복 없음)'과 '식별성(정확한 매칭)'이 특히 중요하다 (Loshin 2010a).

#### 7.3.3.1 MDM 특화 KPI

| KPI | 정의 | 임계값 예시 |
|---|---|---|
| 중복률(Duplication Rate) | 동일 실체가 중복 표현된 비율 | ≤ 1% |
| 매칭 정확도(Match Accuracy) | 매칭 규칙의 정밀도(Precision) | ≥ 95% |
| 매칭 회수율(Match Recall) | 매칭 규칙의 재현율(Recall) | ≥ 90% |
| 골든 레코드 완성도 | 필수 속성 채움 비율 | ≥ 95% |
| 교차 시스템 일치율 | 시스템 간 마스터 데이터 일치 | ≥ 98% |
| 스튜어드 승인 대기 시간 | 평균 승인 소요 시간 | ≤ 24시간 |

## 7.4 실무 적용

### 7.4.1 MDM 도구 선정

2024년 기준 주요 MDM 도구는 다음과 같이 분류된다. Gartner (2024c) 는 리더 그룹으로 Informatica, IBM, SAP, Oracle, Reltio를 지목한다.

| 구분 | 주요 도구 | 특징 |
|---|---|---|
| 엔터프라이즈 상용 | Informatica MDM, IBM InfoSphere, SAP MDG, Oracle MDM | 포괄적, 검증된 안정성 |
| 클라우드 네이티브 | Reltio, Semarchy, Profisee (Cloud) | 빠른 배포, SaaS 모델 |
| 산업 특화 | Stibo Systems (PIM), Riversand (제품) | 특정 도메인 최적화 |
| 오픈소스 | Pimcore, Talend Open Studio, Splink (매칭) | 낮은 비용, 유연성 |

### 7.4.2 실무 팁

> 팁 1: 고객 도메인부터 시작하라
>
> MDM 도입 도메인 우선순위는 조직 상황에 따라 다르지만, **고객(Customer) 도메인**이 가장 흔한 시작점이다. 이유는 (1) 비즈니스 임팩트 가시화 용이, (2) 중복 문제가 매우 명확, (3) 마케팅 ROI 개선 등 정량화 쉬움, (4) 이해관계자 협력 유도 용이. 첫 프로젝트의 성공이 이후 확장의 기반이 된다 (Berson and Dubov 2011).

> 팁 2: 매칭 규칙은 반복적으로 조정하라
>
> 완벽한 매칭 규칙을 처음에 설계하려 하지 말라. 초기 규칙을 배포한 후 False Positive(오매칭)와 False Negative(누락)를 검토하고 규칙을 점진적으로 조정하라. 일반적으로 3~6개월의 튜닝 기간이 필요하며, 이 기간의 데이터 스튜어드 검토가 매우 중요하다.

> 팁 3: 원천 시스템을 무리하게 바꾸지 말라
>
> MDM 도입 초기에 원천 시스템의 대규모 개편을 시도하면 프로젝트가 실패할 확률이 매우 높다. **Registry나 Consolidation 방식으로 시작**하여 원천 시스템 개입을 최소화하고, 조직이 MDM에 익숙해진 후 점진적으로 Coexistence나 Centralized로 이동하라 (Loshin 2010a). Big Bang은 대부분 Big Fail로 이어진다.

> 팁 4: 스튜어드 인센티브를 설계하라
>
> 데이터 스튜어드는 종종 본업 외 겸직으로 부여된다. 별도 인센티브 없이는 형식적 활동에 그치기 쉽다. 우수 스튜어드 표창, KPI에 반영, 명시적 시간 할당(예: 주 4시간) 등의 지원 체계가 필요하다. Ladley (2019) 은 '보상 없는 스튜어드십은 강제 노동'이라 표현한다.

> ⚠ 주의: 정치적 이슈로서의 마스터 데이터
>
> '우리 부서가 우리 고객 데이터의 오너'라는 부서별 주장이 흔하다. MDM은 필연적으로 부서 간 권한과 책임을 재조정한다. 이는 순수 기술 이슈가 아니라 조직 정치의 문제이다. 경영진 후원 없이는 어떤 아키텍처도 정착시킬 수 없다. 특히 오너십 갈등을 조정할 명확한 거버넌스 구조가 프로젝트 시작 전에 확립되어야 한다.

> ⚠ 주의: 개인정보 보호와 MDM
>
> 고객 마스터의 통합은 개인정보 보호법의 핵심 이슈이다. 여러 시스템의 고객 정보를 단일 골든 레코드로 통합하는 것 자체가 개인정보 처리 목적 확대에 해당할 수 있다. GDPR의 '목적 제한 원칙'과 국내 개인정보보호법의 '수집 목적' 제약을 반드시 확인하고, 필요 시 법무팀과 협의하라 (European Data Protection Board 2020a). 특히 마이데이터 시대에는 MDM과 개인정보 관리의 접점 이슈가 더욱 중요해진다.

## 7.5 사례 연구

### 7.5.1 사례 1: 글로벌 통신사 V사의 고객 MDM

#### 7.5.1.1 배경

글로벌 통신사 V사는 인수합병으로 1억 명 이상의 고객 데이터를 여러 국가의 CRM·빌링·마케팅 시스템에 보유하고 있었다. 동일 고객이 시스템별로 다른 ID와 상이한 정보로 존재해 옴니채널 마케팅과 개인화 서비스가 불가능한 상황이었다 (Forrester Consulting 2022).

#### 7.5.1.2 도전 과제

- 15개 국가의 자회사, 30개 이상의 원천 시스템
- 고객 중복률 추정 18%
- 유효 연락처 보유율 71%
- GDPR 준수 요구 (2018년 마감)

#### 7.5.1.3 접근 방법

1. Coexistence 아키텍처 선택 — 원천 유지, 골든 레코드 관리
2. Informatica MDM 도입
3. 확률적 매칭 (이름, 이메일, 전화, 주소 조합)
4. 서바이버십 규칙: CRM 우선, 결측 시 빌링, 최신성
5. 데이터 스튜어드 50명 양성
6. 실시간 데이터 검증 API 구축

#### 7.5.1.4 성과 (18개월)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 중복 고객률 | 18% | 1.2% |
| 유효 연락처 비율 | 71% | 94% |
| 캠페인 응답률 | 기준 100 | 187 |
| GDPR 응대 시간 | 수주 | 수일 |
| 교차 판매 성공률 | 3.2% | 8.7% |

#### 7.5.1.5 교훈

MDM의 성과는 마케팅 ROI라는 명확한 비즈니스 지표로 측정되어야 한다. V사는 캠페인 응답률과 교차 판매 성공률 향상으로 MDM 투자를 정당화했다. 또한 GDPR이라는 규제 마감이 프로젝트 추진력을 유지하는 역할을 했다. 규제 준수와 비즈니스 가치를 동시에 제시할 때 MDM은 가장 강력한 명분을 얻는다.

### 7.5.2 사례 2: 글로벌 제조업체 P사의 제품 MDM

#### 7.5.2.1 배경

글로벌 제조업체 P사는 15개국에서 50,000개 SKU를 생산·판매한다. 제품 정보가 국가별·시스템별로 상이하여 글로벌 카탈로그 구축과 이커머스 확장이 심각한 병목이 되었다. Product Information Management(PIM) 도입이 시급했다 (Stibo Systems 2023).

#### 7.5.2.2 도전 과제

- 국가별 상이한 제품 명명 규칙

- 이커머스용 정보(사진, 마케팅 카피)와 ERP 정보 분리
- 규제 정보(성분, 안전)의 국가별 요구 차이
- 50,000 SKU 마이그레이션 부담

#### 7.5.2.3 접근 방법

1. Stibo Systems PIM 도구 도입
2. 계층적 제품 분류 체계 (5단계)
3. 속성 카탈로그 표준화 — 300+ 표준 속성
4. 워크플로우: 상품 등록 → 스튜어드 검토 → 승인
5. 이커머스·ERP·마케팅 시스템 자동 배포

#### 7.5.2.4 성과

- 신제품 출시 리드타임 8주 → 3주
- 제품 정보 오류로 인한 고객 문의 62% 감소
- 15개국 통합 카탈로그 완성
- 이커머스 매출 47% 증가

#### 7.5.2.5 교훈

제품 MDM(PIM)은 고객 MDM과는 다른 성격을 가진다. 매칭·중복 제거보다 '풍부한 속성 관리와 유통 채널 배포'가 핵심이다. 제품 데이터의 이커머스 활용이 확산되면서 PIM은 별도 성장 시장이 되었으며, 일반 MDM 도구가 아닌 PIM 특화 도구(Stibo, Riversand, Pimcore 등)가 더 효과적인 경우가 많다.

### 7.5.3 사례 3: 국내 대형 은행 K사의 공급업체 MDM

#### 7.5.3.1 배경

국내 대형 은행 K사는 3만 개 이상의 공급업체와 거래하고 있으나, 공급업체 정보가 여러 시스템(구매, 회계, 계약)에 분산되어 있어 중복 등록과 지급 오류가 빈번했다. 특히 사기 리스크와 규제 대응 부담이 커지면서 공급업체 MDM 도입을 결정했다 (금융보안원 2023).

#### 7.5.3.2 주요 이슈

- 동일 공급업체의 중복 등록 (추정 15%)
- 사업자등록번호 오등록으로 인한 세무 이슈
- 규제 대상(제재 국가·개인) 스크리닝 필요
- 결산 시 공급업체 총계 산출 오류

#### 7.5.3.3 접근 방법

1. Centralized 아키텍처 (규제 산업 특성)
2. 사업자등록번호를 유일 키로 지정
3. 공급업체 등록 워크플로우 표준화
4. 외부 신용정보 자동 연계
5. 제재 리스트 자동 스크리닝 (Refinitiv World-Check)
6. KYS(Know Your Supplier) 정기 재검증

#### 7.5.3.4 성과

- 공급업체 중복률 15% → 0.5%
- 지급 오류 90% 감소
- 규제 스크리닝 100% 자동화
- 결산 시간 3일 → 4시간

#### 7.5.3.5 교훈

공급업체 MDM은 사기 방지·규제 준수라는 명확한 비즈니스 가치가 있어 ROI 정당화가 용이하다. 규제 산업(금융, 의료)에서는 Centralized 아키텍처가 적합한 경우가 많으며, 외부 신용정보·제재 리스트와의 연계가 필수적이다. 특히 최근 확산되는 ESG 규제(공급망 실사, EU CSDDD 등)로 인해 공급업체 MDM의 전략적 중요성이 더욱 커지고 있다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **마스터 데이터의 본질** — 마스터 데이터는 조직의 핵심 실체(고객·제품·공급업체·조직·위치)에 대한 공유된 진실이다. 잘못된 마스터 데이터는 그 위에 쌓이는 모든 트랜잭션과 분석을 왜곡한다.
2. **5대 도메인의 특성 이해** — Party·Product·Location·Financial·Reference 5대 도메인은 각기 다른 비즈니스 목적과 관리 특성을 가진다. 도메인별로 적절한 접근이 필요하다.
3. **4대 아키텍처 스타일** — Registry·Consolidation·Coexistence·Centralized의 4대 스타일은 원천 시스템 개입도와 데이터 통제 강도에 따라 다르다. 조직 상황에 맞게 선택하고 점진적으로 진화시켜야 한다.
4. **골든 레코드 생성** — 식별·통합·정제·풍부화의 4단계로 골든 레코드를 만든다. 서바이버십 규칙을 통해 여러 원천의 속성 중 어떤 값을 채택할지 결정한다.
5. **엔티티 해소 기법** — 결정적 매칭·확률적 매칭(Fellegi and Sunter (1969))·ML 기반 매칭 중 조직 상황에 맞는 접근을 선택한다. 규칙은 반복적 튜닝을 통해 개선해야 한다.
6. **스튜어드십의 중요성** — MDM 성공은 기술이 아닌 사람에 달렸다. Business·Technical·Operational 세 유형의 스튜어드가 필요하며, 인센티브 설계가 필수적이다.
7. **비즈니스 가치 중심 접근** — MDM은 IT 프로젝트가 아니라 비즈니스 프로세스 개선 프로젝트이다. 고객 마케팅 ROI, 제품 카탈로그 확장, 공급업체 리스크 관리 등 명확한 비즈니스 지표로 성과를 관리해야 한다.

## 다음 학습 가이드

### 7.5.4 필수 참고 문헌

- Loshin (2010a) — Master Data Management (MDM 이론의 고전)
- Berson and Dubov (2011) — Master Data Management and Data Governance
- Cervo and Allen (2011) — Master Data Management in Practice
- Plotkin (2020) — Data Stewardship
- Christen (2012) — Data Matching (엔티티 해소 기술서)

### 7.5.5 국제 표준

- ISO (2022b) — 마스터 데이터 품질 관련 파트
- ISO/IEC (2023b) — 데이터 요소 표준

### 7.5.6 관련 본서 참조 장

- 제 5장 메타데이터 관리 — MDM 메타데이터
- 제 6장 데이터 품질 관리 — 마스터 데이터 품질
- 제 9장 데이터 거버넌스 체계 — 스튜어드십 조직
- 제 10장 표준화 도구와 기술 — MDM 도구
- 제 11장 실제 적용 사례 — 산업별 MDM 사례

## 연습 문제

1. 여러분 조직의 5대 마스터 데이터 도메인을 매핑하고, 각 도메인의 중복률·품질 이슈를 진단하시오.
2. 여러분 조직에 MDM을 도입한다면 4대 아키텍처 스타일 중 어느 것이 적합한가? 그 이유는 무엇인가?
3. 고객 도메인의 서바이버십 규칙(주소·연락처·이름 등 5~7개 속성)을 설계하시오.

# III 3부: 실무 적용


8 표준화 구현 전략 . . . . . 91
8.1 개요 . . . . . 91
8.2 핵심 개념 . . . . . 92
8.3 방법론 . . . . . 93
8.4 실무 적용 . . . . . 96
8.5 사례 연구 . . . . . 98

9 데이터 거버넌스 체계 . . . . . 101
9.1 개요 . . . . . 101
9.2 핵심 개념 . . . . . 102
9.3 방법론 . . . . . 104
9.4 실무 적용 . . . . . 106
9.5 실무 적용 . . . . . 108
9.6 사례 연구 . . . . . 108

10 표준화 도구와 기술 . . . . . 113
10.1 개요 . . . . . 113
10.2 핵심 개념 . . . . . 114
10.3 방법론 . . . . . 115
10.4 실무 적용 . . . . . 117
10.5 사례 연구 . . . . . 119

11 실제 적용 사례 . . . . . 123
11.1 개요 . . . . . 123
11.2 금융 산업 사례 . . . . . 124
11.3 공공 부문 사례 . . . . . 125
11.4 소매·서비스 산업 사례 . . . . . 127
11.5 헬스케어 산업 사례 . . . . . 129
11.6 제조 산업 사례 . . . . . 130
11.7 종합 교훈 . . . . . 132
11.8 여러분 조직에 적용하기 . . . . . 133

CHAPTER 8

# 표준화 구현 전략

## 8.1 개요

데이터 표준화 구현은 단순한 IT 프로젝트가 아니라 **조직 차원의 변화 관리(Change Management) 활동**이다. 잘 설계된 표준 문서가 책장에 꽂힌 채 아무도 따르지 않는 사례는 너무나 흔하다. Ladley (2019)는 Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program 에서 '표준화 실패의 80%는 기술 문제가 아닌 사람과 프로세스 문제'라고 지적한다.

Forrester Research (2023b) 의 2023년 조사에 따르면, 데이터 표준화 프로젝트의 약 **65%가 초기 목표를 완전히 달성하지 못하며**, 그 중 70%는 기술적 한계가 아닌 조직적 저항, 리더십 부재, 변화 관리 실패 가 원인이다. 즉 **어떤 표준을 만들 것인가**보다 **어떻게 정착시킬 것인가**가 더 중요하다.

### 8.1.1 구현 전략의 핵심 질문

효과적인 표준화 구현 전략을 수립하기 위해서는 다음 다섯 가지 핵심 질문에 답할 수 있어야 한다.

1. WHY — 왜 표준화가 필요한가? 비즈니스 가치는 무엇인가?
2. WHAT — 무엇을 표준화할 것인가? 범위와 우선순위는?
3. HOW — 어떻게 진행할 것인가? 빅뱅 vs 점진적 접근?
4. WHO — 누가 책임지고 누가 참여하는가? R&R은?
5. WHEN — 언제 시작하고 언제 완료할 것인가? 단계별 일정은?

### 8.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 표준화 구현의 3대 축(기술·프로세스·사람)을 균형 있게 설계할 수 있다
- 조직의 데이터 표준화 성숙도를 진단하고 목표를 설정할 수 있다
- 비즈니스 가치 사례(Business Case)를 설득력 있게 작성할 수 있다
- Kotter (2012) 의 8단계 변화 관리 모델을 표준화에 적용할 수 있다
- 저항을 예측하고 효과적으로 관리할 수 있다

## 8.2 핵심 개념

### 8.2.1 표준화 구현의 3대 축

성공적인 데이터 표준화 구현은 **기술(Technology), 프로세스(Process), 사람(People)** 세 가지 축이 균형 있게 작동해야 한다. DAMA International (2017) 은 이 세 축을 '데이터 관리의 삼각대(Tripod)'라고 표현하며, 어느 하나라도 무너지면 전체가 무너진다고 경고한다.

| 축 | 중심 요소 | 전형적 실패 원인 |
|---|---|---|
| 기술 (Technology) | 도구, 시스템, 자동화, 인프라 | 도구 의존, 기술 인프라 부족, 통합 부재 |
| 프로세스 (Process) | 정책, 절차, 워크플로우, 표준 문서 | 프로세스 미정립, 모호한 절차, 강제력 부재 |
| 사람(People) | 역할, 책임, 문화, 역량, 의사소통 | 교육 부족, 저항 미관리, 리더십 부재 |

> **i 도구만으로는 충분하지 않다**
>
> 많은 조직이 비싼 데이터 거버넌스 도구를 구매하면 표준화가 자동으로 해결될 것이라 기대한다. Ladley (2019) 는 이를 '비싼 장난감 함정(Expensive Toy Trap)'이라 부른다. 도구는 이미 정의된 프로세스를 자동화할 때만 가치가 있다. 정의되지 않은 프로세스를 도구로 강제하려는 시도는 거의 항상 실패한다.

### 8.2.2 데이터 표준화 성숙도 모델

조직의 표준화 수준을 객관적으로 평가하기 위한 가장 널리 사용되는 도구는 **5단계 성숙도 모델(Maturity Model)**이다. 이는 카네기멜런대학교 SEI의 CMM(Capability Maturity Model)을 데이터 영역에 적용한 것이다 (DAMA International 2017; Brackett 2012). 조직의 현 위치를 진단하고 목표를 설정하는 출발점이 된다.

| 단계 | 명칭 | 특징 | KPI |
|---|---|---|---|
| 1 | Initial (초기) | Ad-hoc, 표준 부재, 개인 의존 | 없음 |
| 2 | Repeatable (반복) | 일부 표준, 비공식 적용 | 정성 평가 |
| 3 | Defined (정의) | 공식 표준, 체계적 적용, 교육 시행 | 초기 KPI 도입 |
| 4 | Managed (관리) | 성과 측정, 정기 모니터링 | 정량 KPI 추적 |
| 5 | Optimizing (최적화) | 지속 개선, 자동화, 예측·최적화 | 혁신 KPI 도입 |

Gartner (2023c) 의 2023년 조사에 따르면 **대부분의 기업이 1~2단계에 머물러 있으며**, 3단계 이상에 도달한 기업은 약 25%에 불과하다. 4~5단계 기업은 5% 미만이지만, 이들이 데이터 분석 ROI에서 평균 3배 이상의 성과를 보인다.

### 8.2.3 표준화 접근 전략: Top-Down vs Bottom-Up

표준화 추진 방식은 크게 **하향식(Top-Down)과 상향식(Bottom-Up)**으로 구분된다. 두 접근의 장단점을 이해하고 조직 상황에 맞게 결합하는 것이 핵심이다 (Ladley 2019; Seiner 2014).

| 구분 | Top Down | Bottom Up |
|---|---|---|
| 주도 주체 | 경영진, CDO, 거버넌스 위원회 | 현업 부서, 데이터 실무자 |
| 출발점 | 전사 전략, 정책 수립 | 현장 문제 해결 |
| 장점 | 강한 추진력, 통합성, 예산 확보 용이 | 실용성, 수용성, 빠른 효과 |
| 단점 | 현장 괴리, 형식적 준수 위험 | 단편적, 통합 어려움 |
| 적합 조직 | 대기업, 규제 산업, 위계 문화 | 스타트업, 분산 조직, 자율 문화 |

> Seiner의 'Non-Invasive' 접근법
>
> Seiner (2014) 는 Non-Invasive Data Governance 에서 두 접근을 결합한 '비침습적 접근법'을 제안한다. 핵심 원칙은:
>
> 1. 기존 데이터 사용자가 이미 일상적으로 데이터를 책임지고 있다는 인식
> 2. 새 직책 신설보다 기존 역할에 책임 부여
> 3. 정책보다 사람의 자연스러운 흐름에 집중
> 4. 거버넌스 활동을 일상에 보이지 않게 통합

### 8.2.4 빅뱅 vs 점진적 도입

표준화 도입 일정은 크게 빅뱅(Big Bang)과 점진적(Incremental) 방식으로 나뉜다. 두 방식 모두 장단점이 있으며, 대부분의 성공 사례는 점진적 접근을 채택한다.

#### 8.2.4.1 빅뱅 방식의 위험

- 일시에 모든 시스템에 표준 적용 → 막대한 초기 투자
- 실패 시 전사적 영향 → 리스크 집중
- 이해관계자 저항 동시 발생 → 변화 관리 부담
- 초기 학습 효과 활용 불가 → 시행착오 비용 증가

#### 8.2.4.2 점진적 방식의 강점

- 작은 성공으로 추진력 확보
- 실패 시 영향 최소화, 학습 가능
- 이해관계자 신뢰 점진적 구축
- 초기 사례를 활용한 다른 영역 설득

## 8.3 방법론

### 8.3.1 단계별 구현 로드맵

데이터 표준화 구현은 시간 축에 따라 5개 단계로 진행된다. 각 단계는 명확한 산출물과 **게이트 기준(Gate Criteria)**을 가져야 다음 단계로 진행할 수 있다 (DAMA International 2017; Seiner 2014).

#### 8.3.1.1 Phase 1: 진단 및 준비 (0~3개월)

- 현행 데이터 자산 인벤토리 작성
- 표준화 성숙도 진단 (As-Is 평가)
- 이해관계자 매핑 및 분석
- 비즈니스 가치 사례(Business Case) 수립
- 경영진 후원자(Executive Sponsor) 확보

- 초기 예산 및 조직 확보

#### 8.3.1.2 Phase 2: 파일럿 (3~9개월)

- 핵심 도메인 1~2개 선정 (고객, 제품 등)
- 표준 정의서 초안 작성
- 도구 POC(Proof of Concept) 진행
- 초기 데이터 스튜어드 양성
- 측정 가능한 초기 KPI 정의
- 파일럿 성과 정리 및 공유

#### 8.3.1.3 Phase 3: 확장 (9~18개월)

- 5~10개 도메인으로 확장
- 거버넌스 위원회 정착
- 표준 준수 측정 체계 구축
- 교육 프로그램 정례화
- 도구 본격 도입 및 통합
- 우수 사례 공유 및 표준 개선

#### 8.3.1.4 Phase 4: 통합 (18~30개월)

- 전사 핵심 시스템에 표준 적용
- ETL/ELT, CI/CD 파이프라인 통합
- 실시간 품질 모니터링 가동
- 외부 시스템(파트너, 규제기관)과의 표준 정렬
- 데이터 카탈로그 전사 활용 정착

#### 8.3.1.5 Phase 5: 최적화 (30개월~)

- AI 기반 자동화 도입 (메타데이터, 품질 등)
- 지속적 개선 사이클 운영
- 외부 인증 획득 (ISO (2022b), ISO/IEC (2023a) 등)
- 산업 표준 기여 및 외부 사례 발표
- 차세대 기술(AI, 데이터 메시) 대응

### 8.3.2 비즈니스 가치 사례 수립

표준화 투자 승인을 위한 **비즈니스 가치 사례(Business Case)**는 경영진의 의사결정을 지원하는 가장 중요한 문서다. 잘 작성된 가치 사례는 다음 다섯 요소를 포함해야 한다 (Ladley 2019; Brackett 2012).

#### 8.3.2.1 비즈니스 가치 사례의 5대 구성 요소

1. **현황의 비용** — 표준화 부재로 인한 현재 손실 (정량화 필수)
2. **개선 시나리오** — 표준화 후 기대 효과 (재무·운영·전략 차원)
3. **투자 비용** — 도구·인력·교육·외부 컨설팅 (3~5년 전체)
4. **ROI 계산** — 3~5년 누적 효과와 손익분기점
5. **리스크** — 미투자 시 리스크, 투자 시 리스크

#### 8.3.2.2 정량화 기법

| 손실 유형 | 측정 방법 | 예시 |
|---|---|---|
| 재작업 비용 | 오류 데이터 정제 시간 × 인건비 | 월 200시간 × 50,000원 = 1천만원 |

| 손실 유형 | 측정 방법 | 예시 |
|---|---|---|
| 의사결정 오류 | 잘못된 결정으로 인한 매출/비용 영향 | 캠페인 ROI 30% 저하 |
| 규제 벌금 | 컴플라이언스 위반 과징금 | GDPR 위반 시 매출 4% |
| 기회 비용 | 분석 지연으로 인한 시장 기회 손실 | 신제품 출시 지연 3개월 |
| 고객 이탈 | 잘못된 정보로 인한 고객 만족도 저하 | 이탈률 +2%p |

**i** ROI 계산 예시

연 매출 1조원 기업의 데이터 표준화 ROI 시나리오 (Gartner 2021):

- 현재 손실: 매출의 0.5% = 50억원/년 (Gartner 평균치)
- 3년 누적 손실: 약 150억원
- 투자: 도구 20억 + 인력 30억 + 교육 10억 = 60억원
- 3년 후 손실 감소: 60% → 90억원 절감
- **순효과: 90억 - 60억 = 30억원, ROI 50%**

### 8.3.3 우선순위 매트릭스

표준화 대상 영역의 우선순위를 결정할 때 가장 유용한 도구는 **임팩트-노력 매트릭스(Impact-Effort Matrix)**이다. 이는 영향력과 구현 난이도의 2차원으로 항목을 분류한다 (Henderson et al. 2017).

| 구분 | 영향 高 | 영향 低 |
|---|---|---|
| 노력 低 | Quick Win — 즉시 추진 | Fill-in — 여유 시 처리 |
| 노력 高 | Major Project — 단계적 추진 | Reconsider — 재검토 |

#### 8.3.3.1 우선순위 결정 절차

1. 후보 항목 도출 (브레인스토밍, 인터뷰)
2. 각 항목의 영향(Impact) 평가 (5점 척도)
3. 각 항목의 노력(Effort) 평가 (5점 척도)
4. 매트릭스에 배치하고 우선순위 결정
5. 거버넌스 위원회 합의 도출

### 8.3.4 Kotter의 8단계 변화 관리

하버드 비즈니스 스쿨의 John Kotter가 제시한 Leading Change 의 8단계 변화 관리 모델은 데이터 표준화 같은 조직 변화에 가장 널리 적용되는 프레임워크이다 (Kotter 2012). Ladley (2019) 은 이 모델을 데이터 거버넌스 정착에 직접 적용할 것을 권고한다.

| 단계 | 활동 | 표준화 적용 예시 |
|---|---|---|
| 1 | 위기감 조성(Urgency) | 데이터 품질 문제의 비용 가시화 |
| 2 | 변화 추진 연합 구축(Coalition) | CDO + 사업부 임원 연합 |
| 3 | 비전·전략 수립(Vision) | 3년 표준화 비전 선언 |

| 단계 | 활동 | 표준화 적용 예시 |
|---|---|---|
| 4 | 비전 소통(Communicate) | 전사 타운홀, 뉴스레터, 워크숍 |
| 5 | 임파워먼트(Empower) | 스튜어드에게 권한과 자원 부여 |
| 6 | 단기 성과 창출(Quick Wins) | 6개월 내 측정 가능한 성과 |
| 7 | 변화 가속화(Accelerate) | 성공 사례 확산, 추가 도메인 적용 |
| 8 | 조직 문화 정착(Anchor) | 표준 준수도를 평가 지표에 반영 |

### 8.3.5 저항 관리

표준화는 필연적으로 **조직 내 저항**을 유발한다. 저항을 예측하고 효과적으로 관리하지 못하면 가장 잘 설계된 표준도 실패한다 (Kotter 2012).

| 저항 유형 | 전형적 표현 | 대응 전략 |
|---|---|---|
| 불확실성 | '우리 방식이 어떻게 될지 모르겠다' | 명확한 비전·로드맵 공유 |
| 손실 두려움 | '내 자리·역할이 줄어들 것 같다' | 새로운 역할·기회 제시 |
| 기존 투자 방어 | '기존 시스템에 이미 많이 투자했다' | 마이그레이션 경로 제시 |
| 불신 | '이전 프로젝트도 실패했다' | 작은 성공으로 신뢰 회복 |
| 역량 부족 | '우리는 이걸 할 줄 모른다' | 교육·외부 지원 제공 |

## 8.4 실무 적용

### 8.4.1 조직 설계

데이터 표준화를 추진할 조직 구조는 **중앙집중형, 연합형, 분산형** 세 가지 모델 중에서 조직 특성에 맞게 선택한다. Eryurek et al. (2021) 은 Data Governance: The Definitive Guide 에서 연합형(Federated) 모델이 대부분의 중대형 조직에 가장 적합하다고 평가한다.

| 모델 | 구조 | 장점 | 약점 |
|---|---|---|---|
| 중앙집중형(Centralized) | 전담 조직이 모든 표준화 활동 관할 | 강한 통제, 일관성 | 현장 괴리, 병목 |
| 연합형(Federated) | 중앙 + 각 도메인의 스튜어드 네트워크 | 균형, 확장성 | 조정 비용 |
| 분산형(Decentralized) | 각 사업부가 자율적으로 표준 관리 | 유연성, 자율성 | 통합 어려움 |

### 8.4.2 표준화 도구 선정 전략

도구 선정은 표준화 구현의 중요한 결정이지만, **도구 선정에 과도한 시간을 쏟지 말라**. 완벽한 도구를 찾는 동안 표준화는 진전되지 않는다. TDWI (2023a) 는 'Best Practices Report'에서 '80% 적합한 도구를 빠르게 도입하라'고 권고한다.

#### 8.4.2.1 도구 선정 5단계

1. **요구사항 정의** — 필수·우선 기능 목록 작성

2. **후보 도구 식별** — RFI/시장 조사 (3~5개 후보)
3. **POC 진행** — 실제 데이터로 2~4주 검증
4. **평가 매트릭스** — 기능·성능·비용·지원성
5. **최종 결정** — 거버넌스 위원회 승인

### 8.4.3 측정과 추적

구현 진행 상황을 지속적으로 측정하고 보고하는 것이 추진력 유지의 핵심이다.

| 계층 | 지표 유형 | 예시 | 보고 주기 |
|---|---|---|---|
| 전략 | 비즈니스 임팩트 KPI | 데이터 관련 매출, 의사결정 속도 | 분기 |
| 전술 | 표준화 진행 KPI | 표준 적용 도메인 수, 준수율 | 월간 |
| 운영 | 활동 KPI | 표준 정의 건수, 교육 이수자 수 | 주간 |

### 8.4.4 실무 팁

> 팁 1: '왜'를 먼저 답하라
>
> Sinek (2009) 의 Start with Why 가 강조하듯, '왜 표준화가 필요한가?'에 대한 명확한 답이 없으면 어떤 표준도 정착하지 못한다. 비즈니스 가치를 먼저 명확히 하고, 구체적 사례로 설득하라. 추상적인 '데이터 품질 향상'이 아니라 '마케팅 캠페인 ROI 30% 향상' 같은 구체적이고 측정 가능한 가치를 제시해야 한다.

> 팁 2: 작은 성공을 빠르게 만들어라
>
> Kotter (2012) 의 8단계 모델에서 **6단계 'Quick Wins'**가 가장 중요하다. 6개월 안에 측정 가능한 성과를 만들지 못하는 프로젝트는 정치적 지지를 잃는다. 파일럿 도메인을 신중하게 선정하라 — 비즈니스 임팩트가 명확하고, 데이터가 비교적 깨끗하며, 협조적인 이해관계자가 있는 영역이 최적의 후보다.

> 팁 3: 인센티브를 설계하라
>
> 사람들은 평가 받는 것을 한다. 표준 준수가 평가 지표나 인센티브와 연결되지 않으면 형식적 준수에 그친다. Drucker (1954) 의 '측정되지 않으면 관리되지 않는다' 격언은 여기에도 적용된다. 표준 준수율을 부서 평가 지표에 포함시키고, 우수 스튜어드에 대한 인정·보상 체계를 구축하라.

> 팁 4: 표준은 살아있는 문서다
>
> 한 번 만든 표준은 영원하지 않다. 비즈니스 변화, 기술 발전, 새로운 규제에 따라 표준도 진화해야 한다. 정기적 검토(연 1~2회), 변경 요청 처리 절차, 버전 관리, 영향도 분석 등의 운영 체계를 구축해야 한다. 표준이 '돌에 새긴 계명'이 아니라 '살아있는 합의'임을 조직 전체가 인식해야 한다.

⚠ 주의: 컨설팅 의존의 함정

외부 컨설팅으로 표준을 만들고 떠나면 조직에 노하우가 남지 않는다. 컨설팅은 (1) 코칭과 지식 이전, (2) 외부 전문성 보완, (3) 일시적 자원 확보의 세 가지 목적으로만 활용해야 하며, 핵심 의사결정과 정착 활동은 반드시 내부 인력이 주도해야 한다. '표준화는 컨설팅사가 한다'는 인식이 형성되면 컨설팅 종료 후 조직 차원의 추진력이 소실된다.

⚠ 주의: 표준의 형식화

표준 문서가 만들어졌지만 아무도 따르지 않는 상태를 '형식화(Formalization)'라 한다. 이는 가장 흔하고 위험한 실패 패턴이다. 형식화를 방지하려면 (1) 표준 위반에 대한 결과(consequence)가 명확해야 하고, (2) 표준 준수가 일상 업무 흐름에 자연스럽게 통합되어야 하며, (3) 표준 적용에 따른 인센티브가 제공되어야 한다.

## 8.5 사례 연구

### 8.5.1 사례 1: 글로벌 보험사 P사의 5년 표준화 여정

#### 8.5.1.1 배경

글로벌 보험사 P사는 인수합병으로 18개국에서 사업을 운영하게 되었다. 각국 자회사가 독립적으로 운영되어 동일한 '계약', '청구', '고객' 개념이 국가별로 다르게 정의되었고, 그룹 차원의 위험 평가와 IFRS 17(보험 회계 신규 표준) 준수에 심각한 어려움을 겪고 있었다 (Deloitte 2023).

#### 8.5.1.2 도전 과제

- 18개국 자회사의 독자적 시스템과 데이터 정의
- IFRS 17 준수 마감 (3년 내)
- 국가별 규제 환경 차이
- 기존 시스템 운영 중단 불가
- 문화적 차이로 인한 변화 관리 부담

#### 8.5.1.3 접근 방법

1. Year 1: CDO 임명, 비즈니스 가치 사례 수립, 본사·핵심 3개국 파일럿
2. Year 2: 표준 데이터 모델 수립, 거버넌스 위원회 가동, 추가 5개국 확장
3. Year 3: 데이터 카탈로그 구축, IFRS 17 준비 완료, 전체 18개국 확장
4. Year 4: ETL 파이프라인 표준 적용, 실시간 품질 모니터링
5. Year 5: AI 기반 자동화 도입, ISO 8000 인증 획득

#### 8.5.1.4 성과 (5년 후)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 그룹 보고서 작성 기간 | 12주 | 2주 |
| IFRS 17 준수율 | 준비 안 됨 | 100% 준수 |
| 국가 간 데이터 호환성 | 약 25% | 95% |
| 데이터 관련 운영 비용 | 기준 100 | 55 |
| 성숙도 단계 | Level 1 | Level 4 |

#### 8.5.1.5 교훈

글로벌 표준화는 단거리 경주가 아닌 마라톤이다. 5년이라는 긴 시간 동안 추진력을 유지한 것은 (1) CDO의 일관된 리더십, (2) 단계별 성과 가시화, (3) 각국 자회사의 자율성 존중과 표준 강제의 균형, (4) IFRS 17이라는 외부 마감의 활용 덕분이었다. 특히 빅뱅 방식을 피하고 점진적으로 확장한 것이 핵심 성공 요인이었다.

### 8.5.2 사례 2: 중견 제조업체 D사의 실패와 재도전

#### 8.5.2.1 배경

국내 중견 제조업체 D사는 2018년 첫 번째 데이터 표준화 프로젝트를 시작했으나 1년 만에 좌초되었다. 그 후 원인 분석을 거쳐 2020년 재도전하여 성공한 사례다 (한국데이터산업진흥원 2022).

#### 8.5.2.2 첫 번째 시도 (2018)의 실패 원인

- IT 부서 단독 추진 — 비즈니스 부서 참여 부재
- 외부 컨설팅 의존 — 내부 역량 미축적
- 빅뱅 방식 채택 — 6개월 내 전사 적용 시도
- 측정 지표 부재 — 진행 상황 가시화 실패
- 경영진 후원 부족 — 중간에 우선순위 하락

#### 8.5.2.3 재도전 (2020)의 접근

1. CEO 직속 데이터 거버넌스 위원회 신설
2. 재무 부서 단일 도메인 파일럿 (6개월)
3. 내부 데이터 스튜어드 양성 (외부 코칭 활용)
4. 월간 임원 보고로 가시성 확보
5. 3년에 걸친 단계적 확장 계획

#### 8.5.2.4 교훈

데이터 표준화 실패는 흔하지만 회복할 수 있다. 실패 경험은 두 번째 시도의 강력한 자산이 된다. 단, 같은 실수를 반복하지 않으려면 첫 번째 실패의 원인을 정직하게 분석하고 공유하는 문화가 필요하다. D사의 경우 'lessons learned' 워크숍을 통해 조직 전체가 실패의 교훈을 학습한 것이 재도전의 출발점이었다.

### 8.5.3 사례 3: 공공기관의 거버넌스 정착

#### 8.5.3.1 배경

한국의 한 광역자치단체는 데이터 기반 행정 추진을 위해 2021년 표준화 프로젝트를 시작했다. 행정안전부의 공공데이터 표준화 가이드를 기반으로 자체 거버넌스 체계를 구축했다 (행정안전부 2023b).

#### 8.5.3.2 주요 추진 내용

1. 부단체장 직속 '데이터 정책관' 신설
2. 부서별 데이터 스튜어드 100여 명 지정 (겸직)
3. 핵심 행정 데이터 50종 표준화 시범
4. 표준 준수율을 부서 평가에 5% 반영
5. 우수 부서 표창 및 인센티브 부여

#### 8.5.3.3 성과 (2년 후)

- 표준 적용 데이터셋 50종 → 250종
- 부서 간 데이터 공유 요청 처리 시간 80% 단축
- 정책 의사결정용 데이터 분석 빈도 3배 증가

- 주민 대상 공공데이터 만족도 35% 향상

#### 8.5.3.4 교훈

공공 부문은 인사 평가의 영향력이 크다는 특성을 활용할 수 있다. 표준 준수를 부서 평가의 명시적 지표로 설정한 것이 빠른 정착의 핵심이었다. 또한 표창과 인센티브를 통한 '긍정적 강화'가 처벌 위주의 접근보다 효과적이었다. 민간 기업도 KPI·인사 평가와의 연계를 적극 고려할 가치가 있다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **변화 관리로서의 표준화** — 데이터 표준화 구현은 기술 프로젝트가 아닌 조직 변화 관리 활동이다. 실패의 80%는 사람과 프로세스 문제에서 기인하며, '어떤 표준'보다 '어떻게 정착'시킬 것인가가 중요하다.
2. **기술·프로세스·사람의 3대 축** — 성공적 표준화는 세 축이 균형 있게 작동해야 한다. 도구만으로는 충분하지 않으며, 어느 하나가 부족해도 전체가 무너진다.
3. **성숙도 모델 활용** — 5단계 성숙도 모델로 조직의 현 위치를 진단하고 목표를 설정한다. 대부분 조직은 1~2단계이며, 3단계 이상이 표준화의 실질적 가치 창출이 시작되는 지점이다.
4. **점진적 접근의 우월성** — 빅뱅보다 점진적 접근이 거의 항상 우월하다. 작은 성공으로 추진력을 확보하고, 학습 효과를 활용하며, 리스크를 최소화한다.
5. **비즈니스 가치 사례의 중요성** — 정량화된 비즈니스 가치 사례 없이는 어떤 표준화도 시작되지 못한다. 현재 손실, 개선 효과, 투자 비용, ROI, 리스크의 5요소를 명확히 제시해야 한다.
6. **변화 관리 모델의 적용** — Kotter (2012) 의 8단계 변화 관리 모델은 데이터 표준화에 직접 적용 가능하다. 특히 위기감 조성, Quick Wins 창출, 조직 문화 정착의 단계에 집중해야 한다.
7. **측정과 인센티브의 결합** — 측정 없이는 관리되지 않으며, 인센티브 없이는 행동이 변하지 않는다. 표준 준수를 평가 지표와 연결하는 것이 정착의 핵심 메커니즘이다.

## 다음 학습 가이드

### 8.5.4 필수 참고 문헌

- Ladley (2019) — Data Governance: How to Design, Deploy, and Sustain
- Kotter (2012) — Leading Change (변화 관리의 고전)
- Seiner (2014) — Non-Invasive Data Governance
- Eryurek et al. (2021) — Data Governance: The Definitive Guide
- Sinek (2009) — Start with Why

### 8.5.5 관련 본서 참조 장

- 제 9장 데이터 거버넌스 체계 — 거버넌스 조직 상세
- 제 10장 표준화 도구와 기술 — 도구 선정 심화
- 제 11장 실제 적용 사례 — 추가 사례 연구

## 연습 문제

1. 여러분 조직의 데이터 표준화 성숙도를 5단계 모델로 진단하고, 1년 후·3년 후 목표 단계를 설정하시오.
2. 여러분 조직에 표준화 투자를 제안하는 비즈니스 가치 사례의 핵심 요소(현황 손실, 개선 효과, ROI)를 정량적으로 작성하시오.
3. Kotter의 8단계 중 여러분 조직에서 가장 취약할 것으로 예상되는 단계는 무엇이며, 어떻게 보완할 것인가?

CHAPTER 9

# 데이터 거버넌스 체계

## 9.1 개요

**데이터 거버넌스(Data Governance)**는 조직이 데이터를 **전략적 자산으로 관리하기 위한 정책·프로세스·조직 체계의 총체**이다. Ladley (2019) 는 데이터 거버넌스를 '데이터에 관한 의사결정 권한과 책임을 정의하는 활동'으로 규정하며, 이를 데이터 관리의 '헌법'에 비유한다. 헌법이 없이 법률만 존재하는 국가가 혼란에 빠지듯이, 거버넌스 없는 데이터 표준화는 오래 지속될 수 없다.

DAMA International (2017) 은 데이터 거버넌스를 데이터 관리의 **11개 지식 영역 중 중앙에 위치한 영역**으로 도식화한다. 데이터 아키텍처, 모델링, 저장, 보안, 통합, 품질, 메타데이터, 마스터 데이터, 문서 관리 등 다른 모든 영역이 거버넌스의 정책·표준·의사결정에 의해 지배된다는 의미이다.

그러나 거버넌스는 종종 오해된다. 많은 조직에서 데이터 거버넌스는 '통제·규제·관료제'와 동일시되어 저항을 받는다. Eryurek et al. (2021) 는 Data Governance: The Definitive Guide 에서 '진정한 거버넌스는 통제가 아니라 지원(enablement)이다'라고 강조한다. 좋은 거버넌스는 데이터 활용을 어렵게 하는 것이 아니라 쉽게 만든다.

### 9.1.1 왜 지금 거버넌스가 중요한가

2020년대 들어 데이터 거버넌스는 그 어느 때보다 중요해졌다. 다음 다섯 가지 요인이 이를 견인한다 (Eryurek et al. 2021).

1. **규제 강화** — GDPR, 개인정보보호법, BCBS 239 (Basel Committee on Banking Supervision 2013) 등 데이터 관련 규제의 급증
2. **AI/ML 확산** — 학습 데이터의 품질·편향·출처가 새로운 거버넌스 대상 (Sambasivan et al. 2021)
3. **클라우드·분산 데이터** — 여러 클라우드·리전에 분산된 데이터의 통합 관리
4. **데이터 민주화** — 전사 데이터 활용 확대에 따른 통제 필요성
5. **데이터 유출 리스크** — 사이버 보안 사고와 규제 벌금의 증가

### 9.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 데이터 거버넌스의 4대 활동 영역과 그 상호관계를 설명할 수 있다
- 3계층 거버넌스 조직 구조를 설계할 수 있다
- 데이터 오너, 스튜어드, 카스토디안의 역할을 구분하고 지정할 수 있다
- 데이터 정책을 체계적으로 작성할 수 있다
- Seiner (2014) 의 Non-Invasive 접근법 등 다양한 거버넌스 스타일을 조직에 맞게 선택할 수 있다

## 9.2 핵심 개념

### 9.2.1 데이터 거버넌스의 4대 활동 영역

DAMA International (2017) 은 데이터 거버넌스의 활동 영역을 다음 네 가지로 정의한다. 이 네 영역은 상호 연결되어 있으며, 어느 하나만 강조되면 거버넌스는 불균형해진다.

| 활동 영역 | 설명 | 주요 산출물 |
|---|---|---|
| 정책 및 표준 | 데이터에 관한 조직 정책·표준 수립 | 데이터 정책서, 표준 문서, 가이드라인 |
| 데이터 자산 관리 | 데이터의 식별·분류·가치 평가·라이프사이클 | 데이터 카탈로그, 자산 분류 체계 |
| 조직 및 R&R | 역할·책임·의사결정 권한 정의 | RACI 매트릭스, 역할 정의서 |
| 컴플라이언스 및 감사 | 규제 준수·정기 감사·리스크 관리 | 감사 보고서, 컴플라이언스 대시보드 |

### 9.2.2 거버넌스 조직의 3계층 구조

Ladley (2019) 의 권고에 따른 표준적인 거버넌스 조직 구조는 **전략·전술·운영 3계층**으로 구성된다. 각 계층은 서로 다른 의사결정 수준과 활동 주기를 가지며, 정보와 이슈가 계층 간에 원활히 흐르도록 설계되어야 한다.

| 계층 | 명칭 | 구성 | 주요 활동 | 주기 |
|---|---|---|---|---|
| 전략 (Strategic) | 데이터 위원회 (Data Council) | CDO, CIO, CFO 등 C-level | 전략, 정책, 예산 승인 | 분기 1회 |
| 전술 (Tactical) | 데이터 거버넌스 위원회 | 데이터 오너, IT 리더 | 표준 승인, 갈등 조정 | 월 1~2회 |
| 운영(Operational) | 데이터 스튜어드십 커뮤니티 | 실무 스튜어드들 | 일상 운영, 품질 관리 | 주간·상시 |

> **i** 3계층 간 정보 흐름
>
> - **상향(Bottom-Up)**: 이슈·성과·개선 제안이 운영 → 전술 → 전략으로 흐른다
> - **하향(Top-Down)**: 전략·정책·의사결정이 전략 → 전술 → 운영으로 흐른다
> - 이 양방향 흐름이 원활하지 않으면 상층은 현장을 모르고, 하층은 방향을 잃는다

### 9.2.3 핵심 역할과 책임

데이터 거버넌스에는 여러 역할이 정의되어 있으며, 각 역할의 명확한 구분이 거버넌스 성공의 핵심이다. DAMA International (2017) 과 Ladley (2019) 의 권고를 종합한 주요 역할은 다음과 같다.

| 역할 | 책임 | 권한 수준 |
|---|---|---|
| CDO (Chief Data Officer) | 전사 데이터 전략 총괄, 최고 의사결정 | 임원 |
| 데이터 오너 (Data Owner) | 특정 도메인 데이터의 비즈니스 결정 | 부서장/임원 |
| 데이터 스튜어드 (Data Steward) | 일상적 품질 관리·표준 적용 | 실무 책임자 |
| 데이터 카스토디안 (Data Custodian) | 기술적 관리(저장·보안·백업) | IT 실무자 |
| 데이터 사용자 (Data User) | 데이터 활용·이슈 피드백 | 전 임직원 |
| 데이터 아키텍트 | 데이터 아키텍처 설계·표준 정의 | 전문가 |

> ! 오너 vs 스튜어드 vs 카스토디안
>
> 이 세 역할은 자주 혼동되지만 명확히 구분되어야 한다:
>
> - **데이터 오너**: '무엇을' 결정 (비즈니스 의미)
> - **데이터 스튜어드**: '어떻게' 관리 (일상 운영)
> - **데이터 카스토디안**: '어디에' 저장 (기술 운영)

### 9.2.4 CDO(Chief Data Officer)의 부상

CDO(최고 데이터 책임자)는 2010년대 이후 급격히 확산된 임원 역할이다. NewVantage Partners (2023) 의 조사에 따르면, **포춘 1000대 기업의 82%가 CDO 또는 CDAO(Chief Data & Analytics Officer)를 임명**했다. 이는 2012년의 12%에서 크게 증가한 수치이다.

#### 9.2.4.1 CDO의 주요 책임

- 전사 데이터 전략 수립 및 실행
- 데이터 거버넌스 체계 구축 및 운영
- 데이터 관련 규제 준수 총괄
- 데이터 기반 의사결정 문화 조성
- 데이터 관련 예산·인력 관리
- 이사회 및 경영진 대상 데이터 관련 보고

### 9.2.5 데이터 자산 분류

거버넌스의 기본 활동 중 하나는 **데이터 자산 분류(Data Classification)**이다. 모든 데이터를 동일하게 관리할 수는 없기에, 민감도·중요도·규제 요건에 따라 분류하고 차등적으로 관리해야 한다 (ISO/IEC 2022).

| 분류 등급 | 설명 | 관리 요구사항 |
|---|---|---|
| 공개(Public) | 외부 공개 가능한 정보 | 기본적 무결성 관리 |
| 내부(Internal) | 조직 내부용 정보 | 접근 통제, 인증 |

| 분류 등급 | 설명 | 관리 요구사항 |
|---|---|---|
| 기밀(Confidential) | 특정 담당자만 접근 가능 | 암호화, 감사 로그, 세밀한 권한 |
| 극비(Restricted) | 최고 수준 보호 대상 (개인정보, 영업비밀) | 종단 암호화, 격리, 상시 모니터링 |

## 9.3 방법론

### 9.3.1 Seiner의 Non-Invasive 접근법

Seiner (2014) 가 제안한 Non-Invasive Data Governance 접근법은 기존 조직 구조와 업무 흐름을 최대한 활용하여 거버넌스 부담을 최소화하는 방법론이다. 전통적 거버넌스가 종종 '새로운 조직·새로운 절차·새로운 부담'으로 인식되는 반면, Non-Invasive 접근은 '이미 존재하는 것을 활용하고 공식화'하는 데 집중한다.

#### 9.3.1.1 Non-Invasive의 4가지 원칙

1. 기존 데이터 사용자가 이미 일상적으로 데이터를 책임진다는 인식
2. 새로운 직책 신설보다 기존 역할에 책임 부여
3. 정책보다 사람의 자연스러운 흐름에 집중
4. 거버넌스 활동은 보이지 않게 통합

> Non-Invasive의 강점
>
> 이 접근법은 특히 중소·중견 기업, 자율적 문화의 조직, 그리고 이전 거버넌스 시도가 실패한 조직에서 효과적이다. 조직에 '새로운 무게'를 부과하지 않고 이미 하고 있는 일을 인정·공식화·강화하기 때문이다. 다만 강력한 규제 준수 요구가 있는 산업(금융·의료)에서는 이 접근만으로는 부족할 수 있다.

### 9.3.2 거버넌스 구축 절차

Ladley (2019) 의 권고에 따른 표준적 거버넌스 구축 절차는 7단계로 구성된다.

#### 9.3.2.1 Phase 1: 현황 진단 (0~2개월)

- 거버넌스 성숙도 평가 (5단계 모델 기준)
- 기존 데이터 정책·표준·조직 검토
- 이해관계자 인터뷰 및 문화 진단
- 갭 분석 및 우선순위 도출

#### 9.3.2.2 Phase 2: 목표 설정 (2~3개월)

- 거버넌스 비전 선언 (Vision Statement)
- 3~5년 로드맵 수립
- 성공 지표(Success Metrics) 정의
- 경영진 승인 및 자원 확보

#### 9.3.2.3 Phase 3: 조직 구성 (3~4개월)

- 데이터 거버넌스 위원회 구성
- 데이터 오너·스튜어드 지정
- R&R 문서화 및 공식 임명
- 조직도 및 커뮤니케이션 채널 설계

#### 9.3.2.4 Phase 4: 정책 수립 (4~6개월)

- 데이터 정책 프레임워크 설계
- 핵심 정책 작성 (품질, 보안, 접근, 유지)
- 가이드라인·절차서 개발
- 승인 워크플로우 정의

#### 9.3.2.5 Phase 5: 도구 배포 (6~9개월)

- 메타데이터 카탈로그 도입
- 거버넌스 워크플로우 도구 구축
- 품질 모니터링 시스템
- 정책 게시 및 검색 포털

#### 9.3.2.6 Phase 6: 운영 정착 (9~18개월)

- 정기 위원회 회의 정례화
- 월간·분기 거버넌스 보고서
- 교육 프로그램 상시 운영
- 이슈 트래킹 및 해결

#### 9.3.2.7 Phase 7: 지속 개선 (18개월~)

- KPI 모니터링 및 개선 계획
- 정기 검토 및 정책 개정
- 우수 사례 공유 및 확산
- 외부 인증 준비 (ISO (2022b), ISO/IEC (2023a) 등)

### 9.3.3 데이터 정책 작성

데이터 정책(Data Policy)은 조직의 데이터 관리 원칙을 공식화한 문서이다. 좋은 정책은 다음 요소를 반드시 포함해야 한다 (Ladley 2019; Eryurek et al. 2021).

#### 9.3.3.1 데이터 정책 표준 구성 요소

1. **적용 범위** — 조직, 데이터 유형, 시스템
2. **정책 목적과 원칙** — '왜' 이 정책이 존재하는가
3. **정의** — 사용된 주요 용어의 정의
4. **역할과 책임** — 정책과 관련된 R&R
5. **정책 내용** — 구체적 규칙과 요구사항
6. **절차와 워크플로우** — 정책 이행 방법
7. **위반 시 조치** — 준수 실패의 결과
8. **예외 처리** — 예외 신청·승인 절차
9. **검토·개정 주기** — 정책 유지 관리
10. **승인 및 발효 정보** — 승인자, 발효일, 버전

#### 9.3.3.2 정책 계층 구조

| 계층 | 성격 | 예시 |
|---|---|---|
| Policy (정책) | 포괄적 원칙 선언 | '데이터 품질 정책' |
| Standard (표준) | 구체적 규정 | '모든 마스터 데이터는 6대 품질 차원으로 측정한다' |

| 계층 | 성격 | 예시 |
|---|---|---|
| Procedure (절차) | 실행 단계 | '품질 이슈 발생 시 티켓 등록 → 스튜어드 검토' |
| Guideline (가이드) | 권장 사항 | '명명 시 3단 구조를 권장한다' |

### 9.3.4 RACI 매트릭스 활용

RACI 매트릭스는 데이터 거버넌스에서 역할과 책임을 명확히 하는 강력한 도구이다 (Project Management Institute 2021). 각 활동에 대해 네 가지 관계를 정의한다.

| 코드 | 의미 | 설명 |
|---|---|---|
| R | Responsible (실행자) | 실제 작업을 수행하는 사람 (한 명 이상) |
| A | Accountable (책임자) | 최종 책임을 지는 사람 (반드시 한 명) |
| C | Consulted (자문) | 의사결정 전 의견을 구하는 사람 |
| I | Informed (통보) | 결정 후 통보받는 사람 |

#### 9.3.4.1 RACI 매트릭스 예시: 신규 데이터 표준 등록

| 활동 | CDO | 데이터 오너 | 스튜어드 | IT |
|---|---|---|---|---|
| 표준 초안 작성 | I | A | R | C |
| 표준 검토 | I | A | R | C |
| 표준 승인 | A | C | R | I |
| 시스템 반영 | I | C | C | R,A |
| 교육 실시 | I | A | R | C |

## 9.4 실무 적용

### 9.4.1 거버넌스 도구

Gartner (2024a) 의 2024년 기준 주요 데이터 거버넌스 도구는 다음과 같이 분류된다. 도구는 거버넌스 체계 구축을 돕는 지원 수단이며, 조직·프로세스 없이 도구만으로는 거버넌스가 성립하지 않음을 유의해야 한다.

| 영역 | 주요 도구 | 특징 |
|---|---|---|
| 종합 거버넌스 플랫폼 | Collibra, Informatica Axon, OvalEdge | 정책·용어·워크플로우 통합 |
| 데이터 카탈로그 | Alation, Atlan, DataHub, OpenMetadata | 메타데이터 중심 접근 |
| 정책·워크플로우 | ServiceNow, Jira (커스텀) | 요청·승인·감사 흐름 |
| 감사·모니터링 | Immuta, Privacera, BigID | 접근 제어·개인정보 발견 |
| 개인정보 관리 | OneTrust, TrustArc | GDPR·CCPA 준수 |

### 9.4.2 거버넌스 KPI

거버넌스의 성과는 측정 가능한 KPI로 관리되어야 한다. 무엇을 측정하느냐가 무엇이 개선되느냐를 결정한다 (Sebastian-Coleman 2018).

| 카테고리 | KPI | 예시 임계값 |
|---|---|---|
| 조직 | 데이터 스튜어드 지정률 | 핵심 도메인의 100% |
| 정책 | 표준 준수율 | ≥ 90% |
| 교육 | 임직원 교육 이수율 | ≥ 85% |
| 운영 | 거버넌스 위원회 정기 개최 | 월 1회 이상 |
| 품질 | 핵심 데이터 품질 점수 | ≥ 95% |
| 이슈 | 거버넌스 이슈 평균 해결 시간 | ≤ 5 영업일 |
| 규제 | 규제 감사 지적 건수 | 0건 |

### 9.4.3 실무 팁

**팁 1: 거버넌스는 규칙이 아니라 문화다**

강제적 정책만으로는 거버넌스가 정착하지 않는다. **데이터 가치에 대한 공유된 인식, 자발적 협력, 상호 존중**이 진정한 거버넌스의 기반이다. Eryurek et al. (2021) 은 '문화 없는 거버넌스는 종이 위의 헌법과 같다'고 표현한다. 정기 뉴스레터, 우수 사례 공유, 성공 스토리 홍보 등 문화 조성 활동이 정책 문서 못지않게 중요하다.

**팁 2: 작은 위원회부터 시작하라**

처음부터 거대한 거버넌스 위원회를 만들면 형식화된다. 참석자가 많을수록 실질적 의사결정보다 정보 공유에 그치기 쉽다. 핵심 도메인의 작은 그룹(5~7명) 위원회로 시작해 구체적 성과를 만들고 점진적으로 확장하는 것이 효과적이다. 특히 초기에는 '결정'하는 위원회여야 하며, '보고'하는 위원회여서는 안 된다.

**팁 3: 이사회 수준의 관심을 확보하라**

데이터가 조직 전략의 핵심이 된 시대에, 데이터 거버넌스는 이사회 수준의 관심을 받아야 한다. Basel Committee on Banking Supervision (2013) 는 이를 명시적으로 요구하며, 다른 산업에서도 유사한 압력이 증가하고 있다. CDO는 정기적으로(최소 분기 1회) 이사회에 데이터 관련 상황을 보고할 수 있어야 한다.

**팁 4: 이슈 해결 프로세스를 명확히 하라**

거버넌스는 완벽한 규칙을 만드는 것이 아니라, 예외와 갈등을 해결하는 절차를 가지는 것이다. '이 데이터를 어떻게 정의할 것인가'를 두고 부서 간 이견이 있을 때, 누가 어떻게 결정하는지의 절차가 명확해야 한다. 이러한 갈등 해결 능력이 거버넌스의 진정한 가치이다.

⚠ 주의: 명목상 거버넌스의 함정

회의는 있으나 실질적 의사결정이 없는 '명목상 거버넌스(Governance Theater)'는 가장 흔한 실패 패턴이다. 예방하려면 (1) 위원회 의사결정 권한을 문서로 명시, (2) 결정 사항은 반드시 공식 회의록으로 기록, (3) 결정에 대한 이행 상황을 다음 회의에서 확인, (4) 결정 없이 종료되는 회의를 최소화해야 한다. 이슈가 계속 이월된다면 의사결정 구조 자체를 재검토해야 한다.

⚠ 주의: IT 편향의 위험

데이터 거버넌스는 IT 부서만의 활동이 아니다. IT가 주도하는 거버넌스는 종종 '기술 관점'에 편향되어 비즈니스 요구와 괴리된다. 데이터 거버넌스 위원회는 반드시 비즈니스 부서 대표가 절반 이상을 차지해야 하며, 이상적으로는 위원장을 비즈니스 임원이 맡는 것이 좋다.

## 9.5 실무 적용

실무 팁

(실무 팁을 작성하세요)

## 9.6 사례 연구

### 9.6.1 사례 1: BCBS 239 준수를 위한 글로벌 은행의 거버넌스

#### 9.6.1.1 배경

글로벌 시스템적 중요 은행(G-SIB)들은 2013년 발표된 Basel Committee on Banking Supervision (2013) '효과적인 리스크 데이터 집계 및 리스크 보고 원칙'을 준수해야 한다. 이 원칙은 14개의 데이터 거버넌스 원칙을 정의하며, 이사회 수준의 데이터 거버넌스 책임을 명시적으로 요구한다.

#### 9.6.1.2 BCBS 239 핵심 요구사항

- 이사회는 리스크 데이터 집계 및 보고 능력에 궁극적 책임
- 리스크 데이터 관리에 대한 명확한 R&R 정의
- 데이터 품질 정책의 문서화
- 정기적 컴플라이언스 자기 평가
- 규제 기관에 대한 정기 보고

#### 9.6.1.3 한 글로벌 은행의 대응

1. 이사회 산하 '데이터 감독 위원회(Data Oversight Committee)' 신설
2. CDO를 이사회 정기 보고자로 지정
3. 전 세계 30개국에 지역 데이터 거버넌스 위원회 구축
4. 리스크 데이터 원장(Risk Data Register) 관리
5. 연 2회 이사회 대상 데이터 거버넌스 보고

#### 9.6.1.4 성과

- 규제 감사에서 우수 사례로 인정
- 데이터 관련 규제 벌금 5년간 0건
- 리스크 보고서 정확도 및 적시성 대폭 향상
- 데이터 거버넌스 문화의 조직 정착

#### 9.6.1.5 교훈

규제 준수는 데이터 거버넌스 정착의 강력한 동력이 된다. BCBS 239의 명시적이고 검증 가능한 이사회 책임 요구는 CDO 조직의 위상을 크게 높였다. 다른 산업에서도 유사한 이사회 수준 데이터 책임이 강화되는 추세이므로, 규제와 함께 발전하는 거버넌스 전략을 수립할 필요가 있다.

### 9.6.2 사례 2: 미국 대형 소매 기업의 거버넌스 여정

#### 9.6.2.1 배경

미국의 한 대형 소매 기업은 매장 5,000개, 온라인 채널, 물류 센터에서 매일 수억 건의 고객 데이터를 생성한다. 2015년까지 부서별로 독립적으로 데이터를 관리했으나, 고객 옴니채널 경험 제공을 위한 전사 데이터 거버넌스 필요성이 부각되었다 (MIT Sloan CISR 2021).

#### 9.6.2.2 도전 과제

- 부서별 사일로 문화
- 데이터 오너십 부재
- 고객 정보의 부서 간 불일치
- 규제(캘리포니아 CCPA 등) 준수 부담

#### 9.6.2.3 접근 방법

1. Non-Invasive 접근법 채택 — 기존 조직 활용
2. CDO 신설, 마케팅·영업·물류에서 파견받은 겸직 위원회 구성
3. 핵심 고객 마스터부터 시작 (파일럿)
4. 우수 데이터 스튜어드 시상 프로그램 도입
5. 연 1회 ‘Data Day’ 전사 이벤트로 문화 조성

#### 9.6.2.4 성과 (3년 후)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 고객 통합 뷰 정확도 | 62% | 94% |
| 부서 간 데이터 요청 처리 시간 | 2주 | 2일 |
| CCPA 응대 시간 | 수주 | 24시간 이내 |
| 직원 거버넌스 인식 (설문) | 37% | 82% |

#### 9.6.2.5 교훈

Non-Invasive 접근이 대기업 소매업에도 효과적일 수 있다. 새로운 조직을 만들기보다 기존 부서의 데이터 담당자를 공식화·연결하는 접근이 저항을 최소화했다. 특히 ‘문화’ 조성 활동(Data Day, 시상 프로그램)이 정책 문서보다 정착에 더 크게 기여했다.

### 9.6.3 사례 3: 한국 공공기관의 데이터 거버넌스

#### 9.6.3.1 배경

한국의 한 대형 공공기관은 2020년 정부의 데이터 3법 개정에 따라 체계적인 데이터 거버넌스 도입이 필요했다. 특히 개인정보 보호 강화와 공공데이터 개방 확대라는 상반된 요구를 동시에 충족해야 하는 과제가 있었다 (한국지능정보사회진흥원 2023).

#### 9.6.3.2 주요 추진 내용

1. 데이터 관리 규정 제정 및 이사회 승인

2. 이사장 직속 CDO 임명 (정보화 담당 임원 겸직)
3. 부서별 데이터 관리 책임관 지정 (총 45명)
4. 월간 데이터 거버넌스 협의체 정기 운영
5. 데이터 자산 분류 체계 4단계 도입 (공개~극비)
6. 연 2회 데이터 거버넌스 성과 이사회 보고

#### 9.6.3.3 성과 (2년 후)

- 전사 데이터 카탈로그 등록 자산 3,500건
- 공공데이터 개방 건수 2배 증가
- 개인정보 침해 사고 0건 유지
- 행정안전부 데이터 관리 평가 최우수 등급

#### 9.6.3.4 교훈

공공기관의 데이터 거버넌스는 규제 준수와 시민 서비스 향상이라는 두 가지 요구를 균형있게 충족해야 한다. 한국 공공기관의 사례는 명확한 R&R, 정기적 이사회 보고, 그리고 '개방'과 '보호'의 균형점 설정이 성공의 열쇠임을 보여준다. 특히 이사장 직속 CDO 임명이 거버넌스의 실질적 권한을 확보하는 데 결정적이었다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **거버넌스의 본질** — 데이터 거버넌스는 조직이 데이터를 자산으로 관리하기 위한 정책·프로세스·조직 체계의 총체이다. 데이터 관리의 '헌법'에 해당하며, 다른 모든 데이터 관리 활동의 기반이 된다.
2. **4대 활동 영역** — 정책·표준, 자산 관리, 조직·R&R, 컴플라이언스·감사가 거버넌스의 4대 영역이다. 어느 하나만 강조되면 거버넌스는 불균형해진다.
3. **3계층 조직 구조** — 전략(데이터 위원회) - 전술(거버넌스 위원회) - 운영(스튜어드십)의 3계층 구조가 표준적이다. 계층 간 원활한 정보 흐름이 성공의 열쇠이다.
4. **역할의 명확한 구분** — 데이터 오너('무엇을'), 스튜어드('어떻게'), 카스토디안('어디에')의 역할이 명확히 구분되어야 한다. 역할 혼동은 책임 회피와 이슈 방치로 이어진다.
5. **CDO의 부상** — 포춘 1000대 기업의 82%가 CDO 또는 CDAO를 임명했다 (NewVantage Partners 2023). CDO는 조직 데이터 전략의 실행 책임자이자 이사회 대상 데이터 대변인 역할을 수행한다.
6. **Non-Invasive 접근의 유용성** — 새 조직·새 절차보다 기존 조직·역할을 활용·공식화하는 Seiner (2014) 의 Non-Invasive 접근이 특히 문화적 저항이 있는 조직에서 효과적이다. 다만 강한 규제 환경에서는 보완이 필요하다.
7. **거버넌스는 문화다** — 정책과 도구만으로는 거버넌스가 정착하지 않는다. 데이터 가치에 대한 공유된 인식, 자발적 협력의 문화 조성이 정책 문서 못지않게 중요하다.

## 다음 학습 가이드

### 9.6.4 필수 참고 문헌


- Ladley (2019) — Data Governance: How to Design, Deploy, and Sustain
- Eryurek et al. (2021) — Data Governance: The Definitive Guide
- Seiner (2014) — Non-Invasive Data Governance
- Sebastian-Coleman (2018) — Navigating the Labyrinth

### 9.6.5 규제 및 프레임워크

- Basel Committee on Banking Supervision (2013) — 금융 산업 리스크 데이터 원칙
- ISO/IEC (2022) — 정보 보안 관리 시스템 표준

### 9.6.6 관련 본서 참조 장

- 제 5장 메타데이터 관리 — 거버넌스의 정보 기반
- 제 6장 데이터 품질 관리 — 품질 거버넌스
- 제 8장 표준화 구현 전략 — 변화 관리
- 제 12장 AI·ML과 데이터 표준화 — AI 거버넌스

## 연습 문제

1. 여러분 조직의 데이터 거버넌스 3계층 조직 구조를 설계하고, 각 계층의 구성원과 활동 주기를 정의하시오.
2. 여러분 조직의 핵심 데이터 도메인 하나를 선정하여, 데이터 오너·스튜어드·카스토디안을 지정하시오.
3. Non-Invasive 접근이 여러분 조직에 적합한가? 그 이유는 무엇인가? 대안이 있다면 무엇인가?

CHAPTER 10

# 표준화 도구와 기술

## 10.1 개요

데이터 표준화 도구 시장은 지난 10년간 근본적인 변화를 겪었다. 전통적으로 온프레미스 엔터프라이즈 소프트웨어가 시장을 지배했다면, 2020년대 들어 **클라우드 네이티브, 오픈소스, AI 기반 자동화** 도구들이 급성장하며 판도가 바뀌고 있다. Gartner (2024d) 는 메타데이터 관리 시장이 연 19% 성장한다고 보고하며, 이는 소프트웨어 전체 시장 평균의 2배 수준이다.

IDC (2024a) 의 2024년 데이터 관리 소프트웨어 전망에 따르면, **전 세계 데이터 관리 시장은 2028년까지 연평균 16.5% 성장**하여 900억 달러 규모에 이를 것으로 예상된다. 이 성장은 세 가지 큰 흐름에 의해 견인된다: (1) 클라우드 데이터 마이그레이션, (2) DataOps 및 자동화 확산, (3) AI/ML 인프라로서의 데이터 표준화 인식 확산이다.

### 10.1.1 도구 시장의 3대 흐름

1. **클라우드 우선(Cloud-First)** — AWS Glue, Azure Purview, GCP Dataplex 등 클라우드 통합 도구의 부상
2. **DataOps 통합** — 데이터 표준화가 CI/CD 파이프라인의 일부로 자동화 (Cataldo and Foote 2022)
3. **AI 기반 자동화** — Monte Carlo, Anomalo 등 이상 탐지·자동 태깅 도구의 등장

### 10.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 데이터 표준화 도구의 7대 카테고리를 이해하고 각각의 대표 도구를 안다
- 상용·오픈소스·클라우드 네이티브·AI 기반 도구의 강약점을 비교할 수 있다
- 도구 선정을 위한 평가 매트릭스를 설계할 수 있다
- POC(Proof of Concept) 절차를 수립하고 실행할 수 있다
- 모던 데이터 스택의 구성 요소와 통합 아키텍처를 이해한다

## 10.2 핵심 개념

### 10.2.1 데이터 표준화 도구의 7대 카테고리

데이터 표준화를 지원하는 도구는 기능 영역에 따라 다음 7개 카테고리로 분류된다. 각 카테고리는 서로 다른 문제를 해결하며, 성숙한 데이터 조직은 여러 카테고리의 도구를 결합해 활용한다 (Gartner 2023b).

| 카테고리 | 핵심 기능 | 대표 도구 |
|---|---|---|
| 데이터 모델링 | ER 모델·논리·물리 모델 설계 | erwin, ER/Studio, PowerDesigner, DA# |
| 메타데이터 관리 | 메타데이터 수집·저장·검색 | Apache Atlas, DataHub, OpenMetadata |
| 데이터 카탈로그 | 자산 검색·비즈니스 컨텍스트 | Alation, Atlan, Collibra Catalog |
| 데이터 품질 | 품질 규칙·검증·모니터링 | Informatica DQ, Great Expectations, Soda |
| 마스터 데이터 관리 | 골든 레코드·엔티티 해소 | Informatica MDM, Reltio, Profisee |
| 데이터 거버넌스 | 정책·워크플로우·감사 | Collibra, OvalEdge, Informatica Axon |
| DataOps | 파이프라인·CI/CD·오케스트레이션 | dbt, Airflow, Dagster, DataKitchen |

### 10.2.2 도구 분류: 4대 유형별 특성

도구를 배포 모델과 상업 모델에 따라 분류하면 다음 4대 유형으로 나뉜다. 각 유형은 서로 다른 강점과 약점을 가지므로, 조직의 성숙도·규제 환경·인력 상황에 맞게 선택해야 한다 (Gartner 2023b; Forrester Research 2023a).

| 분류 | 강점 | 약점 |
|---|---|---|
| 엔터프라이즈 상용 | 포괄적 기능, 검증된 안정성, 벤더 지원, 규제 준수 인증 | 높은 비용, 도입 복잡성, 벤더 락인 |
| 오픈소스 | 낮은 비용, 커스터마이징, 벤더 락인 없음 | 내부 인력 필요, 커뮤니티 의존, 지원 한계 |
| 클라우드 네이티브 | 확장성, 통합성, 자동 관리, 빠른 배포 | 벤더 락인, 데이터 주권 이슈, 비용 예측 어려움 |
| AI 기반 신규 | AI 자동화, 모던 UX, 빠른 혁신, 이상 탐지 | 성숙도 부족, 검증 부족, 기능 완전성 |

### 10.2.3 모던 데이터 스택

2020년대 초 등장한 '모던 데이터 스택(Modern Data Stack)'은 여러 전문 도구를 조합해 엔드투엔드 데이터 파이프라인을 구축하는 접근이다. Cataldo and Foote (2022) 는 이를 '데이터 관리의 유닉스 철학'이라 부른다 — 각 도구가 하나의 일을 잘 하고, 이들이 표준 인터페이스로 연결된다는 의미이다.

| 레이어 | 역할 | 대표 도구 |
|---|---|---|
| 수집(Ingestion) | 원천 시스템에서 데이터 수집 | Fivetran, Airbyte, Stitch, Segment |
| 저장(Storage) | 데이터 웨어하우스·레이크 | Snowflake, Databricks, BigQuery, Redshift |
| 변환(Transformation) | 데이터 변환·모델링 | dbt, Dataform, Matillion |
| 표준화/품질 | 품질 검증·표준 적용 | Great Expectations, Soda, Monte Carlo |
| 분석(Analytics) | BI·시각화 | Looker, Tableau, Power BI, Mode |
| 거버넌스 | 카탈로그·계보·정책 | Atlan, Collibra, DataHub, Alation |

> 모던 데이터 스택의 철학
>
> 전통적인 '한 벤더의 모든 것' 접근과 달리, 모던 데이터 스택은 각 영역의 최고 도구를 SQL과 REST API 같은 표준 인터페이스로 연결한다. 이는 (1) 특정 벤더 종속을 피할 수 있고, (2) 각 영역에서 최고 성능의 도구를 사용할 수 있으며, (3) 새로운 도구로의 교체가 용이하다는 장점이 있다. 다만 여러 도구를 통합·운영하는 복잡성이 증가하므로 조직의 기술 성숙도가 뒷받침되어야 한다.

### 10.2.4 AI 기반 데이터 표준화 도구

2020년대 중반 등장한 AI 기반 도구들은 기존 도구가 수동으로 처리하던 작업을 자동화한다. 특히 데이터 관측성(Data Observability)과 이상 탐지 영역에서 큰 혁신이 일어나고 있다 (Bilgen 2024).

| 영역 | AI 활용 | 대표 도구 |
|---|---|---|
| 자동 태깅·분류 | ML 모델로 데이터 자산 자동 카테고리화 | Atlan, Castor, Select Star |
| 이상 탐지 | 통계·ML 기반 비정상 데이터 자동 발견 | Monte Carlo, Anomalo, Lightup |
| 자동 계보 추적 | 코드·로그 분석으로 데이터 흐름 추론 | DataHub, OpenLineage |
| 데이터 관측성 | 파이프라인 모니터링·SLA 관리 | Bigeye, Acceldata, Validio |
| 자연어 검색 | LLM 기반 자연어 데이터 검색 | Atlan AI, Select Star Copilot |

## 10.3 방법론

### 10.3.1 도구 선정 7대 기준

데이터 도구 선정 시 다음 7대 기준을 균형 있게 평가해야 한다. Gartner의 시장 가이드가 제시하는 표준적 평가 프레임워크이다 (Gartner 2024d, 2023b).

1. **기능 적합성** — 조직 요구사항과 도구 기능의 일치도
2. **확장성** — 데이터 양·사용자 수·시스템 증가에 대한 대응력
3. **통합성** — 기존 시스템·다른 도구와의 연동
4. **사용성** — 비즈니스 사용자의 접근성, 학습 곡선
5. **거버넌스** — 정책 관리·감사·이력 추적 기능
6. **TCO(Total Cost of Ownership)** — 라이선스·구축·운영·교육 총 비용

7. **벤더 신뢰성** — 재무 상태, 제품 로드맵, 커뮤니티, 지원 품질

### 10.3.2 평가 매트릭스 설계

객관적인 도구 비교를 위해 가중치 평가 매트릭스(Weighted Scoring Matrix)를 활용한다. 각 기준에 조직의 우선순위에 따라 가중치를 부여하고, 후보 도구별로 점수를 매긴다 (Gartner 2023b).

| 기준 | 가중치 | 도구 A | 도구 B | 도구 C |
|---|---|---|---|---|
| 기능 적합성 | 25% | 4.0 | 3.5 | 4.5 |
| 확장성 | 15% | 4.5 | 3.0 | 4.0 |
| 통합성 | 15% | 3.5 | 4.0 | 3.5 |
| 사용성 | 10% | 3.0 | 4.5 | 3.0 |
| 거버넌스 | 10% | 4.0 | 3.5 | 4.5 |
| TCO (역가중치) | 15% | 3.0 | 4.0 | 2.5 |
| 벤더 신뢰성 | 10% | 5.0 | 3.5 | 4.0 |
| 가중 총점 | 100% | 3.85 | 3.68 | 3.83 |

### 10.3.3 POC(Proof of Concept) 절차

실제 도구 선정 전에는 POC를 통해 실제 환경에서의 성능과 적합성을 검증해야 한다. TDWI (2023a)는 POC 실패의 주요 원인을 '실제 데이터가 아닌 샘플 데이터 사용, 시나리오 부족, 평가 기준 부재'로 지목한다. POC는 다음 6단계 절차를 따라야 한다.

1. POC 목적 및 성공 기준 명확화 (측정 가능한 지표)
2. 실제 조직 데이터로 POC 시나리오 설계 (3~5개)
3. 후보 도구 축소 (RFI 검토 후 2~3개 후보)
4. POC 환경 구축 및 각 도구별 2~4주 검증
5. 평가 매트릭스 완성 및 점수 산정
6. 이해관계자 검토·최종 결정·계약

> **i** POC 성공 기준 예시
>
> 좋은 POC 성공 기준의 특징: 측정 가능하고 시간 제한이 있으며 이해관계자가 합의한 것.
>
> - 대량 데이터 처리 성능: 1억 건 데이터의 프로파일링 30분 이내 완료
> - 사용성 지표: 비기술 사용자가 3시간 교육으로 기본 검색 수행 가능
> - 통합 검증: 기존 데이터 웨어하우스 3개와 자동 연동 성공
> - 규칙 정의: 표준 품질 규칙 50개를 24시간 내 정의·실행

### 10.3.4 통합 아키텍처 설계

여러 도구를 결합할 때 통합 아키텍처가 관건이다. Andreessen Horowitz (2023)는 '현대 데이터 인프라의 emerging architectures' 보고서에서 다음 3원칙을 제시한다.

#### 10.3.4.1 통합 아키텍처 3대 원칙

**1원칙: 메타데이터 우선(Metadata-First)**

데이터를 중앙 위치로 옮기려 하지 말고 메타데이터로 통합하라. 데이터 이동은 비용이지만, 메타데이터 통합은 유연성을 제공한다.

**2원칙: 오픈 표준 활용**

OpenLineage, OpenAPI, Apache Iceberg 등 오픈 표준을 활용하여 도구 간 상호운용성을 확보하라. 벤더 종속적 형식은 향후 마이그레이션의 큰 부담이 된다.

**3원칙: API 우선(API-First)**

모든 도구는 REST API 또는 SQL 인터페이스로 접근 가능해야 한다. GUI만 제공하는 도구는 자동화와 통합에 심각한 한계가 있다.

## 10.4 실무 적용

### 10.4.1 도구 도입 체크리스트

도구 도입 결정 시 기술적·비즈니스·조직적 관점에서 종합적으로 검토해야 한다 (Ladley 2019).

#### 10.4.1.1 기술적 체크리스트

- 데이터 소스 지원 (RDBMS, NoSQL, 클라우드, 빅데이터)
- 데이터 형식 지원 (관계형, JSON, XML, Parquet, Iceberg)
- 규칙 정의 방식 (GUI, SQL, Python, YAML)
- 자동화 지원 (CI/CD, API, 웹훅)
- 확장성 및 성능 (수평 확장, 대용량 처리)
- 보안 (인증, 권한, 암호화, 감사)
- 고가용성 및 재해 복구

#### 10.4.1.2 비즈니스 체크리스트

- 총소유비용(TCO) — 3~5년 관점
- 도입·교육·마이그레이션 리소스
- 벤더 재무 안정성 및 시장 지위
- 고객 사례 및 참조 가능 조직
- 계약 조건 (해지, SLA, 데이터 소유권)
- 규제 및 컴플라이언스 지원 (GDPR, HIPAA)

#### 10.4.1.3 조직적 체크리스트

- 내부 역량 (도구 운영에 필요한 스킬)
- 교육 및 지원 체계
- 변화 관리 계획
- 기존 시스템과의 병행 운영 전략

### 10.4.2 벤더 락인 대응 전략

벤더 락인은 도구 선정 시 항상 고려해야 하는 리스크다. 특히 클라우드 네이티브 도구는 특정 클라우드에 종속되는 경향이 강하다. 다음 5가지 전략으로 락인을 완화할 수 있다 (Andreessen Horowitz 2023).

1. 오픈 데이터 형식 사용 — Iceberg, Delta, Parquet 등
2. 표준 SQL 및 REST API 우선
3. 메타데이터·규칙의 이식 가능한 형식 (YAML, JSON) 저장
4. 주요 자산의 정기 백업·엑스포트
5. 계약에 데이터 반환 조항 명시

### 10.4.3 실무 팁

> 팁 1: 완벽한 도구는 없다
>
> 모든 요구사항을 만족하는 도구는 존재하지 않는다. TDWI (2023a) 조사에 따르면 **핵심 요구사항 80%를 충족하는 도구를 빠르게 도입하고 부족한 부분은 보완**하는 전략이 완벽한 도구를 6개월 이상 찾는 전략보다 훨씬 효과적이다. 도구 선정에 매몰되면 정작 표준화 자체는 진전되지 않는다.

> 팁 2: 사람·프로세스가 먼저, 도구는 나중
>
> Ladley (2019) 은 '도구만 도입하고 거버넌스 없는 조직은 비싼 장난감만 가진 셈'이라고 표현한다. 사람과 프로세스 정비 없이 도구만으로는 표준화가 정착하지 않는다. 데이터 스튜어드가 지정되지 않은 상태에서 카탈로그 도구를 도입하면 결국 아무도 사용하지 않는 값비싼 데이터베이스가 될 뿐이다.

> 팁 3: 오픈소스와 상용을 조합하라
>
> 핵심 기능은 상용 도구로 안정성을 확보하고, 부수적 기능이나 실험적 영역은 오픈소스로 유연성을 확보하는 하이브리드 전략이 대부분의 조직에 효과적이다. 예를 들어 거버넌스 플랫폼은 Collibra 같은 상용을 사용하되, 데이터 품질 검증은 Great Expectations 오픈소스를 활용할 수 있다.

> 팁 4: 도구는 정착까지 최소 18개월
>
> 고가 도구를 구매해도 6개월 만에 큰 성과를 기대하지 말라. 도구가 조직에 진정으로 정착하려면 최소 18개월이 필요하다. 사용자 교육, 자산 등록, 프로세스 통합, 문화 정착에 시간이 걸린다. 경영진에게 이 현실을 미리 알리고 장기적 관점의 지원을 확보해야 한다.

> ⚠ 주의: 도구 통합의 복잡성
>
> 여러 카테고리의 도구를 결합하면 통합 비용이 라이선스 비용을 초과할 수 있다. 특히 각 도구가 서로 다른 메타모델과 API를 사용하면 통합 코드 유지보수가 큰 부담이 된다. 단일 벤더의 통합 플랫폼 vs 최고 도구 조합(Best-of-Breed) 사이의 균형을 신중히 평가해야 한다. 일반적으로 중소 조직은 통합 플랫폼, 대기업은 Best-of-Breed가 유리하다.

> ⚠ 주의: 클라우드 표준 ≠ 벤더 표준
>
> AWS Glue Catalog 표준만 따르면 GCP·Azure로 이동할 때 큰 비용이 든다. AWS Data Catalog 형식은 사실상 AWS 전용 표준이며, 다른 클라우드와 호환되지 않는다. OpenLineage, Apache Iceberg, DCAT 같은 진정한 오픈 표준과 벤더별 도구의 균형을 고려하라. 다중 클라우드 전략을 갖고 있다면 특히 오픈 표준을 우선해야 한다.

## 10.5 사례 연구

### 10.5.1 사례 1: 글로벌 소매 W사의 모던 데이터 스택 구축

#### 10.5.1.1 배경

글로벌 소매 기업 W사는 15년간 온프레미스 데이터 웨어하우스와 상용 ETL 도구를 사용해왔다. 매장 5,000개, 온라인 채널, 물류 센터에서 매일 30TB의 데이터가 생성되면서 기존 시스템의 확장성 한계가 명확해졌다. 또한 데이터 팀이 신규 요청 대응에 6~8주가 걸려 비즈니스 요구를 만족시키지 못하는 상황이었다 (MIT Sloan CISR 2023).

#### 10.5.1.2 도전 과제

- 온프레미스 시스템의 확장성 한계
- 데이터 분석 요청 처리 지연 (6~8주)
- 벤더 락인으로 인한 라이선스 비용 급증
- 데이터 팀 30명 대비 500명 이상의 데이터 소비자

#### 10.5.1.3 접근 방법

1. Snowflake로 데이터 웨어하우스 클라우드 마이그레이션
2. Fivetran으로 100+ 원천 시스템 연결
3. dbt로 SQL 기반 데이터 변환 표준화
4. Great Expectations로 품질 규칙 코드화
5. Atlan으로 데이터 카탈로그 및 거버넌스
6. Monte Carlo로 데이터 관측성 (SLA 관리)

#### 10.5.1.4 성과 (18개월)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 신규 데이터 요청 처리 | 6~8주 | 3~5일 |
| 데이터 파이프라인 오류율 | 12% | 1.5% |
| 연간 데이터 인프라 비용 | 기준 100 | 62 |
| 자체 분석 가능 사용자 | 80명 | 450명 |

#### 10.5.1.5 교훈

모던 데이터 스택 도입은 단순한 도구 교체가 아니라 데이터 팀의 운영 방식 자체를 바꾸는 전환이다. W사의 성공 요인은 (1) SQL 기반 접근으로 학습 곡선 완화, (2) 오픈 표준 활용으로 벤더 락인 최소화, (3) 데이터 스튜어드 네트워크 구축을 통한 카탈로그 활용 문화 정착이었다. 특히 dbt의 SQL 접근이 기존 인력의 재교육을 최소화하는 핵심 역할을 했다.

### 10.5.2 사례 2: 한국 중견 제조업체의 오픈소스 도입

#### 10.5.2.1 배경

국내 중견 제조업체 M사는 매출 5,000억원 규모로, 상용 엔터프라이즈 데이터 관리 도구 라이선스 비용(연 8억원)이 큰 부담이었다. 2022년 오픈소스 기반의 데이터 표준화 체계 재구축을 결정했다 (한국데이터산업진흥원 2023b).

#### 10.5.2.2 접근 방법

1. Apache Atlas — 메타데이터 관리
2. DataHub — 데이터 카탈로그 (Atlas 후 이전)

3. Great Expectations — 데이터 품질 검증
4. Apache Airflow — 워크플로우 오케스트레이션
5. Grafana — 모니터링 대시보드

#### 10.5.2.3 성과 (2년)

- 연간 라이선스 비용 8억 → 1억 (85% 절감, 오픈소스 지원 계약 포함)
- 데이터 자산 등록 수 3,200건
- 품질 규칙 자동 검증 1,500건
- 내부 데이터 엔지니어 6명 양성

#### 10.5.2.4 교훈

중견 기업에게 오픈소스는 매력적 대안이지만, 내부 엔지니어링 역량이 반드시 필요하다. M사는 초기 6개월간 외부 컨설팅을 활용해 도입한 후, 내부 인력이 운영을 인수하는 '컨설팅 → 지식 이전 → 자립' 3단계 전략을 사용했다. 상용 대비 UI가 세련되지 않고 일부 고급 기능이 부족하지만, 핵심 기능은 충분히 충족되며 커스터마이징의 유연성이 오히려 장점이 되었다.

### 10.5.3 사례 3: 국내 금융사 K사의 하이브리드 접근

#### 10.5.3.1 배경

국내 대형 은행 K사는 규제 준수(Basel Committee on Banking Supervision (2013))와 혁신 사이의 균형을 필요로 했다. 규제 대응은 검증된 상용 도구로, 혁신적 분석은 오픈소스 및 클라우드 도구로 접근하는 하이브리드 전략을 채택했다 (금융보안원 2023).

#### 10.5.3.2 도구 스택

| 영역 | 규제 대응 상용 | 혁신 분석 오픈 클라우드 |
|---|---|---|
| 데이터 웨어하우스 | Oracle Exadata (온프레) | AWS Redshift, Snowflake |
| 데이터 카탈로그 | Collibra (규제 자산) | DataHub (분석 자산) |
| 데이터 품질 | Informatica DQ | Great Expectations |
| 거버넌스 | Collibra | 내부 개발 대시보드 |
| MDM | Informatica MDM | - |
| 오케스트레이션 | IBM Control-M | Apache Airflow |

#### 10.5.3.3 교훈

규제 산업에서는 '올바른 도구'의 선택이 규제 관점과 혁신 관점에서 다를 수 있다. K사의 하이브리드 접근은 각 영역의 요구사항에 맞는 도구를 사용하되, 메타데이터 통합 레이어(자체 개발)로 두 세계를 연결했다. 이는 벤더 락인 리스크와 벤더 지원 신뢰성 사이의 균형점을 찾은 실용적 접근이다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **도구 시장의 급격한 변화** — 데이터 표준화 도구 시장은 클라우드 네이티브, 오픈소스, AI 기반 자동화로 재편되고 있다. 전통적 엔터프라이즈 도구만으로는 현대의 요구를 충족할 수 없다.
2. **7대 카테고리의 이해** — 데이터 모델링·메타데이터·카탈로그·품질·MDM·거버넌스·DataOps 7대 카테고리는 서로 다른 문제를 해결하며, 성숙한 조직은 이들을 조합해 활용한다.

3. **4대 유형의 강약점** — 엔터프라이즈 상용·오픈소스·클라우드 네이티브·AI 기반 도구는 각기 다른 강점을 가진다. 조직의 규모·성숙도·규제 환경에 맞게 선택해야 한다.
4. **모던 데이터 스택** — 각 영역의 최고 도구를 표준 인터페이스로 연결하는 접근이 확산되고 있다. 유연성이 큰 반면 통합·운영 복잡성도 증가한다.
5. **AI 기반 자동화의 부상** — 이상 탐지, 자동 태깅, 자동 계보 추적 등 AI 기반 기능이 급속히 확산되고 있다. 향후 5년간 이 영역이 도구 시장의 핵심 성장 동력이 될 것이다.
6. **체계적인 도구 선정** — 7대 기준의 가중치 평가 매트릭스, POC를 통한 검증, 통합 아키텍처 3원칙(메타데이터 우선·오픈 표준·API 우선)에 따라 체계적으로 선정해야 한다.
7. **도구는 만능이 아니다** — 가장 좋은 도구도 사람·프로세스가 뒷받침되지 않으면 무용지물이다. 도구는 이미 정의된 표준화 활동을 자동화·확장할 때 진정한 가치를 발휘한다.

## 다음 학습 가이드

### 10.5.4 주요 시장 보고서

- Gartner (2024d) — Market Guide for Data Management Solutions
- Gartner (2023b) — Magic Quadrant for Metadata Management Solutions
- IDC (2024a) — IDC Worldwide Data Management Software Forecast
- Forrester Research (2023a) — Forrester Wave: Data Governance Solutions

### 10.5.5 실무 참고 자료

- Cataldo and Foote (2022) — DataOps: The Authoritative Edition
- Bilgen (2024) — Building a Modern Data Catalog
- TDWI (2023a) — TDWI Best Practices Reports
- Andreessen Horowitz (2023) — Andreessen Horowitz Emerging Architectures

### 10.5.6 관련 본서 참조 장

- 제 5장 메타데이터 관리 — 카탈로그 도구 상세
- 제 6장 데이터 품질 관리 — 품질 도구 상세
- 제 7장 마스터 데이터 관리 — MDM 도구 상세
- 제 9장 데이터 거버넌스 체계 — 거버넌스 도구 상세
- 제 13장 클라우드 환경의 표준화 — 클라우드 네이티브 도구

## 연습 문제

1. 여러분 조직의 표준화 도구를 7대 카테고리별로 매핑하고, 각 카테고리의 성숙도를 5단계로 평가하시오.
2. 여러분 조직에 신규 데이터 카탈로그 도입을 제안한다면, 어떤 후보 도구를 검토할 것이며 어떤 POC 시나리오를 설계할 것인가?
3. 상용·오픈소스·클라우드 네이티브 중 여러분 조직에 가장 적합한 유형은 무엇이며, 그 이유는 무엇인가?

CHAPTER 11

# 실제 적용 사례

## 11.1 개요

이론과 방법론만으로는 데이터 표준화의 진정한 가치를 이해하기 어렵다. 실제 조직에서 어떻게 표준화가 시작되고, 어떤 저항을 극복하며, 어떤 성과를 만들어냈는지의 구체적 이야기가 방법론에 생명을 불어넣는다. Harvard Business Review (2023) 는 'Why Data Initiatives Fail' 특집에서 '데이터 프로젝트의 87%가 초기 목표를 달성하지 못한다'고 보고한다. 그러나 성공한 13%의 사례에는 공통 패턴이 있으며, 이를 학습하는 것이 본 장의 목적이다.

McKinsey Global Institute (2023) 는 'The data-driven enterprise of 2025' 보고서에서 데이터 표준화가 확립된 조직은 그렇지 않은 조직 대비 **3.2배 높은 EBITDA 성장률**을 보인다고 분석한다. 이러한 차이의 뒤에는 각 산업의 특성에 맞춘 정교한 표준화 전략이 있다.

### 11.1.1 본 장의 사례 선정 원칙

- 산업 다양성: 금융·공공·소매·헬스케어·제조 등 다양한 산업 포함
- 규모 다양성: 글로벌 대기업부터 국내 중견기업까지
- 성공과 실패의 균형: 실패 사례에서도 중요한 교훈이 있다
- 정량적 성과: 가능한 한 측정 가능한 결과 제시
- 최신성: 최근 5년 이내 사례 중심

### 11.1.2 본 장의 학습 목표

- 여러 산업의 데이터 표준화 접근 방식을 비교할 수 있다
- 성공과 실패 요인의 공통 패턴을 식별할 수 있다
- 여러분 조직에 적용 가능한 벤치마킹 요소를 도출할 수 있다
- 규제와 비즈니스 요구의 균형점을 이해한다

## 11.2 금융 산업 사례

### 11.2.1 사례 1: 글로벌 은행 G사의 BCBS 239 준수 여정

#### 11.2.1.1 배경

글로벌 시스템적 중요 은행(G-SIB)으로 지정된 G사는 2013년 발표된 Basel Committee on Banking Supervision (2013) '효과적인 리스크 데이터 집계 및 리스크 보고 원칙' 준수 마감이 다가오면서 전사 데이터 표준화 프로젝트를 시작해야 했다. 30개국 지사, 12개 핵심 뱅킹 시스템, 50개 이상의 데이터 마트가 존재하는 복잡한 환경이었다.

#### 11.2.1.2 도전 과제

- BCBS 239의 14개 원칙 준수 (2016년 마감)
- 동일 고객이 시스템마다 다른 ID로 존재
- 규제 보고서 작성에 8주 소요 (요구: 2주)
- 리스크 데이터의 계보 추적 불가능
- 연 규제 벌금 리스크 5억 달러

#### 11.2.1.3 접근 방법

1. 이사회 산하 '데이터 감독 위원회' 신설, CDO를 정기 보고자로 지정
2. ISO (2022c) 금융 메시징 표준을 기반으로 통합 데이터 모델 수립
3. 리스크 데이터 원장(Risk Data Register) 구축 — 400개 핵심 요소
4. Collibra 도입으로 전사 데이터 카탈로그 및 계보 관리
5. Informatica MDM으로 고객 마스터 통합 (30개국 순차 적용)
6. 데이터 스튜어드 800명 양성 (지역별 배치)
7. 실시간 데이터 품질 대시보드 (30개 도메인, 200개 KPI)

#### 11.2.1.4 성과 (3년)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 규제 보고서 작성 | 8주 | 2주 |
| 고객 통합 뷰 생성 | 수일 | 실시간 |
| 데이터 변환 매핑 수 | 약 450개 | 약 60개 |
| 신규 시스템 통합 비용 | 기준 100 | 35 |
| 규제 벌금 (5년 누적) | N/A (예상 5억$) | 0 |
| BCBS 239 준수 등급 | 미달 | Level 4 (우수) |

#### 11.2.1.5 교훈

빅뱅 방식이 아닌 단계적 접근이 성공의 핵심이었다. 30개국 자회사를 한 번에 표준화하려는 시도는 실패로 끝났을 것이다. 신규 시스템은 표준을 의무화하고, 기존 시스템은 매핑 레이어로 연결한 후 점진적으로 마이그레이션하는 전략이 효과적이었다. 규제 준수라는 '외부 마감'이 조직 내 정치적 저항을 극복하는 강력한 동력이 되었다.

### 11.2.2 사례 2: 국내 대형 은행 K사의 마이데이터 대응

#### 11.2.2.1 배경

2020년 8월 개정된 신용정보법 시행으로 국내 금융 산업은 마이데이터(본인신용정보관리업) 시대를 맞이했다. 국내 대형 은행 K사는 마이데이터 사업자 인가를 받고 서비스를 출시해야 했으나, 고객 데이터가 여러 계정계 시스템에 분산되어 있어 API 표준화가 시급한 과제였다 (금융결제원 2023).

#### 11.2.2.2 주요 과제

- 금융결제원의 마이데이터 표준 API 준수
- 100개 이상의 금융기관과 데이터 호환
- 계정계·정보계·인터넷뱅킹 시스템 통합
- 실시간 API 응답 (0.5초 이내)
- 개인정보 보호법 준수 (동의 관리 강화)

#### 11.2.2.3 접근 방법

1. 금융결제원 마이데이터 표준 API 100% 준수
2. Data Fabric 아키텍처 도입 — 데이터 이동 없이 통합
3. Federated API Gateway 구축 — 시스템별 API를 표준 API로 변환
4. 동의 관리 시스템 (Consent Management) 자체 개발
5. Redis 기반 캐싱으로 응답 속도 최적화

#### 11.2.2.4 성과 (1년)

- 서비스 출시 200일 만에 가입자 500만 명 돌파
- API 응답 평균 320ms (SLA 대비 36% 여유)
- 타 기관 데이터 통합 완료율 98%
- 월평균 API 호출 15억 건
- 금융권 마이데이터 사업 시장 점유율 1위

#### 11.2.2.5 교훈

규제가 만들어낸 표준화 요구가 오히려 비즈니스 혁신의 기회가 되었다. 기존 데이터 자산의 표준화 미비가 마이데이터 시대의 진입 장벽이 되었지만, 이를 계기로 근본적인 데이터 통합 아키텍처 개선을 이룬 것이다. Data Fabric 접근은 '데이터를 옮기지 않고 통합'이라는 관점 전환의 좋은 예시이다.

## 11.3 공공 부문 사례

### 11.3.1 사례 3: 미국 인구조사국의 메타데이터 표준화

#### 11.3.1.1 배경

미국 인구조사국(U.S. Census Bureau)은 1990년대부터 ISO/IEC (2023b) 기반 메타데이터 레지스트리를 구축한 대표 사례다. 10년마다 실시하는 인구조사, 매년 실시하는 American Community Survey, 그리고 수백 개의 통계 프로그램에서 생산되는 데이터의 일관성 확보와 국제 통계 비교 가능성이 목표였다 (U.S. Census Bureau 2020).

#### 11.3.1.2 장기 접근 (30년)

**1990년대: 초기 구축**

- ISO/IEC 11179 채택 및 정부 맞춤 확장
- 초기 표준 데이터 요소 약 500개 등록

**2000년대: 시스템화**

- 자체 메타데이터 레지스트리 시스템(MDR) 개발
- USGS 등 타 기관과의 표준 통합

**2010년대: 개방·확장**

- Data.gov 플랫폼과의 연계
- DCAT(Data Catalog Vocabulary) 표준 채택
- 데이터 요소 12,000개 이상으로 확장

**2020년대: 통합·자동화**

- Enterprise Data Lake와의 통합
- AI 기반 자동 메타데이터 태깅
- 연방 통계 기관 간 데이터 공유 표준 주도

#### 11.3.1.3 성과 (30년 누적)

| 지표 | 1990년 | 현재 |
|---|---|---|
| 표준화 데이터 요소 | 약 500개 | 약 12,000개 |
| 기관 간 데이터 공유 시간 | 수개월 | 수일~실시간 |
| 국제 비교 가능 데이터 | 30% | 85% |
| 시민 데이터 접근 건수 | 수천 건/년 | 수백만 건/월 |
| 연방 기관 표준 채택 | 인구조사국만 | 20+ 기관 |

#### 11.3.1.4 교훈

데이터 표준화는 단거리 경주가 아니라 세대에 걸친 마라톤이다. 인구조사국의 30년에 걸친 지속적 투자와 개선이 오늘날 미국 정부 데이터 생태계의 기반이 되었다. 특히 기술 트렌드가 바뀌어도 ISO/IEC 11179라는 근본 표준을 일관되게 적용해온 것이 장기적 성공의 핵심이다. 표준의 안정성이 시스템의 진화를 가능하게 한다.

### 11.3.2 사례 4: 한국 행정안전부의 공공데이터 표준화

#### 11.3.2.1 배경

한국은 2013년 공공데이터법 시행 이후 세계에서 가장 적극적인 공공 데이터 개방 정책을 추진해왔다. 행정안전부 (2023a) 를 주무 기관으로 하여 중앙부처, 광역·기초 자치단체, 공공기관 등 수천 개 기관의 데이터 표준화를 담당한다.

#### 11.3.2.2 주요 산출물

- 공공데이터 공통표준용어: 약 6,000개
- 공공데이터 표준 데이터셋: 약 1,500종
- 자치단체 행정정보 표준 데이터: 64종
- 공공데이터 품질관리 매뉴얼
- 공공데이터 표준 API 규격

#### 11.3.2.3 추진 체계

한국지능정보사회진흥원 (2023) 에 따르면 행정안전부는 표준화 추진을 위해 다음 3단 체계를 운영한다.

1. **정책 계층**: 행정안전부 공공데이터정책관
2. **전문 지원**: 한국지능정보사회진흥원(NIA)

3. **실행 지원**: 기관별 데이터 담당자 및 컨설팅단

#### 11.3.2.4 성과 (10년)

| 지표 | 2013년 | 2023년 |
|---|---|---|
| 개방 공공데이터 | 약 5,000건 | 약 90,000건 |
| 표준 데이터셋 | 0종 | 1,500종 |
| OECD 오픈 데이터 순위 | 5위 | 1위 (2019) |
| 활용 신청 건수 | 약 3,000건/년 | 약 60만 건/년 |
| 파생 서비스 수 | 약 100개 | 약 3,000개 |

#### 11.3.2.5 도전 과제와 한계

표준 수립과 실제 준수 사이의 격차가 여전히 존재한다. 특히 표준 적용에 따른 기존 시스템 개편 비용이 각 기관에 그대로 남기 때문에 실무 저항이 상당하다. 표창과 평가 가점 등 인센티브 제도로 대응하고 있지만, 근본적으로는 표준 적용에 따른 예산 지원이 필요하다는 지적이 있다.

#### 11.3.2.6 교훈

정부 주도 표준화는 강력한 추진력을 갖지만, 각 기관이 체감하는 가치가 없으면 형식적 준수에 그친다. 표준 수립과 인센티브 제공이 함께해야 하며, 특히 표준 적용에 따르는 실무 부담을 정부가 일부 부담하는 지원 체계가 중요하다. 한국의 OECD 오픈 데이터 순위 1위 달성은 성공을 보여주지만, 그 배후에는 지속적 정책 지원과 각 기관의 노력이 있다.

## 11.4 소매·서비스 산업 사례

### 11.4.1 사례 5: 글로벌 통신사 V사의 고객 데이터 품질 개선

#### 11.4.1.1 배경

글로벌 통신사 V사는 1억 명 이상의 고객 데이터를 보유하고 있었지만, 인수한 다양한 자회사로부터 통합된 고객 데이터에 심각한 품질 문제가 있었다. 마케팅 캠페인 성과 저조와 GDPR 준수 요구가 동시에 압박을 가하며 전사적 데이터 품질 개선 프로젝트가 시작되었다 (European Data Protection Board 2023).

#### 11.4.1.2 도전 과제

- 중복 고객 추정 비율: 18%
- 유효 연락처 보유율: 71%
- GDPR 요청 응대 지연 (평균 수주)
- 실시간 마케팅 자동화 요구
- 여러 자회사의 상이한 고객 데이터 스키마

#### 11.4.1.3 접근 방법

1. 6대 차원 기반 품질 진단 (3개월)
2. Informatica DQ 도입 및 정제 규칙 자동화
3. 엔티티 해소(Entity Resolution) 알고리즘 적용
4. 실시간 데이터 검증 API 구축
5. 데이터 스튜어드 50명 양성
6. Golden Record 생성 및 하위 시스템 동기화

#### 11.4.1.4 성과 (18개월)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 중복 고객률 | 18% | 1.2% |
| 유효 연락처 비율 | 71% | 94% |
| 캠페인 응답률 | 기준 100 | 187 |
| GDPR 응대 시간 | 수주 | 수일 |
| 마케팅 ROI | 기준 100 | 156 |
| 고객 만족도 (NPS) | +12 | +34 |

#### 11.4.1.5 교훈

데이터 품질은 IT 프로젝트가 아니라 비즈니스 변화 관리 프로젝트이다. 마케팅 ROI 향상이라는 명확한 비즈니스 가치를 제시하고 데이터 스튜어드 양성을 통해 조직 차원의 책임 의식을 확보한 것이 성공의 핵심이었다. 특히 규제(GDPR) 준수와 비즈니스 가치(마케팅 ROI)를 동시에 달성한 이중 목표 설정이 예산 확보와 지속적 지원의 근거가 되었다.

### 11.4.2 사례 6: 미국 대형 소매 기업의 옴니채널 데이터 통합

#### 11.4.2.1 배경

미국의 한 대형 소매 기업은 매장 5,000개, 온라인 채널, 물류 센터에서 매일 수억 건의 고객 데이터를 생성했다. 그러나 각 채널이 독립적으로 운영되어 동일 고객이 채널마다 다르게 인식되었고, 옴니채널 고객 경험 제공이 불가능했다 (MIT Sloan CISR 2021).

#### 11.4.2.2 도전 과제

- 부서별 데이터 사일로 문화
- 고객 정보의 부서 간 불일치
- CCPA(캘리포니아 소비자 개인정보 보호법) 대응 요구
- 실시간 재고 및 개인화 요구

#### 11.4.2.3 접근 방법

1. Non-Invasive 거버넌스 접근 채택
2. CDO 신설, 마케팅·영업·물류 겸직 위원회 구성
3. 고객 마스터부터 단계적 시작 (파일럿)
4. Snowflake 기반 고객 데이터 플랫폼(CDP) 구축
5. 우수 데이터 스튜어드 시상 프로그램
6. 연 1회 전사 'Data Day' 이벤트로 문화 조성

#### 11.4.2.4 성과 (3년)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 고객 통합 뷰 정확도 | 62% | 94% |
| 부서 간 데이터 요청 처리 | 2주 | 2일 |
| CCPA 응대 시간 | 수주 | 24시간 이내 |
| 옴니채널 전환율 | 기준 100 | 142 |

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 직원 거버넌스 인식 (설문) | 37% | 82% |

#### 11.4.2.5 교훈

새로운 조직을 만들기보다 기존 부서의 데이터 담당자를 공식화·연결하는 Non-Invasive 접근이 대기업 소매업에도 효과적이었다. 특히 'Data Day' 같은 문화 조성 활동이 정책 문서보다 정착에 더 크게 기여했다. 임원 겸직 위원회가 부서 간 이해관계 조정에 결정적 역할을 했다.

## 11.5 헬스케어 산업 사례

### 11.5.1 사례 7: 미국 대형 의료 시스템의 환자 안전

#### 11.5.1.1 배경

미국의 한 대형 의료 시스템(30개 병원, 연 환자 500만 명)은 환자 안전 사고의 상당수가 데이터 오류에서 기인함을 인식했다. Johns Hopkins 대학의 Makary and Daniel (2016) 연구에 따르면 미국에서 의료 오류로 인한 사망은 연간 약 25만 명에 달하며, 이는 심장질환·암에 이은 세 번째 사망 원인이다.

#### 11.5.1.2 데이터 오류 유형

- 환자 식별 오류 (동명이인, 오타)
- 약물 처방 데이터 오류
- 알레르기 정보 누락
- 검사 결과의 부서 간 전달 지연

#### 11.5.1.3 접근 방법

1. 환자 마스터 인덱스(EMPI) 구축 — 환자 식별 일관성
2. HL7 FHIR 표준 채택 — 시스템 간 데이터 일관성
3. 약물 데이터베이스 표준화 (RxNorm)
4. 실시간 데이터 검증 (처방 시점 약물 상호작용 확인)
5. 의료진 데이터 품질 교육 (연 8시간 의무)
6. 데이터 오류 보고 무처벌 정책 (Just Culture)

#### 11.5.1.4 성과 (2년)

| 지표 | 변화율 |
|---|---|
| 환자 식별 오류 | 87% 감소 |
| 처방 오류 | 64% 감소 |
| 환자 만족도 | 31% 향상 |
| 의료 분쟁 건수 | 45% 감소 |
| 의료 소송 비용 | 58% 감소 |
| 직원 만족도 | 22% 향상 |

#### 11.5.1.5 교훈

데이터 품질이 직접적으로 인간 생명·안전과 연결되는 영역에서는 기술적 통제만으로 부족하다. 데이터를 입력하는 사용자(의사·간호사)의 인식 변화와 실시간 검증의 결합이 필수적이다. 특히 오류 보고

에 대한 무처벌 정책(Just Culture)이 실질적 오류 감소의 핵심이었다. 처벌 두려움 없이 오류를 보고할 수 있어야 근본 원인 파악과 재발 방지가 가능하다.

### 11.5.2 사례 8: 국내 대학병원의 CDISC 표준화

#### 11.5.2.1 배경

국내 대학병원 S병원은 신약 개발을 위한 임상시험을 다수 수행하고 있었으나, 각 연구실이 독자적인 데이터 형식을 사용해 국제 임상시험 참여와 규제기관 제출에 큰 어려움을 겪었다. FDA와 EMA는 임상시험 데이터에 대해 CDISC(Clinical Data Interchange Standards Consortium) 표준을 요구한다 (CDISC 2023a).

#### 11.5.2.2 도전 과제

- 연구실별 상이한 데이터 형식
- CDISC SDTM(Study Data Tabulation Model) 준수 요구
- 국제 다기관 임상시험 참여 확대
- 연구자의 표준화 학습 부담

#### 11.5.2.3 접근 방법

1. CDISC 표준화 지원 센터 신설
2. 연구자용 CDISC 교육 프로그램 (연 2회)
3. SDTM 매핑 도구 라이선스 확보
4. 표준 CRF(Case Report Form) 템플릿 개발
5. 데이터 매니저 팀 확대 (4명 → 15명)

#### 11.5.2.4 성과 (3년)

- CDISC 준수 임상시험 5개 → 45개
- 국제 다기관 시험 참여 3배 증가
- FDA/EMA 데이터 제출 지연 사례 0건
- 신규 임상시험 유치 40% 증가 (연간)
- 연구비 수주 규모 65% 증가

#### 11.5.2.5 교훈

국제 표준 준수는 단순한 규제 대응이 아니라 국제 시장 참여의 입장권이다. S병원은 CDISC 표준화를 통해 국제 다기관 시험 참여와 글로벌 제약사 협력의 기반을 마련했다. 표준화가 곧 비즈니스 기회 확대로 직결되는 명확한 사례이다. 특히 표준화 지원 조직(센터)의 신설이 개별 연구자의 부담을 줄이는 결정적 역할을 했다.

## 11.6 제조 산업 사례

### 11.6.1 사례 9: 글로벌 제조 T사의 IoT 데이터 표준화

#### 11.6.1.1 배경

글로벌 자동차 부품 제조사 T사는 30개국 80개 공장에서 매일 페타바이트 규모의 IoT 센서 데이터를 생성한다. 스마트팩토리 전환을 위해 ISA-95(Enterprise-Control System Integration) 표준 기반의 통합 데이터 아키텍처 구축이 필요했다 (ISA 2018a).

#### 11.6.1.2 도전 과제

- 공장마다 상이한 MES/SCADA 시스템
- 센서 데이터 형식 표준화 부재

- 실시간 데이터 처리 요구 (100만 건/초)
- 공장별 데이터 주권 요구

#### 11.6.1.3 접근 방법

1. ISA-95 기반 표준 데이터 모델 수립
2. OPC UA 프로토콜로 센서 데이터 표준화
3. Edge Computing으로 공장별 사전 처리
4. Kafka 기반 스트리밍 아키텍처
5. 공장별 시맨틱 레이어 구축
6. Digital Twin 표준 (Asset Administration Shell) 채택

#### 11.6.1.4 성과 (2년)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 예측 유지보수 정확도 | 62% | 89% |
| 설비 다운타임 | 기준 100 | 43 |
| 생산 계획 정확도 | 78% | 94% |
| 신규 공장 통합 기간 | 18개월 | 6개월 |
| 에너지 효율 | 기준 100 | 121 |

#### 11.6.1.5 교훈

제조 산업의 IoT 데이터 표준화는 표준의 선택이 곧 아키텍처의 방향성을 결정한다. ISA-95와 OPC UA라는 잘 정립된 산업 표준을 채택함으로써 벤더 락인을 피하고 확장성을 확보할 수 있었다. 특히 Edge Computing으로 각 공장의 데이터 주권을 유지하면서도 글로벌 표준을 적용하는 하이브리드 접근이 문화적·규제적 이슈를 동시에 해결했다.

### 11.6.2 사례 10: 한국 중견 제조업체의 실패와 재도전

#### 11.6.2.1 배경

국내 중견 제조업체 D사는 2018년 첫 번째 데이터 표준화 프로젝트를 시작했으나 1년 만에 좌초되었다. 그 후 원인 분석을 거쳐 2020년 재도전하여 성공한 사례이다 (한국데이터산업진흥원 2022).

#### 11.6.2.2 첫 번째 시도 (2018)의 실패 원인

- IT 부서 단독 추진 — 비즈니스 부서 참여 부재
- 외부 컨설팅 의존 — 내부 역량 미축적
- 빅뱅 방식 채택 — 6개월 내 전사 적용 시도
- 측정 지표 부재 — 진행 상황 가시화 실패
- 경영진 후원 부족 — 중간에 우선순위 하락

#### 11.6.2.3 재도전 (2020)의 접근

1. CEO 직속 데이터 거버넌스 위원회 신설
2. 재무 부서 단일 도메인 파일럿 (6개월)
3. 내부 데이터 스튜어드 양성 (외부 코칭 활용)
4. 월간 임원 보고로 가시성 확보
5. 3년에 걸친 단계적 확장 계획
6. 실패 사례집 발간 및 전사 공유

#### 11.6.2.4 성과 (3년)

- 전 부서 데이터 표준화 완료
- 데이터 관련 의사결정 시간 70% 단축
- 연간 데이터 관련 오류 90% 감소
- 내부 데이터 전문가 15명 양성

#### 11.6.2.5 교훈

데이터 표준화 실패는 흔하지만 회복할 수 있다. 실패 경험은 두 번째 시도의 강력한 자산이 된다. 단, 같은 실수를 반복하지 않으려면 첫 번째 실패의 원인을 정직하게 분석하고 공유하는 문화가 필요하다. D사의 경우 'lessons learned' 워크숍을 통해 조직 전체가 실패의 교훈을 학습한 것이 재도전의 출발점이었다. 실패를 숨기지 않는 조직 문화가 진정한 학습을 가능하게 한다.

## 11.7 종합 교훈

### 11.7.1 성공 사례의 공통 패턴

10개 사례에서 도출한 성공의 공통 요인 6가지는 다음과 같다. 이는 MIT CISR의 데이터 관리 성숙도 연구와도 일치한다 (MIT Sloan CISR 2021).

| 성공 요인 | 설명 |
|---|---|
| 명확한 비즈니스 가치 | 규제 준수, 매출 증대, 비용 절감 등 정량화된 목표 |
| 점진적 접근 | 빅뱅 방식 회피, 파일럿 → 확장 단계별 진행 |
| 거버넌스 우선 | 기술보다 R&R, 정책, 조직 정비가 우선 |
| 도구와 사람의 균형 | 도구만으로는 정착 불가, 스튜어드 양성 필수 |
| 측정 가능한 KPI | 측정 없이는 개선 없음, 대시보드 상시 공개 |
| 외부 표준 활용 | 조직 단독보다 ISO, BCBS, HL7 등 검증된 표준 채택 |

### 11.7.2 실패 사례의 공통 패턴

실패한 데이터 표준화 프로젝트에서 반복적으로 관찰되는 공통 원인 6가지이다 (Harvard Business Review 2023).

| 실패 원인 | 설명 |
|---|---|
| IT 단독 추진 | 비즈니스 참여 없이 IT만의 프로젝트로 진행 |
| 빅뱅 방식 | 짧은 기간에 전사 적용을 시도하여 파일럿 학습 불가 |
| 측정 부재 | 진행 상황 가시화 없이 추진, 성과 입증 어려움 |
| 경영진 이탈 | 중간에 우선순위 하락, 예산·자원 축소 |
| 도구 우선 | 거버넌스 없이 비싼 도구 도입, 활용도 저조 |
| 문화 무시 | 저항 관리 없이 강제 적용, 형식화·표류 |

### 11.7.3 산업별 특화 교훈

| 산업 | 핵심 성공 요인 | 주의점 |
|---|---|---|
| 금융 | 규제 준수를 동력으로 활용, 이사회 수준 거버넌스 | 체크박스 컴플라이언스 함정 |
| 공공 | 정책·인센티브 결합, 장기적 지속성 | 형식적 준수, 실질 활용 부족 |
| 소매/서비스 | 고객 통합 뷰, 옴니채널 경험 | 채널 간 사일로 문화 |
| 헬스케어 | 환자 안전과 직결, 무처벌 문화 | 데이터 오류가 곧 인명 리스크 |
| 제조 | 산업 표준(ISA-95, OPC UA) 활용, Edge 아키텍처 | 벤더 락인, 공장별 저항 |

## 11.8 여러분 조직에 적용하기

### 11.8.1 벤치마킹 워크시트

본 장의 사례를 여러분 조직에 적용하기 위한 6단계 벤치마킹 프레임워크이다 (MIT Sloan CISR 2021).

1. 여러분 조직과 가장 유사한 사례 2~3개 선정
2. 각 사례의 초기 상황, 접근 방법, 성과를 상세히 분석
3. 여러분 조직과의 유사점과 차이점 매핑
4. 적용 가능한 요소와 그렇지 않은 요소 구분
5. 적용 시 예상되는 저항과 대응 방안 도출
6. 초기 3개월 액션 플랜 수립

### 11.8.2 시작을 위한 체크리스트

💡🚀 데이터 표준화 시작 체크리스트

첫 단계로 다음을 확인하세요:

- ☐ 표준화의 비즈니스 가치가 명확한가?
- ☐ 경영진 후원자가 확보되었는가?
- ☐ 파일럿 도메인 1~2개를 선정했는가?
- ☐ 성공 지표(KPI)가 정의되었는가?
- ☐ 데이터 오너·스튜어드가 지정되었는가?
- ☐ 6개월 내 첫 성과를 낼 수 있는가?
- ☐ 변화 관리 계획이 있는가?

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **사례는 이론에 생명을 부여한다** — 10개 사례를 통해 방법론이 실제 조직에서 어떻게 작동하는지 확인했다. 각 산업의 특성이 표준화 접근에 반영되어야 한다.
2. **점진적 접근이 거의 항상 우월하다** — 빅뱅 방식으로 성공한 사례를 찾기 어렵다. 파일럿 도메인에서 성공 사례를 만들고 점진적으로 확장하는 것이 검증된 패턴이다.
3. **규제는 강력한 동력이 된다** — BCBS 239, GDPR, CDISC, CCPA 등 외부 규제 요구는 조직 내 정치적 저항을 극복하는 강력한 지렛대가 된다.

4. **글로벌 표준 활용의 가치** — ISO/IEC (2023b), ISO (2022c), HL7 FHIR, ISA-95 등 검증된 산업 표준의 채택이 조직 단독의 표준보다 훨씬 효과적이다.
5. **문화가 최종 결정 요인** — 가장 좋은 방법론과 도구를 갖추어도 조직 문화가 뒷받침되지 않으면 정착하지 않는다. 리더십, 인센티브, 교육의 결합이 핵심이다.
6. **실패는 회복 가능하다** — 첫 시도의 실패는 흔하지만 학습을 통해 두 번째 시도에서 성공할 수 있다. 실패의 원인을 정직하게 분석하고 공유하는 문화가 필요하다.
7. **측정과 인정의 힘** — 정량적 KPI로 성과를 측정하고, 우수 스튜어드를 공식적으로 인정하는 것이 지속적 정착의 원동력이다.

## 다음 학습 가이드

### 11.8.3 추가 사례 자료

- MIT Sloan CISR (2021) — MIT Sloan CISR 데이터 관리 연구
- Harvard Business Review (2023) — Harvard Business Review 데이터 관리 특집
- McKinsey Global Institute (2023) — McKinsey Global Institute 보고서
- 한국데이터산업진흥원 (2022) — 한국데이터산업진흥원 사례집

### 11.8.4 관련 본서 참조 장

- 제 8장 표준화 구현 전략 — 사례의 방법론 배경
- 제 9장 데이터 거버넌스 체계 — 거버넌스 구조
- 제 10장 표준화 도구와 기술 — 사례에서 활용된 도구
- 제 14장 국제 표준 심화 — 사례에서 언급된 표준의 상세

## 연습 문제

1. 본 장의 10개 사례 중 여러분 조직과 가장 유사한 사례 2개를 선정하고, 그 이유와 벤치마킹 요소를 정리하시오.
2. 여러분 조직에서 이전에 시도한 데이터 관련 프로젝트의 성공/실패 요인을 본 장의 공통 패턴과 대조하여 분석하시오.
3. 여러분 산업의 특화 교훈에서 아직 적용하지 못한 요소는 무엇이며, 이를 도입하기 위한 3개월 액션 플랜을 수립하시오.

# IV 4부: 고급 주제


12 AI·ML과 데이터 표준화 . . . . . 137
12.1 개요 . . . . . 137
12.2 핵심 개념 . . . . . 138
12.3 방법론 . . . . . 140
12.4 실무 적용 . . . . . 141
12.5 사례 연구 . . . . . 143

13 클라우드 환경의 표준화 . . . . . 147
13.1 개요 . . . . . 147
13.2 핵심 개념 . . . . . 148
13.3 방법론 . . . . . 150
13.4 실무 적용 . . . . . 151
13.5 사례 연구 . . . . . 153

14 국제 표준 심화 . . . . . 157
14.1 개요 . . . . . 157
14.2 ISO/IEC 11179: 메타데이터 등록소 심화 . . 158
14.3 ISO 8000: 데이터 품질 파트별 심화 . . . . . 160
14.4 ISO/IEC 25012: 데이터 품질 모델 심화 . . . 161
14.5 ISO/IEC 42001: AI 관리 시스템 심화 . . . . 162
14.6 표준 조합과 통합 활용 . . . . . 164
14.7 사례 연구 . . . . . 166

CHAPTER 12

# AI·ML과 데이터 표준화

## 12.1 개요

인공지능과 머신러닝은 지난 10년간 비즈니스의 운영 방식을 근본부터 바꿔놓았다. 데이터로부터 학습하는 시스템이 의사결정, 예측, 자동화의 핵심 동력이 되면서, 그 동력의 원천인 데이터의 품질과 표준화는 기존 IT 시스템 시대와는 비교할 수 없을 만큼 중요해졌다. Google 연구팀의 Sambasivan et al. (2021) 은 'Everyone wants to do the model work, not the data work' 라는 논문에서 **AI 프로젝트 실패의 92%가 데이터 문제에서 기인**한다고 보고한다.

전통적 정보 시스템에서 데이터의 역할은 '기록(record)'이었다. AI·ML 시대에 데이터의 역할은 '학습 재료(learning material)'로 확장된다. 모델은 데이터의 패턴을 학습하고 일반화하여 새로운 상황에 적용한다. 이때 학습 데이터에 내재된 편향, 결측, 불일치는 단순히 보고서의 부정확함을 넘어, **모델이 잘못된 의사결정을 자동화하고 확산**시키는 결과를 낳는다. 'Garbage In, Garbage Out'은 AI 시대에 'Garbage In, Garbage Automated'로 변형된다.

이러한 인식은 규제로 이어지고 있다. 2024년 발효된 European Parliament (2024) 는 고위험 AI 시스템에 대해 학습 데이터의 품질·거버넌스·문서화를 명시적으로 요구하며, 위반 시 매출의 최대 7% 벌금을 부과할 수 있다. NIST (2023) 와 ISO/IEC (2023a) 도 데이터 표준화를 AI 시스템 신뢰성의 핵심 요건으로 규정한다.

### 12.1.1 핵심 관점

AI는 데이터 표준화를 대체하지 않는다. 오히려 **더 정교하고, 더 다층적이며, 더 검증된 표준화를 요구**한다. 이 장은 다음 세 가지 관점 전환을 다룬다:

1. 데이터 표준화 대상의 확장 (컬럼 → 피처·데이터셋·모델)
2. 새로운 품질 차원의 등장 (편향성·공정성·설명가능성)
3. 조직 협업의 확장 (데이터 팀 + ML 팀 + 법무·윤리 팀)

### 12.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- AI·ML 프로젝트가 데이터에 요구하는 새로운 품질 차원을 식별할 수 있다
- 학습 데이터셋, 피처, 모델 메타데이터의 표준화 체계를 설계할 수 있다
- Model Card와 Datasheet 표준을 조직에 적용할 수 있다
- MLOps 파이프라인에 데이터 표준화 검증을 통합하는 방법을 설명할 수 있다
- AI 거버넌스와 데이터 거버넌스의 접점 이슈를 다룰 수 있다
- 규제(European Parliament (2024), ISO/IEC (2023a))가 요구하는 데이터 문서화 수준을 이해한다

## 12.2 핵심 개념

### 12.2.1 학습 데이터의 5대 표준화 차원

학습 데이터는 단순히 '많은 데이터'가 아니라 특정 문제 도메인을 대표하면서 다양성을 가진 정교한 자산이다. Stanford의 Andrew Ng이 주도한 Data-Centric AI 운동은 모델 성능 향상의 열쇠가 알고리즘이 아닌 데이터 품질에 있다는 통찰을 확산시켰다 (Ng 2021). 학습 데이터 표준화는 다음 5개 차원에서 이루어진다.

| 차원 | 정의 | 표준화 산출물 |
|---|---|---|
| 대표성(Representativeness) | 타겟 모집단의 통계적 분포를 반영 | 샘플링 기준서 |
| 균형성(Balance) | 클래스·세그먼트별 적절한 비율 | 클래스 분포표 |
| 신선도(Freshness) | 현재 시점의 현실 반영 | 데이터 갱신 주기 표준 |
| 라벨 일관성<br>(Label Consistency) | 라벨이 어노테이터·시점에 무관하게 일관 | 어노테이션 가이드, 합의도 |
| 출처 추적성(Provenance) | 각 데이터 포인트의 출처·변형 이력 | 데이터 계보(Lineage) 문서 |

### 12.2.2 피처(Feature) 표준화

피처는 원시 데이터로부터 모델이 학습 가능한 형태로 변환된 입력 변수이다. 전통적 표준화가 '컬럼' 단위였다면 AI 시대의 표준화는 **'피처' 단위**로 확장된다. Hutter et al. (2019) 에 따르면, 피처 엔지니어링의 부실이 AI 모델 실패의 약 40%를 차지한다.

#### 12.2.2.1 피처 정의서의 표준 구성 요소

1. 피처 ID 및 명칭
2. 비즈니스 정의 (자연어 설명)
3. 원천 데이터 매핑
4. 변환 로직 (SQL, Python, 의사코드)
5. 값의 도메인 (연속형/범주형, 범위)
6. 통계적 특성 (평균·분산·분포, 모니터링 기준값)
7. 의존성 (사용 모델 목록, 다운스트림 영향)
8. 소유자 및 SLA (지연·품질 보장)

> 피처 스토어 (Feature Store)
>
> 피처 표준화의 인프라적 구현체이다. Tecton, Feast, Databricks Feature Store, AWS SageMaker Feature Store 등이 대표 솔루션이다. Uber가 2017년 발표한 Michelangelo 플랫폼이 피처 스토어 개념의 시초로 여겨진다 (Uber Engineering 2017). 핵심 가치는 (1) 학습용과 추론용 피처가 동일 변환 로직을 거치도록 보장(Training-Serving Skew 방지), (2) 조직 차원의 피처 재사용 가능, (3) 피처 계보와 감사 이력 자동화이다.

### 12.2.3 모델 메타데이터: Model Card

AI 시대에는 모델 자체가 메타데이터 관리 대상이 된다. Mitchell 등이 FAT* 2019 컨퍼런스에서 제안한 Model Card 는 모델 문서화의 표준이 되었다 (Mitchell et al. 2019). Google, Microsoft, Meta 등 주요 AI 기업이 자사 모델에 Model Card를 필수적으로 발행하고 있으며, Hugging Face 등 오픈소스 모델 허브에서도 표준 양식으로 채택되었다.

| 섹션 | 내용 |
|---|---|
| 모델 상세 | 모델명, 버전, 개발 조직, 릴리스 일자, 라이선스, 문의처 |
| 의도된 용도 | 주 사용 사례, 대상 사용자, 사용 범위 외 시나리오 |
| 학습 데이터 | 데이터셋명, 크기, 수집 기간, 전처리, 배포 |
| 평가 데이터 | 평가 데이터셋, 평가 지표, 세그먼트 분석 |
| 성능 지표 | 전체·세그먼트별 성능, 신뢰 구간, 벤치마크 비교 |
| 윤리적 고려사항 | 잠재적 편향, 위험 시나리오, 완화 조치 |
| 주의사항 및 권고 | 알려진 한계, 권장 사용 조건, 유지보수 이력 |

### 12.2.4 데이터셋 카드(Dataset Card)

데이터셋도 모델과 마찬가지로 표준화된 문서를 가져야 한다. Gebru et al. (2021) 이 제안한 Datasheets for Datasets 는 AI 학습 데이터셋의 표준 문서 양식이다. 이 표준은 데이터셋의 수집 동기부터 유지보수까지 전 생애주기를 다룬다.

#### 12.2.4.1 Dataset Card의 9대 섹션

- 기본 정보: 데이터셋명, 버전, 라이선스, 배포처
- 수집 동기: 왜 만들어졌는가, 자금 출처, 제작 조직
- 구성: 인스턴스 수·유형·분할 비율(train/val/test)
- 수집 프로세스: 방법, 도구, 참여자, 동의 여부
- 전처리: 정제·필터링·이상치 처리·비식별화
- 라벨링: 라벨 정의, 어노테이터 정보, 합의도(IAA)
- 알려진 편향: 편향 원인, 영향 세그먼트, 완화 노력
- 사용 권장 사항: 권장·금지 사용 사례, 라이선스 조건
- 유지·관리: 유지보수 주체, 갱신 주기, 폐기 계획

### 12.2.5 새로운 품질 차원: 편향성과 공정성

전통적 6대 품질 차원(완전성·정확성·일관성·유효성·유일성·적시성)에 AI는 두 가지 새로운 차원을 추가한다. IBM Research (2023) 은 이를 'AI 시대의 데이터 품질 확장'이라 부른다.

#### 12.2.5.1 편향성(Bias)의 유형

| 유형 | 정의 | 예시 |
|---|---|---|
| 역사적 편향 | 과거 사회적 불평등이 데이터에 반영 | 채용 데이터의 성별 편향 |
| 표본 편향 | 수집 방법으로 인한 데이터 집단 편향 | 온라인 설문의 연령대 편향 |
| 측정 편향 | 측정 도구·방법의 체계적 오류 | 센서 정확도의 조건별 차이 |
| 라벨 편향 | 어노테이터의 주관·문화적 배경 영향 | 감정 분석의 문화권 차이 |
| 집계 편향 | 이질적 하위 집단을 하나로 취급 | 인종별 성능 차이 은폐 |

#### 12.2.5.2 공정성(Fairness) 측정 지표

공정성은 모델 출력이 보호 속성(성별, 인종, 연령 등)에 따라 차별적이지 않은지의 평가이다. 여러 통계적 정의가 존재하며, 이들은 상호 배타적이어서 모두를 동시에 만족할 수 없다는 '불가능성 정리(Impossibility Theorem)'가 있다 (Kleinberg et al. 2017).

- **통계적 동등성(Statistical Parity)**: 그룹 간 양성 예측 비율 동등
- **기회 균등(Equal Opportunity)**: 실제 양성 중 예측 양성 비율 동등
- **예측 동등성(Predictive Parity)**: 예측 양성 중 실제 양성 비율 동등
- **Calibration**: 예측 확률과 실제 확률의 일치

## 12.3 방법론

### 12.3.1 AI-Ready 표준화 프레임워크

전통적 표준화를 AI 시대에 맞게 확장한 6레이어 프레임워크는 다음과 같다. 각 레이어는 서로 다른 표준화 대상과 산출물을 가지며, 아래 레이어의 표준이 확립되어야 위 레이어가 안정적으로 작동한다 (IBM Research 2023).

| 레이어 | 표준화 대상 | 주요 산출물 |
|---|---|---|
| 1. 원천 데이터 | 수집·기록 표준 | 데이터 사전, 원천 계보 |
| 2. 데이터셋 | 학습/검증/평가 구성 | 데이터셋 카드 |
| 3. 피처 | 정의·변환·검증 | 피처 정의서, 피처 스토어 |
| 4. 모델 | 학습·평가·문서화 | Model Card, 성능 리포트 |
| 5. 파이프라인 | 재현성·자동화 | 파이프라인 명세, 실험 이력 |
| 6. 거버넌스 | 조직·정책·모니터링 | AI 윤리 정책, 감사 체계 |

### 12.3.2 데이터 계보와 재현성

AI 모델 결과는 사용된 데이터의 정확한 버전, 변환, 분할에 따라 달라진다. 따라서 재현 가능한 AI를 위해 다음 세 요소가 모두 추적되어야 한다 (Sculley et al. 2015).

1. **데이터 버전** — DVC, LakeFS, Pachyderm 등의 도구 활용
2. **코드 버전** — Git 커밋 해시 기록
3. **환경 정의** — 라이브러리 버전, 난수 시드, 하드웨어 사양

! 데이터 버전 관리는 필수

AI 규제가 강화되는 추세에서, '이 모델이 어떤 데이터로 학습되었는가'를 입증하지 못하면 감사·소송·인증 절차에서 심각한 문제가 발생한다. NIST (2023) 와 ISO/IEC (2023a) 모두 데이터 버전 관리를 명시적으로 요구한다. 특히 금융·의료 등 규제 산업에서는 학습 데이터의 완전한 재현 가능성이 배포 승인의 전제조건이 되고 있다.

### 12.3.3 데이터 드리프트 모니터링

모델은 학습 시점의 데이터 분포를 가정하고 예측한다. 시간이 흐르면 현실의 데이터 분포가 학습 시 분포와 달라지는 **데이터 드리프트(Data Drift)**가 발생한다. 이는 모델 성능 저하의 주요 원인이며, 지속적 모니터링이 필수적이다 (Rabanser et al. 2019).

| 지표 | 측정 대상 | 임계값 |
|---|---|---|
| PSI | 피처 분포의 안정성 | PSI 〉 0.2 → 경고 |
| KS-Statistic | 두 분포 간 차이 | p 〈 0.05 → 의심 |
| KL Divergence | 확률 분포 간 차이 | 도메인별 설정 |
| 피처 결측률 변화 | 결측 패턴 변화 | ±5%p 변동 시 검토 |
| 예측 분포 변화 | 모델 출력 분포 변화 | 전월 대비 ±10% |
| 개념 드리프트 | 입력-출력 관계 변화 | 성능 저하 관찰 시 |

#### 12.3.3.1 드리프트 유형

- **공변량 드리프트(Covariate Drift)**: 입력 분포 변화, 관계는 유지
- **개념 드리프트(Concept Drift)**: 입력과 출력의 관계 자체 변화
- **라벨 드리프트(Label Drift)**: 타겟 분포 변화

## 12.4 실무 적용

### 12.4.1 단계별 도입 로드맵

AI 데이터 표준화 체계 구축은 조직의 AI 성숙도에 맞춰 단계적으로 추진해야 한다. Google, Microsoft, DataRobot 등의 사례를 종합한 표준 로드맵은 다음과 같다 (Databricks 2024b).

| 단계 | 기간 | 주요 활동 | 주요 산출물 |
|---|---|---|---|
| Stage 1: 파일럿 | 0~6개월 | 핵심 데이터셋 카드 작성, AI 윤리 정책 초안 | 1~2개 프로젝트 표준 적용 |
| Stage 2: 인프라 | 6~12개월 | 피처 정의서, 피처 스토어 POC, 실험 추적 | 피처 스토어, MLflow 도입 |
| Stage 3: 통합 | 12~18개월 | MLOps에 품질 검증 통합, 드리프트 모니터링 | 자동화된 품질 게이트 |
| Stage 4: 거버넌스 | 18개월~ | 전사 AI 거버넌스, ISO/IEC 42001 인증 | AI 거버넌스 위원회, 감사 체계 |

### 12.4.2 도구 스택

AI 데이터 표준화를 지원하는 도구는 오픈소스와 상용 모두 빠르게 성숙하고 있다 (Huyen 2022).

| 영역 | 오픈소스 | 상용 |
|---|---|---|
| 데이터 카탈로그 | DataHub, OpenMetadata | Collibra, Atlan |
| 피처 스토어 | Feast | Tecton, Databricks FS, AWS SageMaker FS |
| 데이터 버전 관리 | DVC, LakeFS, Pachyderm | W&B Artifacts |
| 실험 추적 | MLflow, Aim | W&B, Comet, Neptune |
| 드리프트 모니터링 | Great Expectations, Evidently AI, Deepchecks | Monte Carlo, Arize, Fiddler |
| 편향·공정성 평가 | Fairlearn, AIF360, What-If Tool | Credo AI, Fiddler |

### 12.4.3 규제 준수 요구사항

2024년 발효된 European Parliament (2024) 는 고위험 AI 시스템에 대해 학습 데이터의 특정 표준을 요구한다.

#### 12.4.3.1 EU AI Act의 데이터 관련 요구사항

- 학습·검증·테스트 데이터셋의 품질 기준 문서화
- 대표성 있는 데이터로부터 편향 위험 식별·완화
- 데이터 거버넌스 절차의 명확한 정의
- 개인정보 처리에 대한 GDPR 준수
- 데이터 계보 및 처리 이력 유지

#### 12.4.3.2 ISO/IEC 42001 (AI 관리 시스템) 요구사항

2023년 12월 발효된 ISO/IEC (2023a) 은 AI 관리 시스템의 국제 표준이다. 다음 데이터 관련 통제를 요구한다.

- AI 시스템의 데이터 요구사항 문서화
- 데이터 품질 프로세스 수립
- 데이터 준비·라벨링의 표준 절차
- 데이터 조작·통합의 통제
- 정기적 데이터 검토 및 감사

### 12.4.4 실무 팁

> 팁 1: '완벽한 표준'의 함정을 피하라
>
> '완벽한 표준 체계를 만든 후 AI를 시작'하는 접근은 거의 항상 실패한다. **최소 표준(Minimum Viable Standard, MVS)**으로 출발하고 실전에서 부족한 부분을 채워가라. Databricks (2024b) 의 사례에 따르면, 초기 3개월간 8개 핵심 필드로 시작한 데이터셋 카드가 18개월 후 자연스럽게 30개 필드로 확장되었다. 조기 성공 경험 없이는 어떤 표준도 정착하지 않는다.

> 팁 2: 데이터 과학자를 파트너로
>
> 표준화가 데이터 과학자의 실험 속도를 늦춘다는 인식이 생기면 어떤 표준도 정착하지 못한다. 표준 도구가 '추가 부담'이 아니라 '실험을 빠르게 하는 인프라'가 되도록 설계하라. 예를 들어 피처 스토어를 통해 이미 검증된 피처를 재사용하면 새 모델의 실험 시간이 70% 단축된다는 Uber의 보고가 있다. 데이터 과학자가 자발적으로 사용하는 도구가 성공한다.

> 팁 3: 자동화 가능한 검증을 우선하라
>
> 사람이 매번 확인해야 하는 표준은 오래 가지 못한다. 데이터 품질, 피처 통계, 편향성 측정을 **CI/CD 파이프라인에 통합**하라. Great Expectations, Evidently AI, Fairlearn 등 오픈소스 도구가 자동 검증을 지원한다 (Huyen 2022). 모델 배포 전 자동 게이트를 통과하지 못하면 배포를 차단하는 방식이 가장 효과적이다.

> 팁 4: 편향 감지는 조기에, 자주
>
> 편향은 배포 후 발견되면 큰 비용이 든다. 데이터 수집 단계부터 편향을 감지하고, 모델 학습 후 세그먼트별 성능을 반드시 측정하라. 특히 보호 속성(성별, 인종, 연령 등)에 대한 성능 차이가 임계값을 넘으면 배포를 보류해야 한다. 배포 후에도 정기적(월 1회 이상) 편향 감사를 수행하라.

> ⚠ 주의: 학습 데이터의 개인정보 처리
>
> AI 학습 데이터는 개인정보보호법, GDPR 등의 적용 대상이다. 특히 '프로파일링' 또는 '자동화된 의사결정'에 사용되는 학습 데이터는 더 엄격한 통제가 필요하다. 비식별화, 합성 데이터, 차분 프라이버시(Differential Privacy), 연합 학습(Federated Learning) 등을 표준 옵션으로 검토하라 (Dwork and Roth 2014). 특히 최근 확산되는 대규모 언어 모델(LLM) 학습에서는 학습 데이터의 저작권과 개인정보 이슈가 심각한 법적 리스크로 부상하고 있다.

> ⚠ 주의: 생성형 AI의 데이터 오염 리스크
>
> 2023년 이후 급속히 확산된 생성형 AI는 새로운 데이터 표준화 이슈를 만들고 있다. AI가 생성한 데이터가 다시 학습 데이터로 사용되는 '모델 붕괴(Model Collapse)' 현상이 연구되고 있으며, 이는 데이터 표준화의 새로운 도전이다. 학습 데이터셋에서 AI 생성 콘텐츠를 식별·태깅하는 표준이 필요하다.

## 12.5 사례 연구

### 12.5.1 사례 1: 글로벌 금융사 K사의 신용평가 모델 표준화

#### 12.5.1.1 배경

글로벌 금융사 K사는 25개국에서 신용평가 AI 모델을 운영하고 있었다. 각국 자회사가 독립적으로 모델을 개발했기에 '소득' 같은 동일 개념이 국가별로 다르게 계산되고, 라벨 정의도 상이했다. 그룹 차원의 리스크 관리와 규제 대응이 어려웠다 (McKinsey Analytics 2023).

#### 12.5.1.2 도전 과제

- 25개국 자회사의 독자적 피처 정의

- 데이터 주권 규제로 인한 중앙집중 저장 불가
- 바젤 규제 대응을 위한 모델 감사 요구
- 문화적 차이로 인한 편향 위험

#### 12.5.1.3 접근 방법

1. 글로벌 피처 사전 구축 — 핵심 신용 피처 80개 정의
2. Federated Feature Store — 데이터는 각국 유지, 정의·로직만 중앙화
3. Phased Migration — 신규 모델부터 표준 적용, 기존 모델은 점진 전환
4. Model Card 의무화 — 모든 배포 모델에 필수 문서화
5. 편향 감사 위원회 — 분기별 모델 편향 검토

#### 12.5.1.4 성과 (18개월 후)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 피처 재사용률 | 12% | 67% |
| 모델 개발 기간 (평균) | 94일 | 41일 |
| 규제 감사 소요 시간 | 3주 | 3일 |
| 국가 간 성능 편차 | 최대 23%p | 최대 8%p |
| 세그먼트 편향 지표 | 미측정 | 표준화·모니터링 |

#### 12.5.1.5 교훈

표준화는 '중앙집권화'가 아니라 '공통 언어의 정착'이다. 데이터를 한 곳에 모으지 않고도 표준 정의와 변환 로직만 공유하면 글로벌 일관성을 확보할 수 있다. 이는 데이터 주권 규제가 강화되는 시대에 중요한 전략이다. 특히 Federated Feature Store 접근은 규제와 표준화의 균형점을 찾는 실용적 해법이다.

### 12.5.2 사례 2: 의료 AI 스타트업의 편향 대응

#### 12.5.2.1 배경

미국의 한 의료 AI 스타트업은 흉부 X-ray 영상 진단 모델을 개발했다. 초기 모델은 학술 벤치마크에서 우수한 성능을 보였으나, 실제 병원 배포 후 특정 인종·연령 그룹에서 성능이 현저히 저하되는 문제가 발견되었다 (Obermeyer et al. 2019).

#### 12.5.2.2 문제 진단

- 학습 데이터의 90%가 3개 대형 병원 데이터
- 아시안·히스패닉 환자 데이터 비율 낮음
- 고령층 데이터 부족
- 의료기기 브랜드에 따른 이미지 특성 차이 미고려

#### 12.5.2.3 대응 방법

1. Dataset Card 도입 — 인구통계·기기 분포 명시적 기록
2. 데이터 다양성 확장 — 10개 병원으로 데이터 파트너십 확장
3. Fairlearn 도구로 세그먼트별 성능 상시 측정
4. Model Card에 성능 편차 명시적 공개
5. 배포 병원에 성능 프로필 사전 제공

#### 12.5.2.4 성과

- 세그먼트 간 민감도 편차 18%p → 4%p
- FDA 승인 획득 (SaMD Class II)
- 고객 병원 만족도 향상

#### 12.5.2.5 교훈

의료 AI는 편향이 곧 환자 안전 리스크가 된다. 학술 벤치마크의 우수 성능이 임상 성능을 보장하지 않는다. 데이터 다양성 확보, 세그먼트별 성능 측정, 투명한 성능 공개가 규제 승인과 임상 정착의 핵심이다. 특히 Model Card와 Dataset Card 같은 표준 문서가 규제기관 소통에 결정적 역할을 했다.

### 12.5.3 사례 3: 국내 대기업의 LLM 도입과 데이터 표준화

#### 12.5.3.1 배경

국내 대기업 S사는 2023년 사내 업무용 LLM 도입을 추진했다. ChatGPT와 유사한 챗봇을 내부 문서로 학습시켜 임직원의 업무 지원에 활용하는 것이 목표였다. 그러나 초기 파일럿에서 예상치 못한 문제들이 발견되었다 (한국지능정보사회진흥원 2024b).

#### 12.5.3.2 발견된 문제들

- 사내 문서의 최신성이 일관되지 않아 오래된 정보 답변
- 동일 개념이 부서마다 다른 용어로 사용됨
- 기밀 문서와 공개 문서의 분류 부재
- 개인정보가 포함된 문서 학습에 대한 우려

#### 12.5.3.3 데이터 표준화 대응

1. 문서 메타데이터 표준화 — 최종 수정일, 유효 기간, 담당자 필드 필수화
2. 사내 용어 사전 구축 — RAG의 컨텍스트로 활용
3. 문서 민감도 4단계 분류 — 공개/내부/기밀/극비 자동 분류
4. 개인정보 자동 감지·마스킹 — 학습 전 전처리 표준화
5. 답변에 출처 문서 자동 첨부 — 검증 가능성 확보

#### 12.5.3.4 교훈

LLM 시대에도 전통적 데이터 표준화 원칙은 유효하며 오히려 더 중요해진다. 메타데이터, 용어 표준, 민감도 분류 같은 기본기가 없으면 최신 AI 기술도 무용지물이다. RAG(Retrieval-Augmented Generation) 접근이 확산되면서 '학습 데이터 표준화'와 '참조 데이터 표준화'의 경계가 흐려지고 있다. 조직의 문서·지식 자산 전체를 표준화 대상으로 인식해야 한다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **AI 시대의 데이터 역할 확장** — 데이터는 '기록'에서 '학습 재료'로 역할이 확장되었다. 데이터 문제는 단순한 보고서 오류가 아니라 자동화된 잘못된 의사결정으로 이어진다.
2. **다층적 표준화 대상** — 표준화 대상이 컬럼 단위에서 피처·데이터셋·모델 단위로 다층화된다. 각 레이어마다 고유한 표준 문서(피처 정의서, Dataset Card, Model Card)가 필요하다.
3. **새로운 품질 차원** — 편향성과 공정성이 전통적 6대 품질 차원에 추가되는 새로운 차원이다. 여러 공정성 정의는 상호 배타적이므로 조직 가치에 따른 선택이 필요하다.
4. **새 인프라의 등장** — 피처 스토어, 데이터 버전 관리, 실험 추적이 새 표준 인프라이다. Feast, DVC, MLflow 등이 오픈소스 대표 도구다.

5. **데이터 드리프트 대응** — 모델은 시간이 흐르면 성능이 저하된다. PSI, KS-Statistic 등을 통한 지속적 모니터링과 자동 재학습 체계가 필수적이다.
6. **규제와 표준의 부상** — European Parliament (2024), ISO/IEC (2023a), NIST (2023) 가 AI 데이터의 표준·품질·거버넌스를 명시적으로 요구한다. 이는 선택이 아닌 의무가 되고 있다.
7. **조직 협업의 확장** — AI 표준화는 기술만의 문제가 아니라 데이터·ML·법무·윤리 팀의 협업이 필수적이다. AI 거버넌스 위원회 같은 교차 조직이 필요하다.

## 다음 학습 가이드

### 12.5.4 규제 및 표준

- European Parliament (2024) — 유럽 인공지능 규제법
- ISO/IEC (2023a) — AI 관리 시스템 표준
- NIST (2023) — NIST AI Risk Management Framework

### 12.5.5 핵심 학술 문헌

- Mitchell et al. (2019) — Model Cards for Model Reporting
- Gebru et al. (2021) — Datasheets for Datasets
- Sambasivan et al. (2021) — Data Cascades in High-Stakes AI
- Kleinberg et al. (2017) — Fair ML의 불가능성 정리

### 12.5.6 실무 도서

- Huyen (2022) — Designing Machine Learning Systems
- Databricks (2024b) — The Big Book of MLOps

### 12.5.7 관련 본서 참조 장

- 제 5장 메타데이터 관리 — Model Card, Dataset Card 기반
- 제 6장 데이터 품질 관리 — AI 시대 품질 차원 확장
- 제 9장 데이터 거버넌스 체계 — AI 거버넌스 접점
- 제 13장 클라우드 환경의 표준화 — MLOps 인프라
- 제 14장 국제 표준 심화 — ISO/IEC (2023a), European Parliament (2024)

## 연습 문제

1. 여러분 조직에서 최근 개발한 AI 모델을 하나 선택하여, Model Card와 Dataset Card 초안을 작성하시오.
2. 여러분 조직의 학습 데이터에서 잠재적 편향 유형 3가지를 식별하고, 각각의 완화 전략을 제시하시오.
3. AI-Ready 표준화 프레임워크 6레이어 중 여러분 조직에서 가장 취약한 레이어는 무엇이며, 어떻게 강화할 것인가?

CHAPTER 13

# 클라우드 환경의 표준화

## 13.1 개요

클라우드 컴퓨팅의 부상은 데이터 관리의 모든 측면을 근본적으로 재편했다. AWS가 2006년 S3를 출시한 이후 20년이 지난 지금, **엔터프라이즈 데이터의 80% 이상이 클라우드에 저장·처리**되고 있으며, 이 비율은 계속 증가하고 있다 (IDC 2024b). 데이터 표준화의 관점에서 클라우드는 새로운 기회이자 새로운 도전이다.

전통적 온프레미스 환경에서는 조직이 직접 통제하는 단일 데이터 센터에서 표준을 적용하면 충분했다. 클라우드 환경에서는 **여러 클라우드 서비스 제공자(CSP), 여러 리전, 여러 데이터 형식, 여러 SaaS 애플리케이션**이 얽혀 있어 표준화의 범위와 복잡성이 급격히 증가한다. Reis and Housley (2022) 은 Fundamentals of Data Engineering 에서 이를 '데이터 관리의 3차원 확장'이라 표현한다 — 볼륨, 다양성, 그리고 지리적 분산이 동시에 확대되는 것이다.

특히 2020년대 이후 **멀티 클라우드 및 하이브리드 클라우드**가 표준이 되면서 데이터 표준화는 새로운 국면에 접어들었다. Gartner의 조사에 따르면 **엔터프라이즈의 89%가 이미 멀티 클라우드 전략을 채택**하고 있으며, 이들 조직은 평균 2.6개의 서로 다른 클라우드를 사용한다 (AWS 2024). 각 클라우드는 자체 데이터 서비스, 메타데이터 형식, 보안 모델을 가지고 있어, 이들 간의 상호운용성을 확보하는 것이 새로운 표준화 과제가 되었다.

### 13.1.1 클라우드 표준화의 3대 도전

클라우드 환경의 데이터 표준화는 다음 세 가지 근본적 도전에 직면한다.

1. **벤더 이질성(Vendor Heterogeneity)** — 각 CSP의 서비스·API·데이터 형식 차이
2. **데이터 주권과 규제(Data Sovereignty)** — 국가·리전별 규제 준수 요구
3. **동적 확장성(Dynamic Scaling)** — 자원의 유동성으로 인한 관리 대상 변화

### 13.1.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 클라우드 데이터 아키텍처의 진화(웨어하우스→레이크→레이크하우스→메시)를 이해한다
- 멀티 클라우드 환경의 데이터 표준화 전략을 수립할 수 있다
- 데이터 레이크하우스와 오픈 테이블 포맷의 역할을 설명할 수 있다
- 데이터 주권과 클라우드 데이터 보호 규제를 이해한다
- 클라우드 네이티브 메타데이터 및 계보(Lineage) 도구를 활용할 수 있다
- FinOps 관점에서 클라우드 데이터 비용을 관리할 수 있다

## 13.2 핵심 개념

### 13.2.1 클라우드 데이터 아키텍처의 진화

지난 40년간 데이터 아키텍처는 네 세대에 걸쳐 진화해왔다. 각 세대는 이전 세대의 한계를 극복하기 위해 등장했으며, 현대 조직은 종종 여러 세대의 아키텍처를 동시에 운영한다 (Reis and Housley 2022; Dehghani 2022).

| 세대 | 등장 시기 | 특징 | 대표 기술 |
|---|---|---|---|
| 1세대: 데이터 웨어하우스 | 1980~2000년대 | 구조화 데이터, ETL 중심, 스키마 온 라이트 | Teradata, Oracle, Netezza |
| 2세대: 데이터 레이크 | 2010년대 초 | 비정형·반정형 지원, 스키마 온 리드, 저비용 | Hadoop, S3, HDFS |
| 3세대: 데이터 레이크하우스 | 2020년대 초 | 레이크 저장소 + 웨어하우스 성능, ACID 트랜잭션 | Databricks, Snowflake, Iceberg, Delta |
| 4세대: 데이터 메시 | 2020년대 중 | 도메인 중심 분산 데이터 오너십, Data as Product | Data Mesh 원칙 + 위 기술 조합 |

### 13.2.2 데이터 레이크하우스(Data Lakehouse)

Armbrust et al. (2021) 이 2021년 CIDR 논문에서 제시한 **레이크하우스** 개념은 데이터 웨어하우스와 데이터 레이크의 장점을 결합하는 새로운 아키텍처 패러다임이다. 저비용 오브젝트 스토리지(S3, Azure Blob) 위에 **ACID 트랜잭션, 스키마 관리, 성능 최적화** 기능을 오픈 테이블 포맷으로 구현한다.

> **i** 레이크하우스의 핵심 원리
>
> 레이크하우스가 표준화 관점에서 중요한 이유는 **오픈 테이블 포맷(Open Table Format)**의 등장이다. Apache Iceberg, Delta Lake, Apache Hudi 등은 오브젝트 스토리지 위에서 관계형 데이터베이스와 유사한 기능을 제공하면서도, 특정 벤더에 종속되지 않는다. Inmon and Srivastava (2020) 은 이를 '데이터 아키텍처의 대통일 이론'이라 부른다.

#### 13.2.2.1 3대 오픈 테이블 포맷 비교

| 포맷 | 창시 | 거버넌스 | 강점 | 주요 사용 |
|---|---|---|---|---|
| Apache Iceberg | Netflix (2017) | Apache 재단 | 숨겨진 파티셔닝, 스키마 진화, 시간 여행 | Netflix, Apple, LinkedIn, Snowflake |
| Delta Lake | Databricks (2019) | Linux Foundation (Delta) | 성능 최적화, Databricks 통합, Z-order | Databricks, Microsoft Fabric |

| 포맷 | 창시 | 거버넌스 | 강점 | 주요 사용 |
|---|---|---|---|---|
| Apache Hudi | Uber (2016) | Apache 재단 | 실시간 upsert, 스트리밍 우선 | Uber, Robinhood, Amazon |

Netflix의 Ryan Blue가 2018년 발표한 Apache Iceberg는 특히 **숨겨진 파티셔닝(hidden partitioning)**과 **스키마 진화** 측면에서 혁신적이었다 (Blue 2018). 사용자가 파티션 컬럼을 명시적으로 관리할 필요 없이 시스템이 자동으로 처리해준다.

### 13.2.3 Data Mesh: 도메인 중심 분산 데이터

Dehghani (2022) 가 제안한 Data Mesh는 중앙집중형 데이터 팀의 한계를 극복하기 위한 새로운 조직·기술 패러다임이다. 4대 원칙에 기반한다.

1. **도메인 중심 데이터 오너십(Domain-Oriented Ownership)** — 데이터를 생성하는 도메인이 관리 책임
2. **Data as a Product** — 데이터를 소비자를 위한 제품으로 취급
3. **셀프 서비스 데이터 인프라(Self-Serve Data Infrastructure)** — 도메인 팀이 자율적으로 활용
4. **연합 컴퓨팅 거버넌스(Federated Computational Governance)** — 중앙 표준 + 도메인 자율

> Data Mesh vs 중앙집중형
>
> Data Mesh가 모든 조직에 적합하지는 않다. Dehghani 자신도 "스타트업이나 소규모 조직에는 중앙집중형이 여전히 최적"이라고 인정한다. 조직 규모, 도메인 다양성, 기술 성숙도가 Data Mesh 도입 결정의 핵심 기준이다.

### 13.2.4 데이터 패브릭(Data Fabric)

Gartner가 제창한 Data Fabric은 분산된 데이터 자산을 지능적으로 통합하는 아키텍처 접근이다 (Gartner 2023a). Data Mesh가 조직적 접근이라면, Data Fabric은 기술적 접근이다. 두 개념은 상호 배타적이 아니라 보완적으로 활용될 수 있다.

| 구분 | Data Mesh | Data Fabric |
|---|---|---|
| 주안점 | 조직·프로세스 | 기술·아키텍처 |
| 핵심 아이디어 | 도메인 중심 분산 오너십 | 지능적 메타데이터 기반 통합 |
| 실행 방식 | 조직 재편, 문화 변화 | AI/ML 기반 자동화 도구 |
| 적합 조직 | 대규모 다도메인 조직 | 이질적 시스템 통합 필요 |

### 13.2.5 오픈 표준의 부상

클라우드 환경에서 벤더 락인을 피하려면 오픈 표준을 우선해야 한다. 최근 5년간 등장한 주요 오픈 표준들은 다음과 같다.

| 영역 | 표준 | 거버넌스 |
|---|---|---|
| 테이블 포맷 | Apache Iceberg, Delta Lake | Apache/Linux Foundation |
| 데이터 계보 | OpenLineage | LF AI & Data |
| 메타데이터 | Apache Atlas, OpenMetadata | Apache Software Foundation |

| 영역 | 표준 | 거버넌스 |
|---|---|---|
| 카탈로그 API | Apache Polaris, Unity Catalog | Apache Foundation |
| 스트리밍 | Apache Kafka, Apache Pulsar | Apache Foundation |
| 쿼리 언어 | SQL, DuckDB, Trino | ISO/ANSI, 오픈소스 |

## 13.3 방법론

### 13.3.1 멀티 클라우드 표준화 전략

멀티 클라우드 환경에서 데이터 표준화는 다음 5개 레이어에서 이루어진다 (AWS 2024).

| 레이어 | 표준화 대상 | 대표 기술 |
|---|---|---|
| L1. 데이터 저장 형식 | 파일 형식, 테이블 포맷 | Parquet, Iceberg, Delta |
| L2. 메타데이터·카탈로그 | 메타모델, 스키마 | Unity Catalog, DataHub, Atlan |
| L3. 데이터 계보·거버넌스 | 계보 표준, 정책 | OpenLineage, Marquez |
| L4. 접근·보안 | IAM, 암호화, 감사 | SAML, OAuth, KMS 표준 |
| L5. 오케스트레이션·통합 | 워크플로우, 데이터 계약 | Airflow, dbt, Data Contracts |

### 13.3.2 데이터 계보의 표준화: OpenLineage

멀티 클라우드 환경에서 데이터 계보 추적은 매우 복잡하다. 각 도구가 자체 계보 형식을 사용하면 통합이 불가능하다. LF AI & Data Foundation (2024) 은 이 문제를 해결하는 오픈 표준이다.

#### 13.3.2.1 OpenLineage의 핵심 구성 요소

- Job: 데이터를 처리하는 작업 (예: SQL 쿼리, Spark job)
- Run: Job의 특정 실행 인스턴스
- Dataset: 입력·출력 데이터 자산
- Facet: 확장 가능한 메타데이터 (스키마, 통계, 품질 등)

주요 벤더가 OpenLineage를 채택하고 있다: Airflow, dbt, Great Expectations, Spark, Flink 등. 이는 도구 간 계보 정보의 상호운용성을 보장한다.

### 13.3.3 데이터 계약(Data Contracts)

분산 환경에서 데이터 생산자와 소비자 간의 명시적 약속인 **데이터 계약**이 새로운 표준화 요소로 부상하고 있다. 마이크로서비스의 API 계약과 유사한 개념이다.

#### 13.3.3.1 데이터 계약의 표준 구성 요소

1. **스키마** — 필드명, 타입, 제약조건
2. **의미적 정의** — 각 필드의 비즈니스 의미
3. **SLA** — 지연 시간, 신선도, 가용성
4. **품질 기준** — 완전성, 정확성, 무결성 임계값
5. **호환성 정책** — 스키마 진화 규칙
6. **책임과 오너십** — 생산자·소비자 정보

### 13.3.4 데이터 주권과 규제

클라우드 환경의 데이터 표준화에서 가장 복잡한 이슈는 **데이터 주권(Data Sovereignty)**이다. 데이터가 어느 국가·리전에 저장·처리되는지에 따라 적용되는 법률이 달라진다.

| 규제 | 핵심 요구 | 영향 |
|---|---|---|
| GDPR (EU) | EU 외부 이전 시 적정성 보장 | EU 리전 저장 필수인 경우 다수 |
| CCPA (미국 캘리포니아) | 판매 옵트아웃, 삭제 권한 | 미국 서부 리전 특별 관리 |
| 개인정보보호법 (한국) | 국외 이전 시 동의·표준계약조항 | 한국 리전 활용 확대 |
| PIPL (중국) | 중요 정보의 국외 반출 통제 | 중국 리전 강제 사용 |
| Schrems II (EU-US 판결) | EU-US 개인정보 이전의 적정성 | EU-US 데이터 프라이버시 프레임워크 |

European Data Protection Board (2020b) 의 2020년 권고사항은 특히 EU-미국 간 데이터 이전에 대한 상세한 지침을 제공한다. 클라우드 사용 조직은 이 권고를 정기적으로 검토해야 한다.

## 13.4 실무 적용

### 13.4.1 멀티 클라우드 도구 스택

멀티 클라우드 환경에서 데이터 표준화를 지원하는 도구는 다음과 같이 분류된다.

| 카테고리 | 주요 도구 | 특징 |
|---|---|---|
| 통합 카탈로그 | Atlan, Collibra, DataHub, Alation | 멀티 클라우드 자산 통합 뷰 |
| 쿼리 페더레이션 | Trino, Starburst, Dremio | 쿼리로 여러 소스 조인 |
| 계보 추적 | OpenLineage, Marquez | 표준 프로토콜 기반 추적 |
| 데이터 이동 | Fivetran, Airbyte, Matillion | SaaS·CSP 간 데이터 파이프라인 |
| 정책 관리 | Immuta, Privacera, Okera | 중앙 정책·CSP별 시행 |
| 관측성 | Monte Carlo, Bigeye, Databand | SLA·품질·비용 통합 모니터링 |

### 13.4.2 FinOps: 클라우드 데이터 비용 관리

클라우드 데이터 관리에서 **FinOps(Financial Operations)**는 새로운 표준화 영역이다. FinOps Foundation (2024) 프레임워크는 다음 6개 원칙을 제시한다.

1. 팀 협업 (Engineering·Finance·Product)
2. 클라우드 비용의 비즈니스 가치 인식
3. 실시간 결정 지원
4. 중앙 팀이 FinOps 활동 주도
5. 사용량 기반 비용 예측
6. 클라우드 변동 비용 모델의 활용

#### 13.4.2.1 데이터 관련 FinOps 실무

- **파티션 프루닝(Partition Pruning)** — 불필요한 데이터 스캔 방지
- **적절한 스토리지 클래스** — Hot/Warm/Cold 계층화
- **쿼리 최적화** — 열 지향 형식(Parquet), 압축

- **컴퓨트 최적화** — Spot 인스턴스, Auto-scaling
- **데이터 계보 기반 절감** — 사용되지 않는 데이터셋 식별·폐기

### 13.4.3 실무 팁

> 팁 1: 오픈 테이블 포맷을 표준으로 채택하라
>
> Apache Iceberg나 Delta Lake 같은 오픈 테이블 포맷을 조직 표준으로 지정하라. 이는 (1) 특정 벤더 종속을 피할 수 있고, (2) 여러 컴퓨트 엔진(Spark, Trino, Flink, DuckDB)이 동일 데이터에 접근 가능하며, (3) 향후 새로운 도구 도입 시 마이그레이션 부담이 크게 줄어든다. Snowflake, Databricks, AWS, Google 모두 Iceberg를 지원하는 추세이므로 (Blue 2018), 이는 안전한 선택이 되고 있다.

> 팁 2: 하나의 카탈로그, 여러 클라우드
>
> 멀티 클라우드에서 여러 카탈로그를 운영하는 것은 재앙이다. **단일 통합 카탈로그**로 모든 클라우드의 데이터 자산을 관리하라. Atlan, DataHub, Collibra 등이 이런 통합을 지원한다. 사용자는 데이터가 어느 클라우드에 있는지 알 필요 없이 통합 검색·거버넌스를 활용할 수 있다.

> 팁 3: 데이터 계약을 코드로 관리하라
>
> 데이터 계약은 문서가 아니라 **코드로 관리**되어야 한다. YAML 또는 JSON으로 계약을 정의하고 Git으로 버전 관리하며, CI/CD 파이프라인에서 자동 검증하라. Data Contract Specification, PayPal의 오픈소스 프로젝트 등이 표준적 접근을 제공한다. 계약 위반 시 배포를 차단하는 게이트가 효과적이다.

> 팁 4: 데이터 주권을 아키텍처로 통제하라
>
> GDPR·PIPL 등 규제는 정책·문서만으로 준수할 수 없다. **아키텍처 자체에 통제 로직**을 내장하라. 예: EU 시민 데이터는 EU 리전만 접근 가능하도록 IAM 정책, 데이터 태깅으로 자동 라우팅, 리전 간 복제 금지 규칙. 이런 통제는 실수와 규제 위반을 방지한다.

> ⚠ 주의: '클라우드 = 표준화 문제 해결' 착각
>
> 클라우드 마이그레이션 자체가 데이터 표준화 문제를 해결해주지는 않는다. 오히려 새로운 표준화 도전을 야기할 수 있다. 온프레미스의 사일로가 클라우드로 그대로 옮겨가면 '클라우드 사일로'가 될 뿐이다. 마이그레이션 전에 표준화 전략을 먼저 수립해야 한다.

> ⚠ 주의: 벤더 종속 클라우드 서비스의 함정
>
> AWS Glue Catalog, Azure Purview, GCP Data Catalog 등 CSP 전용 서비스는 편리하지만 벤더 락인을 강화한다. 특히 카탈로그와 거버넌스는 조직의 데이터 주권과 직결되므로, 벤더 중립적인 도구(OpenMetadata, DataHub, Collibra 등)를 고려하라. 초기 편의성보다 장기적 유연성이 더 중요하다.

> ⚠ 주의: 클라우드 비용의 급격한 확산
>
> 클라우드 데이터 비용은 통제하지 않으면 급격히 확산된다. 특히 (1) 반복 스캔되는 대용량 쿼리, (2) 여러 리전 간 데이터 이동, (3) 미사용 데이터의 무기한 보관, (4) 잘못된 스토리지 계층 선택이 흔한 문제다. FinOps 실무를 초기에 도입하고 CDO·CFO 협업 체계를 구축하라 (FinOps Foundation 2024).

## 13.5 사례 연구

### 13.5.1 사례 1: 글로벌 금융사 A의 멀티 클라우드 데이터 플랫폼

#### 13.5.1.1 배경

글로벌 금융사 A는 규제·리스크 관리로는 AWS를, 데이터 분석·AI로는 Azure를, 개발·실험으로는 GCP를 사용하는 3-클라우드 전략을 채택했다. 각 클라우드가 자체 데이터 서비스를 가져 통합 뷰가 불가능했고, 규제 보고에 심각한 지연이 발생했다 (McKinsey Digital 2023).

#### 13.5.1.2 도전 과제

- 3개 클라우드의 15개 데이터 서비스 사용
- 데이터 계보가 클라우드 경계에서 단절
- 리전 간 데이터 이동으로 인한 규제 리스크
- 연간 클라우드 비용 급증 (전년 대비 47%)

#### 13.5.1.3 접근 방법

1. **Apache Iceberg 표준 채택** — 모든 클라우드에서 동일 형식
2. **Atlan 통합 카탈로그** — 3-클라우드 자산 단일 뷰
3. **OpenLineage 계보 표준** — 클라우드 경계 넘는 추적
4. **데이터 계약** — 도메인 간 명시적 SLA
5. **FinOps 팀 구성** — 재무·엔지니어링 협업

#### 13.5.1.4 성과 (18개월)

| 지표 | 도입 전 | 도입 후 |
|---|---|---|
| 규제 보고 소요 시간 | 2주 | 2일 |
| 데이터 자산 검색 성공률 | 43% | 89% |
| 클라우드 데이터 비용 | 기준 100 | 68 |
| 데이터 이슈 평균 해결 시간 | 3일 | 6시간 |
| 계보 커버리지 | 35% | 92% |

#### 13.5.1.5 교훈

멀티 클라우드는 필연적으로 복잡성을 증가시킨다. 그러나 오픈 표준(Iceberg, OpenLineage)과 벤더 중립적 도구(Atlan)를 활용하면 단일 클라우드 수준의 통합 경험을 제공할 수 있다. 특히 데이터 계약과 FinOps는 멀티 클라우드 성공의 필수 요소이다.

## 13.5.2 사례 2: 한국 대형 소매 B의 레이크하우스 전환

### 13.5.2.1 배경

한국 대형 소매 기업 B는 10년간 사용해온 온프레미스 데이터 웨어하우스(Teradata)의 라이선스 비용과 확장성 한계에 직면했다. 온라인 채널의 급성장으로 데이터가 10배 증가했으나, 기존 시스템으로는 대응이 불가능했다. 2022년 데이터 레이크하우스 전환을 결정했다 (한국데이터산업진흥원 2023a).

### 13.5.2.2 도전 과제

- 온프레미스 → 클라우드 대규모 마이그레이션
- 기존 SQL 자산(10,000+ 쿼리) 재활용 필요
- 데이터 사이언스팀의 Python·Spark 요구
- BI 팀의 SQL·Tableau 요구
- 개인정보 보호법 준수

### 13.5.2.3 접근 방법

1. Databricks + Delta Lake — 통합 레이크하우스 플랫폼
2. SQL 워크로드 우선 마이그레이션 — Databricks SQL 활용
3. Unity Catalog — 데이터·AI 자산 통합 거버넌스
4. 개인정보 태깅 — 자동 마스킹·비식별 처리
5. 점진적 전환 — 부서별 6개월 간격

### 13.5.2.4 성과 (2년)

- 데이터 처리 성능 5~10배 향상
- 라이선스·인프라 비용 45% 절감
- 데이터 사이언스팀 실험 시간 65% 단축
- BI 리포트 생성 시간 78% 단축
- 개인정보 감사 이슈 0건

### 13.5.2.5 교훈

레이크하우스 전환은 단순한 도구 교체가 아니라 데이터 조직 전체의 운영 방식 변화이다. B사는 점진적 전환과 SQL 우선 접근으로 학습 곡선을 완화했다. 특히 Unity Catalog 같은 통합 거버넌스가 기술적 전환과 함께 문화적 정착을 이끌었다. 오픈 형식(Delta Lake, Iceberg)을 선택한 것이 장기적 유연성을 확보하는 결정이었다.

## 13.5.3 사례 3: 스트리밍 기업 C의 실시간 데이터 표준화

### 13.5.3.1 배경

미국의 한 대형 스트리밍 서비스 기업 C는 초당 수백만 이벤트를 처리한다. 배치 데이터 파이프라인만으로는 사용자 개인화·이상 탐지·콘텐츠 추천 요구를 충족할 수 없어, 스트리밍 우선 아키텍처로 재편했다 (Confluent 2024).

### 13.5.3.2 접근 방법

1. Apache Kafka — 중앙 이벤트 백본
2. Schema Registry — 스트리밍 스키마 표준화
3. Apache Iceberg — 스트리밍-배치 통합 저장소
4. Flink — 실시간 스트림 처리
5. 데이터 계약 — 팀 간 이벤트 스키마 계약

#### 13.5.3.3 Schema Registry의 역할

스트리밍 데이터에서 스키마 진화는 특히 어렵다. 생산자가 스키마를 변경하면 수많은 소비자에게 영향이 간다. Schema Registry는 다음을 표준화한다:

- 스키마 버전 관리
- 후방 호환성(backward compatibility) 검증
- 스키마 진화 정책 (add-only, breaking change)
- 스키마 검색 및 문서화

#### 13.5.3.4 교훈

실시간 데이터는 배치와 다른 표준화 요구를 가진다. 특히 스키마 관리, 이벤트 순서, 정확히 한 번 처리(exactly-once) 등이 새로운 표준화 영역이다. Schema Registry, 스트리밍 계약, Iceberg 같은 통합 저장소가 실시간·배치 통합의 열쇠이다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **클라우드 데이터의 지배적 위치** — 엔터프라이즈 데이터의 80% 이상이 클라우드로 이동했으며, 멀티 클라우드가 표준이 되었다. 데이터 표준화는 이 환경의 복잡성에 맞춰 진화해야 한다.
2. **아키텍처의 진화** — 데이터 웨어하우스에서 레이크, 레이크하우스, 그리고 데이터 메시로 아키텍처가 진화하고 있다. 각 세대는 이전의 한계를 극복하며, 조직은 종종 여러 세대를 병행 운영한다.
3. **오픈 테이블 포맷의 부상** — Apache Iceberg, Delta Lake, Apache Hudi 같은 오픈 테이블 포맷이 클라우드 데이터의 새로운 표준이 되었다. 벤더 락인을 피하는 핵심 전략이다.
4. **Data Mesh와 Data Fabric** — 조직적 접근(Data Mesh)과 기술적 접근(Data Fabric)이 상호 보완적으로 활용된다. 조직 상황에 맞는 조합이 필요하다.
5. **오픈 표준의 확산** — OpenLineage, Schema Registry, 데이터 계약 등이 클라우드 데이터 표준화의 새 요소이다. 이들은 조직·도구 경계를 넘는 상호운용성을 제공한다.
6. **데이터 주권과 규제** — GDPR, PIPL 등 데이터 주권 규제가 클라우드 아키텍처 결정에 직접 영향을 미친다. 정책이 아닌 아키텍처 자체에 통제를 내장해야 한다.
7. **FinOps의 부상** — 클라우드 데이터 비용은 급격히 확산된다. 재무·엔지니어링 협업의 FinOps 실무가 필수 표준화 영역이 되었다.

## 다음 학습 가이드

### 13.5.4 필수 참고 문헌

- Reis and Housley (2022) — Fundamentals of Data Engineering
- Dehghani (2022) — Data Mesh
- Inmon and Srivastava (2020) — Building the Data Lakehouse
- Armbrust et al. (2021) — Lakehouse 원조 논문 (CIDR)

### 13.5.5 산업 보고서·프레임워크

- IDC (2024b) — 멀티 클라우드 시장 전망
- Gartner (2023a) — Data Fabric 기술 동향
- FinOps Foundation (2024) — FinOps Framework
- Databricks (2024a) — Lakehouse 구축 가이드

### 13.5.6 오픈 표준

- LF AI & Data Foundation (2024) — 데이터 계보 표준
- Blue (2018) — Apache Iceberg 소개
- Delta Lake, Apache Hudi 공식 문서

### 13.5.7 관련 본서 참조 장

- 제 5장 메타데이터 관리 — 클라우드 카탈로그
- 제 9장 데이터 거버넌스 체계 — 멀티 클라우드 거버넌스
- 제 10장 표준화 도구와 기술 — 클라우드 네이티브 도구
- 제 12장 AI·ML과 데이터 표준화 — 클라우드 MLOps
- 제 14장 국제 표준 심화 — 데이터 주권 규제

## 연습 문제

1. 여러분 조직의 클라우드 데이터 전략을 진단하고, 4세대 아키텍처 진화 중 어느 단계에 있는지 평가하시오.
2. Data Mesh와 Data Fabric 중 여러분 조직에 적합한 접근은 무엇인가? 그 이유는 무엇인가?
3. 여러분 조직이 멀티 클라우드로 전환한다면 벤더 락인 방지를 위해 어떤 5가지 오픈 표준을 채택할 것인가?

CHAPTER 14

# 국제 표준 심화

## 14.1 개요

3장에서 우리는 주요 표준화 프레임워크를 개괄했다. 본 장은 그 중 특히 실무에서 가장 자주 참조되는 국제 표준들을 **심층적으로 다룬다**. ISO/IEC (2023b) 의 메타데이터 등록소 메타모델, ISO (2022b) 의 데이터 품질 파트별 요구사항, ISO/IEC (2008) 의 15개 품질 특성 세부, ISO/IEC (2023a) 의 AI 관리 시스템 구축 요건 등을 표준 원문의 세밀한 수준까지 파고든다.

국제 표준을 '알고 있다'와 '적용할 수 있다'는 큰 차이가 있다. ISO 국제 표준은 통상 60~200페이지의 정교한 요구사항으로 구성되며, 각 조항은 정확한 해석을 요구한다. 표준 문서는 '모호성을 배제하는 것'이 목적이기에, 종종 일상 언어와는 다른 정밀한 용어로 작성된다. 이 장은 실무자가 표준을 실제 프로젝트에 적용할 때 필요한 심층 지식을 제공한다.

### 14.1.1 국제 표준화 기구의 이해

국제 표준화는 여러 기구의 협력을 통해 이루어진다. 각 기구의 역할과 관계를 이해하는 것이 표준 활용의 첫걸음이다 (ISO Central Secretariat 2024).

| 기구 | 약칭 | 관할 영역 |
|---|---|---|
| International Organization for Standardization | ISO | 산업·기술 전반, 데이터 관리 포함 |
| International Electrotechnical Commission | IEC | 전기·전자·정보기술 |
| ISO/IEC Joint Technical Committee 1 | ISO/IEC JTC 1 | 정보기술 공동 위원회 |
| International Telecommunication Union | ITU | 통신 표준 |
| World Wide Web Consortium | W3C | 웹 표준 |
| Internet Engineering Task Force | IETF | 인터넷 프로토콜 |

| 기구 | 약칭 | 관할 영역 |
|---|---|---|
| Object Management Group | OMG | 모델링·소프트웨어 표준 |

데이터 표준화 실무자가 가장 자주 접하는 것은 ISO/IEC JTC 1이 발간하는 표준들이다. 특히 SC 32(Data management and interchange), SC 40(IT service management and governance), SC 42(Artificial intelligence)의 표준이 데이터 관리에 직결된다 (ISO/IEC JTC 1 2024).

### 14.1.2 국제 표준의 개발 과정

ISO 국제 표준은 다음 6단계를 거쳐 개발된다. 각 단계는 통상 6개월~2년 소요되며, 표준 하나의 완성에는 평균 3~5년이 걸린다.

1. Proposal (제안 단계, NP) — 새로운 표준의 필요성 제안
2. Preparatory (준비 단계, WD) — 작업 초안 개발
3. Committee (위원회 단계, CD) — 위원회 검토·투표
4. Enquiry (조회 단계, DIS) — 회원국 공식 조회
5. Approval (승인 단계, FDIS) — 최종 국제 표준 초안 승인
6. Publication (발간 단계, IS) — 국제 표준 공식 발간

> **i** 표준의 개정 주기
>
> 국제 표준은 발간 후 5년마다 재검토된다. 재검토 결과는 (1) 개정 없이 유지(reconfirmation), (2) 수정(amendment), (3) 전면 개정(revision), (4) 폐기(withdrawal) 중 하나이다. 따라서 표준 인용 시 반드시 발간 연도와 판(edition)을 명시해야 하며, 조직은 채택한 표준의 재검토 주기를 추적해야 한다.

### 14.1.3 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- 주요 국제 표준의 세부 요구사항과 구조를 이해한다
- ISO/IEC (2023b) 메타데이터 등록소를 조직에 적용할 수 있다
- ISO (2022b) 의 파트별 요구사항을 조직 상황에 맞게 활용할 수 있다
- ISO/IEC (2008) 의 15개 품질 특성을 측정·관리할 수 있다
- ISO/IEC (2023a) AI 관리 시스템의 인증 과정을 이해한다
- 국제 표준 간의 관계와 조합 활용법을 안다

## 14.2 ISO/IEC 11179: 메타데이터 등록소 심화

ISO/IEC (2023b) 은 조직 간·시스템 간 데이터 요소의 상호운용성을 확보하기 위한 메타데이터 등록소(Metadata Registry, MDR)의 국제 표준이다. 6개 파트로 구성되며, 1994년 초판 이후 여러 차례 개정되어 현재 최신은 2023년 판이다.

### 14.2.1 메타모델의 핵심 구조

ISO/IEC (2023c) 의 메타모델은 데이터 요소를 정교하게 분해·기술하는 개념 체계이다. 핵심 구성 요소는 다음과 같다.

| 구성 요소 | 역할 |
|---|---|
| Data Element (데이터 요소) | 관리 대상의 기본 단위 |

| 구성 요소 | 역할 |
|---|---|
| Data Element Concept (DEC) | 표현 방식과 독립된 의미 단위 (예: '사람의 나이') |
| Object Class (개체 클래스) | 관심 대상 (예: 사람) |
| Property (속성) | 개체의 특성 (예: 나이) |
| Value Domain (값 영역) | 허용되는 값의 집합 (예: 0~150 정수) |
| Conceptual Domain (개념 영역) | 값의 의미적 범위 (예: 성인 나이 범위) |
| Data Element Representation | 실제 저장·표현 형식 |
| Classification Scheme | 분류 체계 |

! DEC의 중요성

Data Element Concept(DEC)는 ISO/IEC (2023b) 의 핵심 혁신이다. 동일한 개념('사람의 나이')이 여러 시스템에서 다른 형식(정수, 문자열, 코드)으로 표현될 수 있다. DEC를 중심에 두면 표현 형식이 달라도 의미적으로 동일한 데이터임을 인식할 수 있다 (ISO/IEC 2023c). 이는 시맨틱 상호운용성(semantic interoperability)의 기반이다.

### 14.2.2 명명 규칙 (Part 5)

ISO/IEC 11179-5는 데이터 요소의 명명(naming)에 대한 상세한 규칙을 제공한다. 표준화된 이름은 이해와 검색을 크게 개선한다.

#### 14.2.2.1 명명의 4원칙

1. Uniqueness (유일성) — 등록소 내에서 유일한 이름
2. Necessity (필요성) — 각 이름은 목적을 가짐
3. Precision (정확성) — 모호함 없는 정의
4. Concise-Language Effect (간결성) — 필요 이상 길지 않음

#### 14.2.2.2 이름의 구성 요소

ISO/IEC 11179 명명 규칙에 따른 데이터 요소 이름은 다음 3부분으로 구성된다.

- Object Class Term (개체 클래스 용어) — 예: '고객'
- Property Term (속성 용어) — 예: '이름'
- Representation Term (표현 용어) — 예: '문자열'

예시: '고객_이름_문자열(Customer_Name_Text)'. 이 3부분 구조는 이름 자체가 데이터의 의미를 명확히 전달하도록 설계되었다.

### 14.2.3 정의 작성 규칙 (Part 4)

ISO/IEC (2004a) 는 데이터 요소의 정의를 어떻게 작성할지의 규칙을 제공한다. 다음은 좋은 정의의 필수 특성이다.

| 특성 | 설명 | 예시 |
|---|---|---|
| Precise (정확) | 모호함 없이 하나의 해석 | '주소' vs '자연인·법인이 통상적으로 거주·소재하는 물리적 위치' |

| 특성 | 설명 | 예시 |
|---|---|---|
| Concise (간결) | 불필요한 수식 없음 | '서비스의 대상이 되는 인물' vs '고객' |
| Distinct (구별) | 다른 개념과 명확히 구별 | '식별번호' vs '개인식별번호' |
| Unencumbered (독립) | 다른 정의에 의존하지 않음 | '다음과 같은 X' 반복 회피 |
| Positive (긍정) | '~이 아닌' 표현 회피 | '단순 상품이 아닌 것' vs '가공된 상품' |
| Non-circular (순환 없음) | 정의 순환 회피 | '주소는 주소지이다' 회피 |

## 14.3 ISO 8000: 데이터 품질 파트별 심화

ISO (2022b) 은 데이터 품질을 다루는 국제 표준 시리즈로, 30개 이상의 파트로 구성된다. 실무자는 조직 상황에 맞는 파트를 선별적으로 활용해야 한다 (ISO 2022a).

### 14.3.1 파트별 개요

| 파트 번호대 | 주제 | 주요 파트 |
|---|---|---|
| 1~99: General | 일반 개념, 용어, 정보 품질 | 8000-1, 8000-8, 8000-61 |
| 100~199: Master Data | 마스터 데이터 품질 | 8000-110, 8000-115, 8000-116 |
| 200~299: Transaction Data | 트랜잭션 데이터 품질 | 8000-220 (개발 중) |
| 300~399: Product Data | 제품 데이터 품질 | 8000-311 (제조 데이터) |
| 1500~1799: Application Ref | 산업별 적용 참조 | 산업 특화 |

### 14.3.2 ISO 8000-8: 정보·데이터 품질 개념

ISO 8000-8은 정보 품질과 데이터 품질의 개념적 기반을 제공한다. 다음 3대 품질 유형을 정의한다 (ISO 2022a).

1. **Syntactic Quality (구문 품질)** — 형식·구조의 정확성
2. **Semantic Quality (의미 품질)** — 실세계와의 일치성
3. **Pragmatic Quality (실용 품질)** — 사용 목적 대비 적합성

이 3대 유형은 서로 독립적이면서도 관련되어 있다. 예: 주소 데이터의 형식이 완벽해도 (구문 품질 만족) 실제 주소와 다르면(의미 품질 실패) 배송에 활용 불가하다(실용 품질 실패).

### 14.3.3 ISO 8000-61: 데이터 품질 관리 프로세스

ISO (2016) 은 데이터 품질 관리의 프로세스 표준이다. Deming의 PDCA 사이클을 데이터 품질에 적용한 것으로, 조직이 준수해야 할 체계적 프로세스를 규정한다.

#### 14.3.3.1 4대 활동 영역

| 활동 | 설명 | 주요 산출물 |
|---|---|---|
| Data Quality Planning | 품질 목표·기준 수립 | 품질 정책, KPI |
| Data Quality Control | 일상적 품질 관리·검증 | 품질 모니터링 로그 |
| Data Quality Assurance | 품질 보증·인증 | 감사 보고서, 인증서 |
| Data Quality Improvement | 지속적 개선 | 개선 프로젝트, 근본 원인 분석 |

### 14.3.4 ISO 8000-110: 마스터 데이터 품질

ISO 8000-110은 마스터 데이터의 품질 요구사항을 상세히 규정한다. 특히 데이터 교환 시 요구되는 4가지 특성을 정의한다.

- Syntactic Compliance (구문 준수) — 정의된 형식 준수
- Semantic Encoding (의미 인코딩) — 표준 용어·코드 사용
- Data Provenance (데이터 출처) — 출처 정보 명시
- Data Accuracy (데이터 정확성) — 실세계와의 일치

특히 ISO 8000-115와 8000-116은 조직 및 위치 마스터 데이터에 대한 세부 요구를 제공한다. 글로벌 마스터 데이터 관리 시 이 표준의 활용이 상호운용성을 크게 향상시킨다.

## 14.4 ISO/IEC 25012: 데이터 품질 모델 심화

ISO/IEC (2008) 는 ISO/IEC 25000 시리즈(SQuaRE)의 일부로, 데이터 품질을 15개 특성으로 정의한다. 이 모델은 데이터 품질의 '측정 가능한 정의'를 제공한다는 점에서 다른 표준과 구별된다.

### 14.4.1 15개 품질 특성

15개 품질 특성은 데이터가 어떻게 사용되는지의 관점에서 3개 범주로 분류된다.

#### 14.4.1.1 범주 1: 고유 특성 (Inherent) — 5개

데이터 자체의 특성으로, 시스템과 무관하게 존재한다.

| 특성 | 정의 |
|---|---|
| Accuracy (정확성) | 실제 세계와 데이터 표현의 일치도 |
| Completeness (완전성) | 필요한 모든 값이 존재 |
| Consistency (일관성) | 표현이 관련 다른 데이터와 모순 없음 |
| Credibility (신뢰성) | 출처와 내용이 진실하다고 인정됨 |
| Currentness (현재성) | 적절한 시점의 데이터 |

#### 14.4.1.2 범주 2: 고유+시스템 종속 특성 — 7개

데이터 자체와 시스템 환경 모두에 의존하는 특성이다.

| 특성 | 정의 |
|---|---|
| Accessibility (접근성) | 장애를 가진 이용자를 포함해 접근 가능 |
| Compliance (준수성) | 표준·규정 준수 |

| 특성 | 정의 |
|---|---|
| Confidentiality (기밀성) | 권한 있는 사용자만 접근 |
| Efficiency (효율성) | 적절한 성능으로 처리 가능 |
| Precision (정밀성) | 명세된 정확도로 표현 |
| Traceability (추적성) | 데이터의 접근·변경 감사 가능 |
| Understandability (이해성) | 사용자가 해석·활용 가능 |

#### 14.4.1.3 범주 3: 시스템 종속 특성 — 3개

주로 시스템 환경에 의존하는 특성이다.

- **Availability (가용성)** — 필요할 때 접근·이용 가능
- **Portability (이식성)** — 다른 환경으로 이전 가능
- **Recoverability (복구성)** — 장애로부터 복구 가능

### 14.4.2 특성별 측정 지표

ISO/IEC (2015b) 는 25012 특성의 측정 지표를 제공한다. 각 특성별로 정량적 측정이 가능하다는 것이 이 시리즈의 강점이다.

| 특성 | 측정 지표 예시 |
|---|---|
| Accuracy | (정확한 레코드 수 / 전체 레코드 수) × 100 |
| Completeness | (NULL 아닌 셀 수 / 전체 셀 수) × 100 |
| Consistency | (제약 만족 레코드 수 / 전체 레코드 수) × 100 |
| Currentness | (적시성 만족 레코드 수 / 전체 레코드 수) × 100 |
| Availability | (가용 시간 / 요구된 총 시간) × 100 |

> **ISO 8000 vs ISO/IEC 25012 vs DAMA**
>
> 세 프레임워크는 상호 보완적이다:
>
> - ISO (2022b) = 프로세스 관점 (PDCA로 관리)
> - ISO/IEC (2008) = 측정 관점 (15개 특성 정의)
> - DAMA-DMBOK 6대 차원 = 실무 관점 (완전성·정확성·일관성·유효성·유일성·적시성)
>
> **실무 접근**: DAMA 6대 차원으로 시작 → ISO/IEC (2008) 의 정밀 특성으로 확장 → ISO (2016) 프로세스로 운영

## 14.5 ISO/IEC 42001: AI 관리 시스템 심화

2023년 12월 발효된 ISO/IEC (2023a) 은 인공지능 관리 시스템(AI Management System, AIMS)의 국제 표준이다. AI 시대의 조직적 관리 요구를 표준화한 최초의 국제 규격이며, 인증 대상 표준(certifiable standard)이다.

### 14.5.1 구조: PDCA와 High-Level Structure

ISO/IEC (2023a) 은 다른 ISO 관리 시스템 표준(ISO 9001 품질, ISO 27001 보안 등)과 동일한 High-Level Structure(HLS)를 따른다. 이는 여러 관리 시스템 표준을 통합적으로 운영할 수 있게 하는 설계이다 (ISO/IEC 2024).

#### 14.5.1.1 10개 조항 구조

1. Scope (적용 범위)
2. Normative References (인용 표준)
3. Terms and Definitions (용어 정의)
4. Context of the Organization (조직 상황)
5. Leadership (리더십)
6. Planning (계획)
7. Support (지원)
8. Operation (운영)
9. Performance Evaluation (성과 평가)
10. Improvement (개선)

### 14.5.2 Annex A: AI 통제 요구사항

ISO/IEC (2023a) 의 핵심은 Annex A의 39개 통제 항목이다. 이들은 9개 범주로 분류되며, 조직은 각 통제를 구현하거나 그 제외 이유를 문서화해야 한다.

| 범주 | 통제 예시 |
|---|---|
| A.2 AI 정책 | AI 정책 수립·유지, 목표 설정 |
| A.3 AI 조직 | AI 책임자·조직 지정, 역할·책임 정의 |
| A.4 AI 리소스 | 인적·기술적 자원 확보, 예산 확보 |
| A.5 AI 영향 평가 | AI 시스템 영향 분석, 이해관계자 식별 |
| A.6 AI 라이프사이클 | 설계·개발·배포·운영·폐기의 통제 |
| A.7 AI 데이터 | 학습·평가 데이터 관리, 품질·거버넌스 |
| A.8 AI 정보 제공 | 이해관계자에 대한 시스템 정보 제공 |
| A.9 AI 시스템 사용 | 책임 있는 사용, 사용자 교육 |
| A.10 제3자 관계 | 외부 파트너·공급업체 관리 |

### 14.5.3 인증 절차

ISO/IEC (2023a) 은 인증 대상 표준이므로, 조직은 공인 인증 기관을 통해 인증을 획득할 수 있다. 인증 절차는 다음 6단계로 진행된다 (International Accreditation Forum 2024).

1. **갭 분석** — 현행 대비 42001 요구사항 갭 식별
2. **AIMS 구축** — 정책·프로세스·문서화
3. **내부 감사** — 구축된 시스템 자체 평가
4. **경영 검토** — 최고 경영자 참여 정식 검토
5. **Stage 1 심사** — 인증 기관의 문서 심사
6. **Stage 2 심사** — 실효성 심사, 인증 발급

인증 후에는 매년 감시 심사와 3년마다 재인증 심사를 받는다. 준비 기간은 조직 규모에 따라 6~18개월이 소요된다.

> **i** ISO/IEC 42001과 관련 표준
>
> ISO/IEC (2023a) 은 여러 관련 표준과 함께 활용된다:
>
> - ISO/IEC 22989 — AI 개념과 용어
> - ISO/IEC 23053 — ML 프레임워크
> - ISO/IEC 23894 — AI 리스크 관리
> - ISO/IEC TR 24028 — AI 신뢰성
> - ISO/IEC 5259 시리즈 — 학습 데이터 품질

## 14.6 표준 조합과 통합 활용

### 14.6.1 표준의 계층적 관계

국제 표준들은 서로 독립적으로 존재하지만, 실무에서는 계층적으로 조합하여 활용된다. 다음은 데이터 관리 표준의 전형적인 계층 구조이다 (Sebastian-Coleman 2018).

| 계층 | 표준 | 역할 |
|---|---|---|
| 최상위: 전략 | DAMA-DMBOK, DCAM | 전사 관리 프레임워크 |
| 중위: 도메인 표준 | ISO 8000, ISO/IEC 42001 | 도메인별 관리 표준 |
| 하위: 기술 표준 | ISO/IEC 11179, ISO/IEC 25012 | 구체적 기술·측정 표준 |
| 산업 표준 | BCBS 239, FHIR, ISO 20022 | 산업 특화 요구사항 |

### 14.6.2 통합 활용 시나리오

#### 14.6.2.1 시나리오 1: 데이터 관리 체계 구축

- 전략: DAMA-DMBOK 프레임워크로 전체 그림 정의
- 품질: ISO (2016) 프로세스, ISO/IEC (2008) 특성
- 메타데이터: ISO/IEC (2023b) 등록소 구축
- 성숙도: DCAM으로 진단·개선

#### 14.6.2.2 시나리오 2: AI 시스템 도입

- 관리 시스템: ISO/IEC (2023a) AIMS
- 학습 데이터: ISO/IEC 5259 시리즈
- 리스크 관리: ISO/IEC 23894
- 데이터 품질: ISO 8000-8, ISO/IEC (2008)

#### 14.6.2.3 시나리오 3: 금융권 데이터 규제 대응

- 규제 요구: Basel Committee on Banking Supervision (2013)
- 메시징 표준: ISO (2022c)
- 데이터 품질: ISO 8000-110 (마스터 데이터)
- 정보 보안: ISO/IEC (2022)

### 14.6.3 실무 팁

**팁 1: 표준 원문을 확보하라**

국제 표준을 활용하려면 반드시 원문을 확보해야 한다. ISO 표준은 iso.org에서 구매하며, 가격은 파트당 100~300 CHF(약 15~40만원) 수준이다. 한국은 KS 표준으로 국내 발간되는 경우도 있어 e나라표준인증을 확인하라 (국가기술표준원 2024). ISO/IEC (2023a) 은 조직당 수백만 원의 투자이지만, 규제 준수와 조직 신뢰도에 비하면 매우 효과적인 투자이다.

**팁 2: 인증과 준수의 차이를 이해하라**

국제 표준은 크게 (1) 인증 대상(ISO 9001, ISO/IEC (2022), ISO/IEC (2023a) 등)과 (2) 준수 참조 대상(ISO/IEC (2023b), ISO/IEC (2008) 등)으로 나뉜다. 인증 대상은 공인 기관을 통해 공식 인증을 획득할 수 있으며 조직 신뢰의 증표가 된다. 준수 참조 대상은 인증 개념이 없이 조직이 스스로 준수를 선언하는 방식이다. 조직의 목표에 따라 두 방식을 구분하여 접근해야 한다.

**팁 3: 여러 표준을 통합 관리하라**

ISO 관리 시스템 표준들(9001·27001·42001·22301 등)은 High-Level Structure를 공유하므로 통합 관리가 가능하다. 각 표준별로 독립적인 관리 체계를 운영하면 중복과 낭비가 크다. **'통합 관리 시스템(Integrated Management System, IMS)'** 접근으로 하나의 정책·문서 체계 아래 여러 표준을 함께 관리하라. 이는 감사·인증 부담도 크게 절감한다.

**팁 4: 표준의 업데이트를 추적하라**

국제 표준은 통상 5년 주기로 재검토되며, 개정 시 새로운 요구사항이 추가된다. 조직이 채택한 표준의 개정 이력을 정기적으로 추적하고, 개정 시 갭 분석을 수행해야 한다. ISO는 표준 개정 소식을 iso.org의 뉴스레터로 제공하며, 국내에서는 국가기술표준원 웹사이트에서 확인할 수 있다.

**⚠ 주의: 표준 인용의 정확성**

표준을 인용할 때는 반드시 정확한 표기가 필요하다. 'ISO/IEC 42001:2023'과 같이 번호·연도·판을 명시해야 한다. 잘못된 인용은 계약서·규제 대응 문서에서 심각한 문제가 될 수 있다. 특히 여러 파트로 구성된 표준(ISO/IEC (2023b) 시리즈)은 파트 번호까지 정확히 표기해야 한다.

**⚠ 주의: 표준의 국가별 채택**

국제 표준(ISO)이 발간되어도 각국의 국가 표준으로 채택되는 데는 시차가 있다. 예를 들어 ISO/IEC (2023a) 은 한국의 KS 표준으로 채택되기까지 1~3년이 걸릴 수 있다. 국내에서는 국가기술표준원(KATS)이 채택 여부를 결정하며, KS로 채택되면 국내 규제·조달에서 활용 가능하다 (국가기술표준원 2024). 국제 표준과 국가 표준의 관계를 조직 내 표준 정책에 명시해야 한다.

## 14.7 사례 연구

### 14.7.1 사례 1: 글로벌 헬스케어 기업의 다중 표준 통합 인증

#### 14.7.1.1 배경

글로벌 헬스케어 기업 H사는 30개국에서 의약품·의료기기를 판매한다. 각국의 규제와 여러 국제 표준을 준수해야 하는 복잡한 환경이었다. 2022년 통합 관리 시스템(IMS) 구축을 결정했다 (BSI Group 2024).

#### 14.7.1.2 적용 표준

- ISO 9001:2015 — 품질 관리 시스템
- ISO/IEC 27001:2022 — 정보 보안
- ISO 13485:2016 — 의료기기 품질
- ISO/IEC 42001:2023 — AI 관리 (신규)
- ISO 8000-110 — 마스터 데이터 품질
- HL7 FHIR — 의료 정보 교환

#### 14.7.1.3 접근 방법

1. HLS 기반 통합 문서 체계 — 정책·매뉴얼·절차의 통합
2. 공통 감사 체계 — 여러 표준의 통합 내부 감사
3. 데이터 품질 표준화 — ISO 8000 마스터 데이터
4. AI 라이프사이클 통제 — 42001 Annex A 적용
5. 동시 인증 심사 — 여러 표준의 동시 인증

#### 14.7.1.4 성과

- 인증 유지 비용 40% 절감
- 내부 감사 시간 60% 단축
- 규제 대응 준비 기간 50% 단축
- 30개국 규제 감사 통과율 100%

#### 14.7.1.5 교훈

여러 국제 표준을 통합 관리하는 것은 초기 구축 부담은 크지만 장기적 효과는 매우 크다. 특히 HLS를 공유하는 ISO 관리 시스템 표준들은 통합 관리에 최적화되어 있다. 핵심은 각 표준의 요구사항을 조직 프로세스에 통합적으로 매핑하는 사전 설계이다.

### 14.7.2 사례 2: 국내 대기업의 ISO/IEC 42001 첫 인증

#### 14.7.2.1 배경

국내 대기업 S사는 사내 AI 시스템(챗봇·문서 분석·품질 예측 등) 도입을 확대하면서 AI 거버넌스 필요성이 부각되었다. 2024년 상반기 ISO/IEC (2023a) 인증 획득을 목표로 프로젝트를 시작했다 (한국지능정보사회진흥원 2024a).

#### 14.7.2.2 도전 과제

- 표준 발효 직후로 국내 인증 사례 부족
- 39개 Annex A 통제의 조직 적용 방식 불명확
- AI 시스템 별 상이한 성숙도
- 기존 ISO/IEC (2022) 과의 통합 필요

#### 14.7.2.3 접근 방법

1. Gap 분석 (3개월) — 현행 vs 42001 요구사항

2. AI 정책 및 조직 구축 (2개월) — CAIO 임명, 위원회 구성
3. Annex A 통제 구현 (6개월) — 39개 통제의 조직 적용
4. 문서화·프로세스 정착 (2개월)
5. 내부 감사 및 경영 검토 (1개월)
6. 공식 인증 심사 (2개월)

#### 14.7.2.4 성과 (14개월 후)

- ISO/IEC 42001 인증 획득 (국내 대기업 최초 사례)
- 사내 AI 시스템 27개 등록·관리
- AI 리스크 평가 100% 완료
- 공공 조달·수출에서 신뢰도 확보

#### 14.7.2.5 교훈

새로운 관리 시스템 표준의 조기 도입은 부담이 크지만, 시장 선점의 효과가 있다. 특히 ISO/IEC (2023a) 은 규제 강화와 함께 조직 신뢰도의 척도가 될 것으로 전망된다. S사의 경험은 (1) 기존 ISO 관리 시스템과의 통합 활용, (2) 경영진의 강력한 후원, (3) 외부 컨설팅과 내부 역량의 결합이 성공 요인이었음을 보여준다.

## 요약

본 장의 핵심 내용을 정리하면 다음과 같다.

1. **국제 표준의 위상** — 국제 표준은 조직 간·국가 간 상호운용성과 규제 준수의 기반이다. ISO/IEC JTC 1 산하 SC 32, SC 40, SC 42가 데이터 관리에 직결되는 표준을 개발한다.
2. **ISO/IEC 11179의 심층 구조** — 메타모델의 핵심은 Data Element Concept(DEC)와 3부분 명명 규칙이다. 시맨틱 상호운용성 확보의 이론적 기반을 제공한다.
3. **ISO 8000의 파트별 다양성** — 일반·마스터 데이터·트랜잭션·제품·산업별 참조 등 30개 이상의 파트로 구성된다. 8000-8(개념), 8000-61(프로세스), 8000-110(마스터 데이터)이 핵심이다.
4. **ISO/IEC 25012의 15개 특성** — 고유 특성 5개, 고유+시스템 종속 7개, 시스템 종속 3개로 데이터 품질을 정밀히 정의한다. ISO/IEC (2015b) 가 측정 지표를 제공한다.
5. **ISO/IEC 42001의 AIMS** — 2023년 발효된 최초의 AI 관리 시스템 국제 표준이다. HLS 구조를 따르며 39개 Annex A 통제로 조직적 AI 관리를 요구한다.
6. **표준 조합의 실무** — 여러 국제 표준은 계층적으로 조합하여 활용된다. 전략(DAMA)-도메인(8000, 42001)-기술(11179, 25012)-산업(BCBS 239, FHIR) 계층이 전형적이다.
7. **통합 관리 시스템(IMS)** — ISO 관리 시스템 표준들은 HLS를 공유하므로 통합 관리가 가능하다. 인증 유지 비용 절감과 조직 효율성 향상의 핵심 접근이다.

## 다음 학습 가이드

### 14.7.3 필수 표준 원문

- ISO/IEC (2023b) ~ ISO/IEC 11179-6 시리즈 (메타데이터 등록소)
- ISO (2022a), ISO 8000-8, ISO (2016), ISO 8000-110 (핵심 파트)
- ISO/IEC (2008), ISO/IEC (2015b) (데이터 품질과 측정)
- ISO/IEC (2023a), ISO/IEC 22989, ISO/IEC 23894 (AI 관리)
- ISO/IEC (2022), ISO 9001 (통합 관리 시스템 기반)

### 14.7.4 표준 활용 참고 자료

- ISO Central Secretariat. Directives Part 1 & 2 — 표준 개발 절차

- Sebastian-Coleman (2018) — 실무 관점
- DAMA International (2017) — 국제 표준의 실무 통합

### 14.7.5 관련 본서 참조 장

- 제 3장 주요 표준화 프레임워크 — 표준 개괄
- 제 5장 메타데이터 관리 — ISO/IEC 11179 실무
- 제 6장 데이터 품질 관리 — ISO 8000, ISO/IEC 25012 실무
- 제 9장 데이터 거버넌스 체계 — 통합 관리 시스템
- 제 12장 AI·ML과 데이터 표준화 — ISO/IEC 42001 실무

## 연습 문제

1. 여러분 조직의 데이터 요소 하나를 선택하여, ISO/IEC 11179의 3부분 명명 규칙(Object Class Term, Property Term, Representation Term)에 따라 명명하고 좋은 정의의 6가지 특성을 만족하는 정의를 작성하시오.
2. ISO/IEC 25012의 15개 품질 특성 중 여러분 조직에서 가장 취약한 5개를 식별하고, ISO/IEC 25024를 참조하여 각각의 측정 지표를 제안하시오.
3. 여러분 조직에 ISO/IEC 42001 인증을 도입한다면, Annex A 39개 통제 중 우선순위 상위 10개를 선정하고 그 근거를 제시하시오.

# V

# 5부: 디플래닉스 방법론


15 디플래닉스 방법론 개요 . . . . . 171
15.1 개요 . . . . . 171
15.2 본 장의 학습 목표 . . . . . 171
15.3 KDSM의 배경 . . . . . 172
15.4 KDSM의 구성 . . . . . 172
15.5 KDSM의 특징 . . . . . 176
15.6 다른 프레임워크와의 관계 . . . . . 177
15.7 KDSM의 적용 . . . . . 177

16 KDSM 데이터 모델링 . . . . . 181
16.1 개요 . . . . . 181
16.2 본 장의 학습 목표 . . . . . 181
16.3 KDSM 모델링 원칙 . . . . . 182
16.4 KDSM 모델링 프로세스 . . . . . 183
16.5 KDSM 명명 규칙 . . . . . 186
16.6 산출물 표준 . . . . . 187
16.7 검토와 승인 . . . . . 189
16.8 지원 도구 . . . . . 190
16.9 사례 연구 . . . . . 192

17 KDSM 메타데이터 관리 . . . . . 195
17.1 개요 . . . . . 195
17.2 본 장의 학습 목표 . . . . . 195
17.3 KDSM 메타데이터 원칙 . . . . . 196
17.4 KDSM 메타데이터 등록소 . . . . . 196
17.5 KDSM 메타데이터 수집·관리 프로세스 . . . 196
17.6 KDSM 메타데이터 활용 . . . . . 196
17.7 도구와 인프라 . . . . . 196
17.8 사례 연구 . . . . . 196

18 KDSM 데이터 품질 관리 . . . . . 197
18.1 개요 . . . . . 197
18.2 본 장의 학습 목표 . . . . . 197
18.3 KDSM 품질 관리 원칙 . . . . . 198
18.4 KDSM 품질 차원 . . . . . 198
18.5 KDSM 품질 측정 체계 . . . . . 198
18.6 KDSM 품질 관리 프로세스 (PDCA) . . . . . 198
18.7 데이터 품질 이슈 대응 . . . . . 198
18.8 도구와 인프라 . . . . . 198
18.9 사례 연구 . . . . . 198

18.9.1 용어 사전 . . . . . 199

CHAPTER 15

# 디플래닉스 방법론 개요

## 15.1 개요

디플래닉스가 개발한 KDSM(KANDLE Data Standardization Methodology)은 국내 조직의 데이터 표준화 실무를 위한 방법론이다. 앞선 1~14장이 특정 방법론에 기대지 않고 데이터 표준화의 개념·프레임워크·실무·고급 주제를 다루었다면, 5부는 그 일반론이 하나의 방법론 안에서 어떻게 구체화되는지를 다룬다. 일반론을 먼저 서술한 것은 KDSM이 어떤 문제를 어떤 방식으로 다루는지를 다른 방법론과 견주어 판단할 수 있게 하기 위해서이다.

KDSM은 **데이터 모델링 방법론**, **메타데이터 관리 방법론**, **데이터 품질관리 방법론**의 세 하위 방법론으로 구성된다. 세 방법론은 모두 Foundation·Guidance·Support·Experience의 **4계층 구조(4 Layered Architecture)**를 공유하며, 각 계층에는 절차를 정의한 문서부터 산출물 양식, 작업 기준, 자동화 소프트웨어, 경험 기록까지 여덟 가지 구성 요소가 놓인다. 이 때문에 KDSM은 지켜야 할 규칙을 적은 문서 한 벌이 아니라, 문서·도구·경험이 서로 맞물려 운용되는 **방법론 생태계(Methodology Ecosystem)**를 지향한다.

본 장은 KDSM 전체의 구조와 원리를 개관한다. 세 방법론 가운데 데이터 모델링 방법론은 제16장(Chapter 16.)에서, 메타데이터 관리와 데이터 품질관리 방법론은 각각 제17장(Chapter 17.)과 제18장(Chapter 18.)에서 다룬다.

> **i** 본서와 방법론 사이트의 관계
>
> KDSM의 각 방법론은 절차·기준·산출물·도구를 규정한 별도의 방법론 사이트(방법론 포털)로 제공된다. 본서의 5부는 그 구조와 원리를 설명하는 데 목적이 있으며, 수행에 필요한 상세 절차는 방법론 사이트가 정본이다. 두 자료의 내용이 다르면 방법론 사이트를 따른다.

## 15.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- KDSM이 등장한 배경과 해결하려는 문제를 설명할 수 있다
- KDSM을 구성하는 세 하위 방법론과 그 상호 관계를 설명할 수 있다
- 4계층 구조와 여덟 가지 구성 요소의 역할을 구분할 수 있다
- KDSM과 DAMA-DMBOK2·ISO/IEC 11179 등 국제 프레임워크의 관계를 설명할 수 있다
- 조직에 KDSM을 적용할 때의 출발점과 전제 조건을 이해한다

## 15.3 KDSM의 배경

### 15.3.1 표준의 부재가 아니라 표준의 미작동

데이터 프로젝트에서 반복적으로 마주하는 문제는 대체로 기술이 아니라 표준에서 비롯된다. 동일한 개념이 시스템마다 다른 이름으로 불리고, 같은 고객번호가 한 테이블에서는 문자형으로, 다른 테이블에서는 숫자형으로 정의된다. 날짜 하나가 `DATE`, `CHAR(8)`, `VARCHAR(10)`의 세 형식으로 공존하는 데이터베이스도 드물지 않다. 이러한 불일치는 조인 실패와 성능 저하, 의미 해석 비용, 영향도 분석의 불가능으로 이어진다(Chapter 1.).

그런데 실제 현장에서 더 흔한 것은 표준이 아예 없는 상태보다 **표준이 문서로는 있으나 작동하지 않는 상태**이다. 표준 문서가 작성된 뒤 프로젝트가 진행되는 동안 아무도 그 문서를 대조하지 않고, 예외가 승인 없이 쌓이며, 결국 표준은 "지켜지지 않는 문서"가 된다. 데이터 모델링 표준이 지켜지지 않는 원인을 진단하고 표준을 도구로 강제하라는 제4장의 권고(Chapter 4.)는 바로 이 상태를 겨냥한다.

### 15.3.2 일반 프레임워크와 실무 사이의 간극

제3장에서 살펴본 DAMA-DMBOK2(DAMA International (2017))와 ISO/IEC 11179(ISO/IEC (2023b)) 같은 프레임워크는 데이터 관리가 **무엇을** 다루어야 하는지를 정의한다. 지식 영역을 나누고, 메타데이터 등록소의 구조를 규정하며, 명명의 원칙을 제시한다. 그러나 프레임워크는 의도적으로 특정 조직의 절차와 양식을 정하지 않는다. 실무자가 내일 아침 무엇을 어떤 양식에 기재하고 누구의 승인을 받아야 하는지는 각 조직이 다시 설계하여야 한다.

국내 조직은 여기에 고유한 조건이 더해진다. 한글 업무 용어를 영문 약어 기반의 물리명으로 옮겨야 하고, 공공 부문을 중심으로 행정안전부의 공공데이터 공통표준(공통표준단어·공통표준용어·공통표준도메인)과의 정합을 고려하여야 한다. 한글 용어를 표준단어 단위로 분해하는 판단, 동음이의어와 이음동의어의 처리, 분류어의 선택처럼 국제 표준이 다루지 않는 판단이 실무에서는 가장 많은 시간을 차지한다.

### 15.3.3 방법론 생태계라는 접근

KDSM은 이 간극을 세 가지 방식으로 메운다. 첫째, 일반 프레임워크의 원칙을 국내 실무의 **절차와 산출물 양식**으로 구체화한다. 둘째, 그 절차와 양식을 기계가 읽어 검사하고 변환할 수 있도록 **라이브러리와 툴킷**을 함께 제공한다. 셋째, 방법론을 적용하면서 얻은 성공과 시행착오를 **명시적 지식 자산**으로 기록하여 다음 개정의 근거로 삼는다. 절차·도구·경험을 하나의 구조로 묶는 이 접근이 KDSM을 방법론 생태계라고 부르는 이유이다.

## 15.4 KDSM의 구성

### 15.4.1 세 하위 방법론

KDSM의 세 하위 방법론은 독립적으로 운용할 수 있으나, 데이터 표준화라는 하나의 목적 아래 서로 연결된다(Figure 15.1).

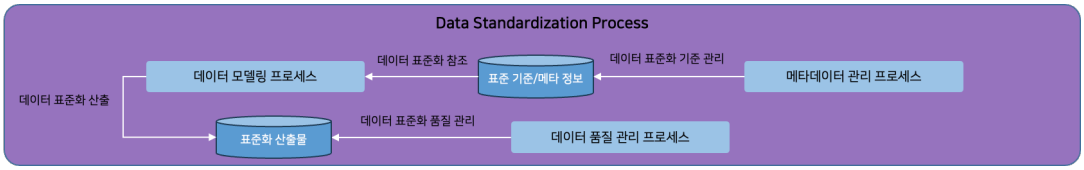



Figure 15.1. 세 하위 방법론의 관계

Table 15.1. KDSM의 세 하위 방법론

| 하위 방법론 | 다루는 것 | 다른 방법론과의 관계 |
|---|---|---|
| 데이터 모델링 방법론 | 표준단어·표준용어·표준도메인·표준코드를 정의하고, 이를 논리·물리 데이터 모델에 적용하는 절차 | 표준 기준과 메타 정보를 참조하여 표준화 산출물을 만든다 |
| 메타데이터 관리 방법론 | 표준 기준과 메타 정보를 수집·관리·활용하는 절차 | 모델링이 참조하는 데이터 표준화 기준을 관리한다 |
| 데이터 품질관리 방법론 | 표준화 산출물과 데이터의 품질을 진단·개선·모니터링하는 절차 | 모델링이 산출한 표준화 산출물의 품질을 관리한다 |

세 방법론 가운데 데이터 모델링 방법론이 가장 앞에 놓인다. 메타데이터 관리는 모델링이 만들어 낸 표준 사전과 모델 정의를 관리 대상으로 삼고, 품질관리는 그 표준을 판정 기준으로 삼기 때문이다. 표준 사전 없이 메타데이터를 관리하면 관리 대상의 이름부터 일관되지 않고, 표준 없이 품질을 진단하면 무엇이 결함인지 판정할 기준이 없다.

## 15.4.2 4계층 구조

세 방법론은 동일한 4계층 구조로 설계된다. 어느 방법론을 익히든 같은 구조로 학습하고 적용할 수 있게 하기 위해서이다(Figure 15.2).

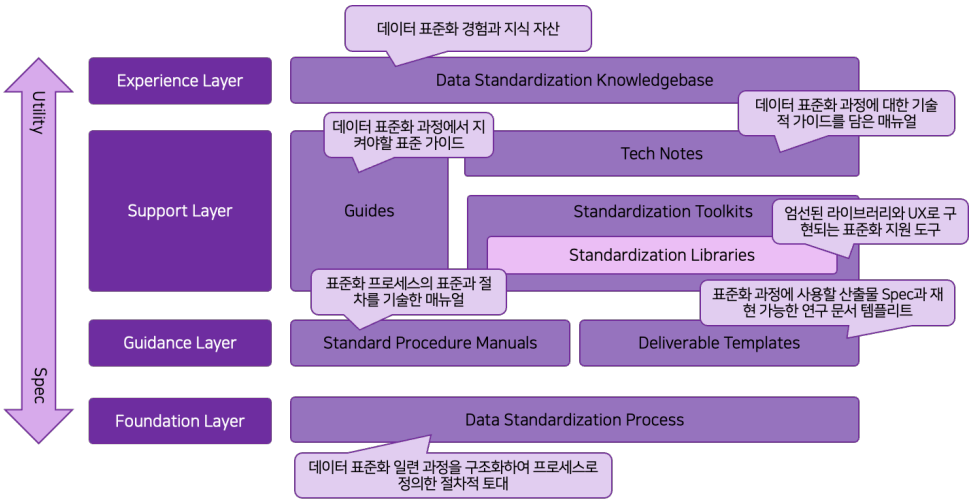



Figure 15.2. KDSM의 4계층 구조와 구성 요소

Table 15.2. KDSM의 4계층

| 계층 | 정의 | 구성 요소 |
|---|---|---|
| Foundation | 표준화 목표를 달성하기 위해 수행해야 하는 일련의 작업 절차와 흐름을 정의하는 핵심 통제 체계 | Data Standardization Process |
| Guidance | 방향성 정렬(Alignment), 품질과 다양성의 균형(Trade-off Control), 의사결정 지원(Decision Support), 일관성·신뢰성 보장(Standardization)을 위한 기준과 절차 | Standard Procedure Manuals · Deliverable Templates |
| Support | 실무의 품질 보장과 생산성 극대화를 위한 규칙, 자동화 도구, 기술 문서 | Guides · Standardization Library · Standardization Toolkits · Tech Notes |
| Experience | 실무자의 노하우와 조직의 과거 데이터 관리 이력으로 축적된, 성공과 시행착오에 대한 명시적 지식 체계 | Data Standardization Knowledgebase |

계층은 아래에서 위로 쌓인다. 아래 계층일수록 사양(Spec)의 성격을 띠고, 위 계층일수록 도구(Utility)의 성격을 띤다(Figure 15.3). 사양은 작업 절차·기준·요건처럼 무엇이 갖추어져야 하는지를 정의한 것이고, 도구는 작업에 쓰이는 구체적인 형상이 있는 문서나 프로그램이다. 아래 계층이 바뀌면 그 위의 계층이 모두 영향을 받으므로, 프로세스를 개정하면 매뉴얼과 템플릿, 가이드와 라이브러리를 차례로 맞추어야 한다.

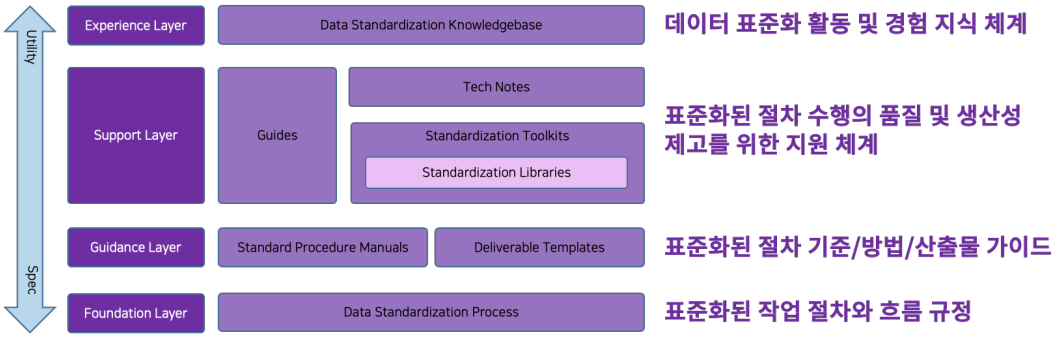



Figure 15.3. 계층별 역할: 아래는 사양, 위는 도구

### 15.4.3 여덟 가지 구성 요소

네 계층에는 여덟 가지 구성 요소가 놓인다. 각 구성 요소의 기능과 형상은 다음과 같다(Figure 15.4).

Table 15.3. KDSM의 여덟 가지 구성 요소

| 계층 | 구성 요소 | 기능 | 형상 |
|---|---|---|---|
| Foundation | 데이터 표준화 프로세스(Data Standardization Process) | 데이터 표준화 프로세스의 정의 | 프로세스 다이어그램 |
| Guidance | 표준 및 절차 매뉴얼(Standard Procedure Manuals) | 데이터 표준화의 표준과 절차 제시 | 문서 |
| Guidance | 산출물 템플릿(Deliverable Templates) | 표준 산출물 템플릿 제공 | 문서 |
| Support | 가이드(Guides) | 데이터 표준화 과정의 표준 가이드 제시 | 문서 |
| Support | 표준화 라이브러리(Standardization Library) | 데이터 표준화를 위한 라이브러리 제공 | 파이썬 라이브러리 등 |
| Support | 표준화 툴킷(Standardization Toolkits) | 데이터 표준화를 위한 분석 도구 제공 | 툴킷·애플리케이션 |
| Support | 테크니컬 노트(Tech Notes) | 데이터 표준화 테크니컬 노트 제공 | 문서 |
| Experience | 데이터 표준화 지식 베이스(Data Standardization Knowledgebase) | 데이터 표준화의 명시적 경험·지식 제공 | 지식 베이스 |

| Layers | Components | Components(KO) | Functions | 형상 | 예시 |
|---|---|---|---|---|---|
| Foundation | Data Standardization Process | 데이터 표준화 프로세스 | 데이터 표준화 프로세스의 정의 | Process Diagram | 데이터 표준화 프로세스,<br>데이터 표준화 관리 프로세스 |
| Guidance | Standard Procedure Manuals | 표준 및 절차 매뉴얼 | 데이터 표준화의 표준과 절차 제시 | Documents | 데이터 표준화 표준 및 절차 매뉴얼 |
| Guidance | Deliverable Templates | 산출물 템플릿 | 표준 산출물 템플릿 제공 | Documents | 테이블정의서, 컬럼정의서,<br>도메인정의서 |
| Support | Guides | 가이드 | 데이터 표준화 과정의 표준 가이드 제시 | Documents | 명명규칙, 단어(약어) 사전,<br>코딩가이드 |
| Support | Standardization Library | 표준화 라이브러리 | 데이터 표준화를 위한 라이브러리 제공 | Python Library,<br>SQL Function | 데이터 품질 진단,<br>데이터 클렌징,<br>데이터 표준화 진단,<br>데이터 정합성 진단 |
| Support | Standardization Toolkits | 표준화 툴 박스 | 데이터 표준화를 위한 분석 툴 박스 제공 | Toolkit / APP | 단어(약어) 추천 앱<br>데이터 표준화 진단 앱 |
| Support | Tech Notes | 테크니컬 노트 | 데이터 표준화 테크니컬 노트 제공 | | 표준사전(단어/용어,도메인,<br>코드) 정의 방법서 |
| Experience | Standardization Knowledgebase | 데이터 표준화 지식 베이스 | 데이터 표준화 명시적 경험/지식 제공 | Knowledge Base | 공공데이터 메타설계 소개서 |

Figure 15.4. 계층별 구성 요소의 기능

구성 요소는 아래에서 정의한 것을 위에서 구체화하는 관계로 연결된다.

- **프로세스**는 표준화 작업을 Phase → Activity → Task의 세 단계로 분해하여, 누가 언제 무엇을 수행하는지와 각 Task의 입력물·산출물을 정의한다.
- **표준 및 절차 매뉴얼**은 프로세스의 각 Task를 어떻게 수행하는지(기준·절차·역할·체크리스트)를 규정하고, **산출물 템플릿**은 그 결과를 어떤 양식에 담는지를 규정한다. 두 구성 요소는 Guidance 계층에서 짝을 이룬다.
- **가이드**는 매뉴얼이 정한 절차에 대해 작업 기준과 판단 원칙을 제시한다. 명명규칙, 표준사전 작성 기준, 코딩 가이드가 여기에 해당한다. 가이드의 조항은 반드시 지켜야 하는 규칙과 따를 것을 권하는 지침으로 구분된다.

- **라이브러리**는 산출물의 사양과 기준에 부합하는 산출물의 생성을 돕거나 자동으로 생산하고, **툴킷**은 같은 기능을 화면과 사용자 경험(UI/UX)으로 제공한다. **테크니컬 노트**는 현장에서 이 도구를 운용하는 방법과 문제 해결 방법을 담는다.
- **지식 베이스**는 아래 세 계층의 구성 요소를 운용하며 얻은 경험 자산과 지식 자산의 총체이다. 방법론 포털, 데이터 표준화 저서(본서), 지식 베이스 웹 사이트가 여기에 해당한다.

라이브러리와 툴킷의 관계에는 KDSM의 설계 원칙 하나가 반영되어 있다. 툴킷은 판단을 다시 구현하지 않고 라이브러리를 호출하여 그 결과를 화면에 전달한다. 따라서 화면으로 수행하든 커맨드라인으로 수행하든 같은 조건이면 같은 결과를 얻는다. 판단 기준이 두 곳에 존재하면 언젠가 서로 어긋나기 때문이다.

## 15.5 KDSM의 특징

### 15.5.1 일관된 구조

세 방법론이 같은 4계층과 같은 구성 요소를 공유하므로, 하나를 익히면 나머지의 구조를 미리 알 수 있다. 데이터 모델링 방법론에서 표준 및 절차 매뉴얼을 찾던 위치에서 메타데이터 관리 방법론의 매뉴얼을 찾을 수 있고, 산출물 템플릿과 가이드도 마찬가지이다. 구조의 일관성은 방법론을 조직에 교육하고 개정하는 비용을 낮춘다.

### 15.5.2 실무 우선

KDSM은 이론적 완결성보다 담당자가 현업에서 바로 적용할 수 있는지를 기준으로 절차와 산출물을 설계한다. 표준 및 절차 매뉴얼은 모든 Task를 개요·표준 기준·단계별 절차·입력물과 산출물·역할과 책임·체크리스트·예시의 같은 형식으로 기술하여, 수행자가 바뀌어도 같은 품질로 작업을 반복할 수 있게 한다. 원 프로세스 정의서에 없어 일반적인 모범사례로 보강한 내용은 따로 표시하여 조직이 확정한 내용과 구별한다.

### 15.5.3 기계가 읽을 수 있는 표준

KDSM은 표준 사전과 산출물을 사람이 읽는 문서인 동시에 기계가 읽는 데이터로 다룬다. 표준단어·표준용어·표준도메인·표준코드는 데이터베이스의 메타테이블로 관리하고, 산출물은 정해진 양식의 엑셀 파일로 작성하여 프로그램이 읽을 수 있게 한다. 표준이 기계가독형으로 관리되면 물리명 생성, 산출물 사이의 대조, 비표준 항목 탐지, DDL과 ERD의 생성을 자동화할 수 있다. KDSM에서 표준화는 문서를 작성하는 일이 아니라 **자동화할 수 있는 메타데이터를 구축하는 일**이다.

이 특징은 "표준은 도구로 강제하라"는 원칙(Chapter 4.)을 방법론의 구조에 넣은 것이다. 표준을 지켰는지는 사람의 기억이 아니라 규칙의 실행 결과로 판정하며, 그 결과에는 무엇을 검사하였고 무엇을 검사하지 않았는지가 함께 남는다.

### 15.5.4 사람의 판단과 기계의 대조를 구분한다

자동화의 범위에는 분명한 경계가 있다. 어떤 낱말을 표준단어로 삼을지, 어떤 엔터티가 필요한지, 경계에 있는 대상을 어느 쪽으로 분류할지는 사람이 정한다. 라이브러리는 그 결정을 여러 문서에 옮겨 적은 값이 서로 어긋나지 않는지 확인하고, 확인을 마친 내용을 다른 형태로 변환하는 일을 맡는다. LLM이 새 표준단어의 초안을 제시하는 기능도 있으나, 초안의 상태는 언제나 후보이며 등재 여부는 사람이 판정한다.

### 15.5.5 경험이 방법론으로 돌아온다

프로세스와 가이드는 개정되기 전까지 같은 내용을 유지하지만, 경험 자산은 프로젝트를 수행할 때마다 늘어난다. 같은 예외가 반복하여 승인되면 그것은 표준이 현실과 괴리되었다는 신호이며, 예외를 계

속 승인할 것이 아니라 표준의 개정을 발의하여야 한다. Experience 계층은 이렇게 늘어난 경험이 다시 가이드와 테크니컬 노트의 개정 근거가 되도록 연결한다.

## 15.6 다른 프레임워크와의 관계

KDSM은 국제 프레임워크를 대체하지 않는다. 프레임워크가 정의한 영역과 원칙 위에서 국내 실무의 절차·양식·도구를 구체화한다. 주요 프레임워크와의 대응은 다음과 같다.

Table 15.4. 국제·국내 프레임워크와 KDSM의 대응

| 프레임워크 | 해당 영역 또는 원칙 | KDSM의 대응 |
|---|---|---|
| DAMA-DMBOK2 (DAMA International (2017)) | 데이터 모델링과 설계 | 데이터 모델링 방법론 |
| DAMA-DMBOK2 | 메타데이터 관리 | 메타데이터 관리 방법론 |
| DAMA-DMBOK2 | 데이터 품질 관리 | 데이터 품질관리 방법론 |
| DAMA-DMBOK2 | 데이터 거버넌스 | 세 방법론에 공통으로 놓인 확정·승인·변경관리 절차와 역할 |
| ISO/IEC 11179-5 (ISO/IEC (2015a)) | 객체어·특성어·표현어의 명명 구조 | 표준용어의 주제어·수식어·분류어 구조 |
| ISO 8000-61 (ISO (2016)) | 프로세스 참조 모델에 기반한 데이터 품질관리 | Phase → Activity → Task로 정의한 데이터 표준화 프로세스 |
| PMBOK (Project Management Institute (2021)) | 산출물 중심의 작업 분해(WBS) | 앞 Task의 산출물이 다음 Task의 입력물이 되는 프로세스 구조 |
| 행정안전부 공공데이터 공통표준 | 공통표준단어·공통표준용어·공통표준도메인 | 표준 사전의 기준 자료. 개정될 때마다 원본과 사전을 대조한다 |

데이터 거버넌스는 KDSM에서 별도의 하위 방법론으로 두지 않는다. 거버넌스는 독립된 활동이 아니라 각 방법론의 절차 안에서 작동하여야 하기 때문이다. 데이터 모델링 방법론에서는 후보 표준을 확정·공표하는 단계, 확정된 표준의 변경을 관리하는 단계, 표준의 등록·수정·폐기를 제안받아 승인 권한을 가진 사람이 판정하는 절차가 거버넌스의 역할을 수행한다. 제9장(Chapter 9.)이 다룬 거버넌스의 원칙이 KDSM에서는 이 절차들로 구현된다.

ISO/IEC 11179-5와의 대응은 표준용어의 구성 문법에서 가장 분명하게 드러난다. 표준용어의 맨 끝에 오는 분류어가 값의 형식과 도메인을 결정하는 구조는 표현어(representation term)가 데이터 요소의 표현 형식을 나타내는 국제 표준의 구조와 같다(Chapter 14.). KDSM의 용어 문법은 이 국제 표준 구조를 한국어 업무 용어에 적용한 것이다.

## 15.7 KDSM의 적용

### 15.7.1 출발점은 데이터 모델링이다

조직에 KDSM을 적용할 때에는 데이터 모델링 방법론에서 시작한다. 메타데이터 관리와 품질관리가 모두 모델링이 만든 표준 사전을 기준으로 삼기 때문이다. 표준 사전을 먼저 세우고, 그 사전으로 논리·물리 모델을 정의하며, 정의된 모델과 운영 데이터베이스의 일치를 점검하는 순서가 자연스럽다.

이미 운영 중인 시스템을 인수한 경우에는 진단부터 시작할 수 있다. 운영 데이터베이스가 표준에서 얼마나 벗어나 있는지를 먼저 측정하여야 표준 사전을 무엇으로 채울지와 개선의 우선순위를 정할 수 있다.

### 15.7.2 전제 조건

KDSM을 적용하려면 다음 조건이 갖추어져야 한다.

1. **표준을 관리하는 역할**: 후보 표준을 검토하고 확정하며 예외를 승인하는 데이터 관리 조직 또는 담당자가 지정되어 있어야 한다. 표준의 확정이 특정 개인의 판단에만 기대면 담당자가 바뀔 때마다 같은 논의가 반복된다.
2. **단일한 표준 사전**: 표준 사전을 파일로 나누어 돌리지 않고 조직 전체가 참조하는 하나의 저장소로 관리하여야 한다. 파일형 정의서는 초기 정의와 검토 단계의 작업본으로 사용하고, 확정된 뒤에는 사전에 등록하여 참조원을 하나로 유지한다.
3. **산출물 양식의 준수**: 산출물을 KDSM이 정한 양식대로 작성하여야 도구가 읽을 수 있다. 양식을 임의로 바꾸면 자동 검증과 변환의 이점을 잃는다.
4. **변경의 절차화**: 확정된 표준의 변경은 반드시 정해진 절차를 거쳐 사전·모델·정의서에 함께 반영되어야 한다. 한 곳만 고치면 그 지점에서 불일치가 생긴다.

## 요약

1. KDSM은 방법론 생태계이다
   - KDSM은 규칙을 적은 문서 한 벌이 아니라, 절차·양식·도구·경험이 맞물려 운용되는 방법론 생태계를 지향한다.
2. 세 하위 방법론
   - 데이터 모델링·메타데이터 관리·데이터 품질관리의 세 방법론으로 구성되며, 표준 사전을 만드는 데이터 모델링 방법론이 나머지 둘의 기준을 제공한다.
3. 4계층과 여덟 구성 요소
   - Foundation·Guidance·Support·Experience의 4계층에 프로세스, 매뉴얼, 템플릿, 가이드, 라이브러리, 툴킷, 테크니컬 노트, 지식 베이스가 놓인다. 아래 계층은 사양, 위 계층은 도구의 성격을 띤다.
4. 기계가 읽을 수 있는 표준
   - 표준 사전과 산출물을 기계가독형으로 관리하여 검증과 변환을 자동화한다. 판단은 사람이, 대조와 변환은 기계가 맡는다.
5. 프레임워크의 구체화
   - KDSM은 DAMA-DMBOK2와 ISO/IEC 11179 등 국제 프레임워크를 대체하지 않고, 그 원칙을 국내 실무의 절차와 양식으로 구체화한다.

## 다음 학습 가이드

### 15.7.3 관련 본서 참조 장

- 제 3장 주요 표준화 프레임워크 — KDSM의 프레임워크 기반
- 제 4장 데이터 모델링 표준 — 데이터 모델링의 일반 원칙
- 제 9장 데이터 거버넌스 체계 — KDSM의 확정·승인 절차가 구현하는 거버넌스 원칙
- 제 16장 KDSM 데이터 모델링 — KDSM 모델링 방법론
- 제 17장 KDSM 메타데이터 관리 — KDSM 메타데이터 관리
- 제 18장 KDSM 데이터 품질 관리 — KDSM 품질 관리

### 15.7.4 KDSM 자료

- KANDLE 데이터 표준화 방법론 포털 — 세 방법론의 절차·매뉴얼·템플릿·가이드를 제공하는 정본
- 데이터 표준화 라이브러리 사용자 매뉴얼(유충현 (2026a)) — 표준화 작업의 기준과 `kdsm` 라이브러리의 사용법
- 데이터 표준화 애플리케이션 사용자 매뉴얼(유충현 (2026b)) — KDSM 웹 애플리케이션의 사용과 구축·운영

## 연습 문제

1. KDSM과 DAMA-DMBOK2 프레임워크의 공통점과 차이점을 분석하시오. 특히 DAMA-DMBOK2가 별도의 지식 영역으로 두는 데이터 거버넌스를 KDSM이 하위 방법론으로 두지 않는 이유를 설명하시오.
2. KDSM의 4계층 구조에서 아래 계층은 사양, 위 계층은 도구의 성격을 띤다. 데이터 표준화 프로세스의 한 Task가 개정되었을 때, 그 영향이 어느 구성 요소에 어떤 차례로 전파되는지 서술하시오.
3. 여러분 조직의 데이터 표준 문서 하나를 골라 그것이 “작동하는 표준”인지 “지켜지지 않는 문서”인지 판단하시오. 판단의 근거와 함께, 그 표준을 기계가 검사할 수 있게 하려면 무엇을 바꾸어야 하는지 제시하시오.
4. 운영 중인 시스템을 인수한 조직이 KDSM을 적용하려 한다. 진단에서 시작하는 경우와 표준 사전 구축에서 시작하는 경우를 비교하고, 어느 쪽이 적절한지 조건을 들어 논하시오.

CHAPTER 16

# KDSM 데이터 모델링

## 16.1 개요

KDSM 데이터 모델링 방법론은 표준 사전을 세우고, 그 사전으로 논리·물리 데이터 모델을 정의하며, 정의된 모델이 표준을 지키는지 점검하기까지의 전 과정을 다루는 방법론이다. KDSM의 세 하위 방법론 가운데 가장 먼저 적용하는 방법론이며, 여기서 만든 표준 사전과 모델 정의가 메타데이터 관리와 데이터 품질관리의 기준이 된다(Chapter 15.).

제 4장 데이터 모델링 표준이 개념·논리·물리 모델의 추상화 수준, 정규화, 명명 원칙 같은 일반 원리를 다루었다면, 본 장은 KDSM이 그 원리를 어떤 **절차**와 **산출물**과 **도구**로 구현하는지를 다룬다. 제4장이 "엔티티와 속성의 이름은 ISO/IEC 11179의 명명 원칙을 따라야 한다"고 말한다면, 본 장은 그 원칙을 따르기 위해 어떤 사전에 무엇을 먼저 등재하고, 어떤 양식에 어떤 값을 기재하며, 그 값이 맞는지를 무엇으로 확인하는지를 설명한다.

KDSM 데이터 모델링 방법론의 핵심은 **이름과 형식을 사람이 매번 다시 정하지 않게 하는 것**이다. 표준단어의 영문약어를 정해 두면 컬럼의 물리명이 계산되고, 표준도메인을 정해 두면 데이터 타입과 길이가 결정된다. 사람이 정하는 것은 단어와 도메인이며, 나머지는 그 결정에서 도출된다.

## 16.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- KDSM 데이터 모델링의 원칙과 그 근거를 설명할 수 있다
- 7 Phase · 14 Activity · 18 Task로 구성된 모델링 프로세스의 흐름과 산출물을 이해한다
- 표준단어·표준용어·표준도메인·표준코드로 이름과 형식을 정하는 KDSM 명명 규칙을 적용할 수 있다
- 데이터 표준·논리 모델·물리 모델의 산출물 양식과 그 관계를 이해한다
- 표준의 확정·검증·변경을 관리하는 검토와 승인 절차를 운영할 수 있다
- 라이브러리와 툴킷이 모델링의 어느 단계를 지원하는지 설명할 수 있다

## 16.3 KDSM 모델링 원칙

### 16.3.1 하나의 사실에는 하나의 이름

같은 사실은 어디에서나 같은 이름으로 부르고, 하나의 이름은 하나의 사실만 가리킨다("One Fact, One Name"). 이 양방향 유일성이 무너지는 지점에서 의사소통의 오류가 발생한다. KDSM은 이 원칙을 표준 사전으로 구현한다. 논리 모델의 속성명은 표준용어의 한글명을, 물리 모델의 컬럼명은 표준용어의 물리명을 사용하며, 논리명과 물리명은 1:1로 대응한다.

### 16.3.2 한 번 정한 것을 여러 곳에 반복하여 기록하지 않는다

같은 사실을 여러 곳에 기록하면 언젠가 어긋난다. 사람이 옮겨 적는 지점마다 불일치가 생길 수 있기 때문이다. KDSM은 앞 단계에서 정한 것이 뒤 단계를 결정하도록 사전을 설계하여 옮겨 적는 지점을 줄인다.

Table 16.1. 결정과 도출의 관계

| 사람이 정하는 것 | 그로부터 도출되는 것 |
|---|---|
| 표준단어의 영문약어 | 표준용어의 물리명(영문약어를 밑줄로 연결) |
| 표준용어의 구성 단어 | 표준용어의 영문명, 물리명, 표준도메인(마지막 분류어에서 유도) |
| 표준도메인 | 컬럼의 데이터 타입과 길이 |
| 표준용어 | 엔터티명·속성명, 테이블·컬럼의 이름 |

그럼에도 옮겨 기록할 수밖에 없는 지점은 남는다. 논리 모델과 물리 모델은 같은 엔터티와 속성을 각각 기록하고, 정의서와 운영 데이터베이스는 같은 테이블을 각각 기술한다. KDSM은 이러한 지점마다 검증 규칙을 두어 두 기록이 서로 어긋나지 않았는지 대조한다.

### 16.3.3 표준 사전은 기계가 읽을 수 있어야 한다

표준단어·표준용어·표준도메인·표준코드는 조직 전체가 참조하는 사전으로, 파일이 아니라 데이터베이스 테이블로 관리하는 것을 원칙으로 한다. 파일형 정의서는 초기 정의와 검토 단계의 작업본으로 쓰고, 확정된 뒤에는 사전에 등록하여 참조원을 하나로 유지한다. 사전이 기계가독형이어야 물리명 계산, 비표준 항목 탐지, 모델과 데이터베이스의 대조를 자동화할 수 있다.

### 16.3.4 판단은 사람이, 대조는 기계가

표준화 작업에는 사람의 판단이 필요한 부분과 규칙이 정해진 부분이 섞여 있다. 어떤 낱말을 표준단어로 삼을지, 어떤 엔터티가 필요한지는 사람이 정한다. 그 결정을 여러 문서에 옮겨 적은 값이 서로 일치하는지 확인하고, 확인을 마친 내용을 DDL이나 ERD로 바꾸는 일은 규칙이 정해져 있으므로 기계가 맡는다.

Table 16.2. 사람의 판단과 라이브러리의 역할

| 사람이 하는 일 | 라이브러리가 하는 일 |
|---|---|
| 표준단어·표준용어를 무엇으로 정할지 결정한다 | 이름을 표준단어 단위로 분해하여 후보를 제시한다 |
| 어떤 엔터티와 속성이 필요한지 결정한다 | 기재한 이름이 사전에 등재되어 있는지 검사한다 |
| 산출물에 값을 기재한다 | 여러 문서에 기재한 값이 서로 일치하는지 대조한다 |
| 개선 대상과 순서를 결정한다 | 표준 준수 수준을 측정하여 제시한다 |

### 16.3.5 국제 표준의 명명 구조를 따른다

KDSM의 명명은 ISO/IEC 11179-5가 정의하는 객체어(object class term)·특성어(property term)·표현어(representation term)의 조합 구조를 따른다(ISO/IEC (2015a)). 표준용어의 주제어는 객체어에, 수식어는 한정어에, 맨 끝의 분류어는 표현어에 해당한다(Chapter 14.). KDSM의 용어 문법은 이 국제 표준 구조를 한국어 업무 용어에 적용한 것이다.

## 16.4 KDSM 모델링 프로세스

### 16.4.1 3단계 분해 구조

KDSM 데이터 모델링 프로세스는 **Phase → Activity → Task**의 세 단계로 정의된다. 각 Task는 하나 이상의 산출물을 생성하고, 앞 Task의 산출물은 다음 Task의 입력물이 된다. 산출물을 중심으로 작업을 분해하는 이 구조는 프로젝트 관리 표준의 작업 분해 방식과 같은 맥락에 있으며(Project Management Institute (2021)), 프로세스 참조 모델에 기반한 관리는 데이터 품질관리 국제 표준의 접근이기도 하다(ISO (2016)).

Table 16.3. 프로세스의 3계층

| 계층 | 설명 | 표기 |
|---|---|---|
| Phase | 프로세스의 최상위 단계 | 1, 2, 3 … |
| Activity | Phase를 구성하는 활동 단위 | 1.1, 1.2 … |
| Task | Activity를 구성하는 최소 수행 단위 | 1.1.1, 1.1.2 … |

### 16.4.2 7 Phase · 14 Activity · 18 Task

데이터 모델링 프로세스는 7개 Phase, 14개 Activity, 18개 Task로 구성된다. 큰 흐름은 **요구사항분석 → 데이터 표준화 정의 → 데이터 표준화 확정 → 데이터 표준화 적용**의 네 단계이며, 표준단어·표준용어·표준도메인·표준코드를 정의하는 네 Phase가 데이터 표준화 정의로 묶인다(Figure 16.1).

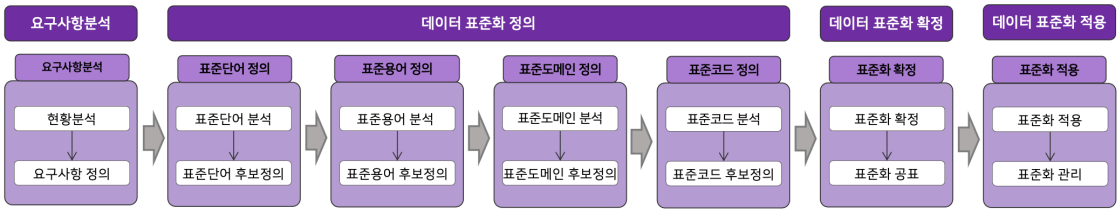



Figure 16.1. 데이터 모델링 프로세스의 생명주기

Table 16.4. 데이터 모델링 프로세스의 구성

| Phase | Activity | Task | 주요 산출물 |
|---|---|---|---|
| 1. 요구사항 분석 | 1.1 현황분석 | 표준화 원칙 이해 · 표준화 범위·방법 정의 | 표준화 대상·범위·방법 |
| | 1.2 요구사항 정의 | 데이터 구성요소 식별 · 표준화 대상 데이터 정의 | 표준화 대상 데이터 형상 |
| 2. 표준단어 정의 | 2.1 표준단어 분석 | 표준단어 분석 | — |
| | 2.2 표준단어 정의 | 표준단어 정의 | 후보 표준단어 |
| 3. 표준용어 정의 | 3.1 표준용어 분석 | 표준용어 분석 | — |
| | 3.2 표준용어 정의 | 표준용어 정의 | 후보 표준용어 |
| 4. 표준도메인 정의 | 4.1 표준도메인 분석 | 표준도메인 분석 | — |
| | 4.2 표준도메인 정의 | 표준도메인 정의 | 후보 표준도메인 |
| 5. 표준코드 정의 | 5.1 표준코드 분석 | 표준코드 분석 | — |
| | 5.2 표준코드 정의 | 표준코드 정의 | 후보 표준코드 |
| 6. 표준화 확정 | 6.1 표준화 확정 | 데이터 표준화 확정 | 확정된 표준화 내역 |
| | 6.2 표준화 공표 | 데이터 표준화 공표 | 표준화 메타데이터(표준사전·표준코드) |
| 7. 표준화 적용 | 7.1 표준화 적용 | 논리 데이터 모델링 · 물리 데이터 모델링 | 논리·물리 데이터 모델과 정의서 |
| | 7.2 표준화 관리 | 표준 준수 여부 점검 · 변경관리 | 비표준 내역, 변경된 표준화 내역 |

요구사항분석(Phase 1)에서 표준화의 원칙·범위·방법을 정하고 표준화 대상 데이터를 식별한다. 이어 표준단어(Phase 2)·표준용어(Phase 3)·표준도메인(Phase 4)·표준코드(Phase 5)를 차례로 분석하고 후보로 정의한다. 표준단어는 이름의 최소 단위이고, 표준용어는 이를 조합한 이름이며, 표준도메인은 용어가 가질 데이터 형식을, 표준코드는 코드성 데이터의 값 체계를 규정한다. 후보 표준은 표준화 확

정(Phase 6)에서 이해관계자의 합의를 거쳐 확정·공표되고, 표준화 적용(Phase 7)에서 논리·물리 데이터 모델에 반영된 뒤 표준 준수 여부 점검과 변경관리로 이어진다.

각 Task의 입력물과 산출물의 연쇄는 Figure 16.2 과 같다. 앞 Task의 산출물이 다음 Task의 입력물이 되는지를 이 표로 확인할 수 있다.

| Phase | Activity | Task | Input | Deliverable |
|---|---|---|---|---|
| 현황분석 | 현황분석 | 표준화 원칙 이해 | | |
| | | 표준화 범위/방법 정의 | | 표준화 대상/범위/방법 |
| | 요구사항 정의 | 데이터 구성요소 식별 | 표준화 대상 데이터 정보 | |
| | | 표준화 대상 데이터 정의 | | 표준화 대상 데이터 형상 |
| 표준단어 정의 | 표준단어 분석 | 표준단어 분석 | 표준단어 사전 | |
| | 표준단어 후보정의 | 표준단어 후보정의 | | 후보 표준단어 |
| 표준용어 정의 | 표준용어 분석 | 표준용어 분석 | 표준용어 사전 | |
| | 표준용어 후보정의 | 표준용어 후보정의 | | 후보 표준용어 |
| 표준도메인 정의 | 표준도메인 분석 | 표준도메인 분석 | 표준도메인 사전 | |
| | 표준도메인 후보정의 | 표준도메인 후보정의 | | 후보 표준도메인 |
| 표준코드 정의 | 표준코드 분석 | 표준코드 분석 | 표준코드 | |
| | 표준코드 후보정의 | 표준코드 후보정의 | | 후보 표준코드 |
| 표준화 확정 | 표준화 확정 | 데이터 표준확정 | 후보 표준화 내역(단어, 용어, 도메인), 표준코드 | 확정된 표준화 내역(단어, 용어, 도메인), 표준코드 |
| | 표준화 공표 | 데이터 표준 공표 | 확정된 표준화 내역(단어, 용어, 도메인), 표준코드 | 표준화 메타 데이터(표준사전, 표준코드) |
| 표준화 관리 | 표준화 적용 | 논리 데이터 모델링 | 확정된 표준화 내역(단어, 용어, 도메인) | 논리데이터모델다이어그램(ERD),<br>엔터티정의서, 애트리뷰트정의서 |
| | | 물리 데이터 모델링 | 논리데이터모델다이어그램(ERD),<br>엔터티정의서, 애트리뷰트정의서, 확정된 표준코드 | 물리데이터모델다이어그램(ERD),<br>데이터베이스정의서, 테이블정의서,컬럼정의서 |
| | 표준화 관리 | 표준 준수 여부 점검 | 논리데이터모델다이어그램(ERD) | 비표준 내역 및 변경된 표준화 내역 |
| | | 변경관리 | 비표준 내역 및 변경된 표준화 내역,<br>표준화 메타 데이터(표준사전, 표준코드),<br>논리데이터모델다이어그램(ERD),<br>엔터티정의서, 애트리뷰트정의서,<br>물리데이터모델다이어그램(ERD),<br>데이터베이스정의서, 테이블정의서, 컬럼정의서 | 표준화 메타 데이터(표준사전, 표준코드),<br>논리데이터모델다이어그램(ERD),<br>엔터티정의서, 애트리뷰트정의서,<br>물리데이터모델다이어그램(ERD),<br>데이터베이스정의서, 테이블정의서, 컬럼정의서 |

Figure 16.2. Task별 입력물과 산출물

### 16.4.3 정의 단계의 산출물은 후보이다

Phase 2~5의 산출물이 모두 “후보”라는 점에 유의하여야 한다. 새로 정의한 단어·용어·도메인·코드는 즉시 확정하지 않고 후보 상태로 먼저 등록하며, 확정은 Phase 6에서 한 번에 이루어진다. 후보 상태로도 모델링 작업을 이어 갈 수 있으므로 검토가 끝나기를 기다리며 작업이 멈추지 않는다. 사전의 각 항목은 후보·확정·폐기의 세 상태를 가지며, 폐기는 행을 삭제하는 것이 아니라 현재 시점에 사용을 금지한다는 표시이다.

### 16.4.4 실제 작업은 논리 모델에서 시작된다

프로세스 표는 표준단어 정의가 표준용어 정의보다 앞서고, 데이터 표준화 정의가 표준화 적용보다 앞서는 것처럼 보인다. 그러나 실무에서 아무런 기준 없이 표준단어부터 등재하는 경우는 드물다. 무엇을 등재할지는 작성할 모델이 결정하기 때문이다.

실제로는 논리 모델을 작성하다가 필요한 용어를 확인하고, 그 용어를 정의하기 위하여 단어·도메인·코드를 조회하며, 사전에 없는 항목은 그 시점에 후보로 등재한다. 사전을 먼저 완성한 뒤에 모델링을 시작하면 사용되지 않을 단어까지 등재하게 된다. 프로세스 표의 차례는 **산출물이 확정되는 논리적 차례**이지 작업자가 손을 대는 시간적 차례가 아니다.

> 실무 팁: 모델이 사전을 요구하게 하라
>
> 표준 사전 구축을 별도의 선행 프로젝트로 떼어 내지 않는다. 모델링을 진행하면서 필요한 용어를 그때 등재하는 방식이 사용되지 않는 단어의 누적을 막는다. 다만 이 방식이 작동하려면 후보 등재와 확정의 절차가 빠르게 순환하여야 한다.

## 16.5 KDSM 명명 규칙

### 16.5.1 네 개의 사전

KDSM의 명명은 네 개의 사전으로 결정된다.

Table 16.5. KDSM의 네 사전

| 사전 | 역할 | 주요 관리 항목 |
|---|---|---|
| 표준단어 | 이름의 최소 단위. 한글명에 영문명과 영문약어를 연결한다 | 단어명, 영문명, 영문약어, 단어구분(일반어·도메인어·분류어), 이음동의어, 금칙어, 상태 |
| 표준용어 | 표준단어를 조합한 이름. 엔터티명·속성명으로 쓰이며 컬럼명이 되는 단위이다 | 용어명, 구성 단어, 물리명, 영문명, 표준도메인, 표준코드, 용어구분, 상태 |
| 표준도메인 | 데이터 타입과 길이를 결정한다. 도메인그룹 → 도메인분류 → 도메인의 3계층이다 | 도메인명, 도메인그룹, 도메인분류, 데이터타입, 길이, 소수점 자릿수, 상태 |
| 표준코드 | 값이 정해진 목록에서만 선택되는 항목의 허용값을 정의한다. 코드마스터와 코드상세로 구성된다 | 코드명, 코드 아이디, 코드 유형, 자릿수, 코드값, 코드값명, 정렬 순서, 상태 |

### 16.5.2 이름이 결정되는 과정

컬럼 이름 `고객여신금액` 하나가 결정되는 과정을 따라가면 네 사전의 관계가 드러난다.

1. **단어를 선택한다.** 표준단어 사전에서 `고객`, `여신`, `금액`을 선택한다. 사전에 없는 낱말이 있으면 그 낱말을 후보 표준단어로 먼저 등재한다.
2. **용어를 구성한다.** 세 단어를 연결하면 표준용어 `고객여신금액`이 된다. 물리명은 각 단어의 영문약어를 밑줄로 연결하여 계산하고, 영문명은 영문 전체명을 공백으로 연결한다.
3. **형식이 결정된다.** 맨 뒤의 분류어 `금액`이 도메인분류 `금액`을 가리키고, 그 아래의 표준도메인 `금액N15`가 `NUMERIC(15)`를 결정한다. 설계자가 15자리를 임의로 지정하는 것이 아니라 분류어를 선택하는 순간 결정된다.
4. **필요하면 코드를 연결한다.** `고객등급코드`처럼 값이 정해진 항목은 표준코드를 함께 참조한다.

물리명은 작성자가 짐작하여 기재하는 값이 아니라 사전에서 계산되는 값이다. `고객주문일자`의 물리명을 흔히 쓰는 약어로 `CUST_ORD_DT`라고 짐작하기 쉬우나, 사전에 등재된 약어가 `ORDR`과 `YMD`라면 실제 물리명은 `CUST_ORDR_YMD`이다. 통상적으로 쓰이는 약어와 사전에 등재된 약어는 다를 수 있으므로, 물리명은 반드시 사전에서 계산하여 확인한다.

### 16.5.3 표준용어의 구성 문법

표준용어는 표준단어 사전에 등록된 단어로만 구성한다. 사전에 없는 단어가 필요하면 표준단어의 등재가 먼저 이루어져야 한다. 용어를 만들면서 일정이 급하다는 이유로 단어의 등재를 건너뛰는 관행이 사전 붕괴의 시작이다.

용어는 데이터의 소속이나 대상을 나타내는 **주제어**, 의미를 한정하는 **수식어**, 그리고 맨 뒤에 하나만 오는 **분류어**로 이루어진다. 분류어는 값의 형식적 성격을 나타내며, `명` · `번호` · `일자` · `금액` · `코드` · `여부` · `수` · `율`이 대표적이다. 분류어가 맨 뒤에 고정되어 있어야 이름만으로 값의 성격을 판단할 수 있다. `고객여신금액`은 금액이고 `고객여신일자`는 날짜이다. 사람이 그렇게 판단하듯 프로그램도 같은 방법으로 판단할 수 있다.

Table 16.6. 혼동하기 쉬운 분류어

| 분류어 | 사용 기준 | 혼동 주의 |
|---|---|---|
| 코드 | 표준코드 사전에 정의된 코드 체계의 값을 담을 때 | 자유 입력 문자열에 '코드'를 쓰지 않는다 |
| 번호 | 업무 규칙에 따라 부여되는 식별 번호 | 연산 대상이 아니므로 문자 타입이다 |
| 일자 · 일시 | 날짜만 필요하면 일자, 시각까지 필요하면 일시 | 정밀도는 요건이 확정한다 |
| 여부 | 참·거짓 두 값의 상태 | 상태가 셋 이상이면 여부가 아니라 상태코드이다 |

### 16.5.4 표준도메인의 명명

표준도메인은 도메인그룹 → 도메인분류 → 도메인의 3계층으로 관리한다. 도메인그룹은 분류어 단위의 큰 구분이고, 도메인분류는 그룹을 세분한 중간 구분이며, 도메인은 타입과 길이 단위이다. 도메인명은 **도메인분류명 + 데이터 타입 약어 + 길이**로 짓는다. `금액N15`는 도메인분류 `금액`에 속하는 `NUMERIC(15)` 도메인이고, `가격N9,2`는 소수 둘째 자리까지의 가격이다. 이름만으로 소속과 물리 형식을 읽을 수 있어야 하기 때문이다.

도메인의 개수는 의식적으로 억제한다. 길이가 조금 다르다는 이유로 도메인을 늘리면 도메인이 사실상 컬럼별 정의로 퇴화한다. 명칭류와 내용류처럼 대표 길이 몇 단계로 수렴시키고, 새 길이 요구는 기존 단계로의 흡수를 먼저 검토한다.

### 16.5.5 금지어와 폐기어

표준이 아닌 낱말을 쓰지 못하게 하는 장치도 사전에 함께 둔다. 같은 뜻이지만 표준이 아닌 낱말은 **이음동의어**로, 표준단어로 등재해서는 안 되는 낱말은 **금칙어**로 기록한다. 폐기한 표준단어에는 대신 사용할 **대체어**를 연결한다. 폐기된 단어·이음동의어·금칙어가 모여 사용해서는 안 되는 낱말 전체, 즉 **금지어**가 되며, 금지어가 들어간 이름을 입력하면 대체어로 바꾼 후보를 함께 제시받는다.

### 16.5.6 모델 객체의 이름

엔터티명과 속성명은 표준용어의 한글명을, 테이블명과 컬럼명은 표준단어 영문약어의 조합을 사용한다. 물리명은 영문 대문자·숫자·밑줄만 사용하고 DBMS 예약어를 쓰지 않는다. 따옴표로 감싸지 않은 식별자를 DBMS마다 대문자 또는 소문자로 바꾸어 저장하므로, 물리명을 대문자로 통일하고 인용 식별자를 만들지 않으면 여러 DBMS에서 같은 이름으로 동작한다.

테이블명 앞에 붙는 접두어는 별도로 다룬다. 예를 들어 `MT_ENTTY`의 `MT_`는 표준용어가 아니라 그 테이블이 어느 영역에 속하는지를 나타내는 모델링 규약이다. 접두어를 사전에 담으면 "물리명은 영문약어의 조합"이라는 규칙이 성립하지 않으므로, 사전은 `ENTTY`까지만 알고 접두어는 검증 규칙이 따로 받는다.

## 16.6 산출물 표준

### 16.6.1 산출물의 체계

KDSM의 산출물 템플릿은 표준 및 절차 매뉴얼과 짝을 이룬다. 매뉴얼이 각 Task를 어떻게 수행하는지를 규정한다면, 템플릿은 그 결과를 어떤 양식에 담는지를 규정한다. 프로세스의 산출물과 템플릿의 대응은 다음과 같다.

Table 16.7. 프로세스 산출물과 템플릿의 대응

| Phase | 산출물 | 산출물 템플릿 | 제공 형태 |
|---|---|---|---|
| 1. 요구사항분석 | 표준화 대상·범위·방법 | 표준화 계획서 | 파일 |
| 1. 요구사항분석 | 표준화 대상 데이터 형상 | 표준화 대상 데이터 정의서 | 파일 |
| 2~5. 데이터 표준화 정의 | 후보 표준단어·용어·도메인·코드 | 표준단어·표준용어·표준도메인·표준코드 정의서 | 파일·DB |
| 6. 표준화 확정·공표 | 표준화 메타데이터 | 표준 확정·공표 대장 | DB |
| 7. 표준화 적용 | 논리 데이터 모델 | ERD, 주제영역·주제영역구성·엔터티·속성 정의서 | 파일 |
| 7. 표준화 적용 | 물리 데이터 모델 | 데이터베이스·테이블·컬럼 정의서 | 파일 |
| 7. 표준화 관리 | 비표준·변경 내역 | 표준 준수 점검 결과서, 변경관리 대장 | 파일 |

템플릿은 정의서·계획서·대장처럼 작성자가 항목을 채우는 **파일형**과, 표준사전처럼 데이터베이스 테이블로 관리하는 **DB형**으로 나뉜다. 두 형태는 항목 구성이 같으므로 어느 매체로 구현하든 같은 템플릿을 적용할 수 있다.

## 16.6.2 세 가지 정의 산출물

정의 산출물은 대상 영역에 따라 세 종류로 묶인다. 데이터 표준, 논리 모델, 물리 모델이다. 각 산출물은 하나의 엑셀 파일이며, 그 안의 시트가 개별 정의서이다.

Table 16.8. DM-DST 정의 산출물 3종

| 문서번호 | 산출물 | 수록하는 정의서 |
|---|---|---|
| DM-DST-001 | 데이터 표준 정의 | 표준단어 · 표준용어 · 표준도메인(도메인그룹·도메인분류 포함) · 표준코드(코드마스터·코드값 상세) 정의서 |
| DM-DST-002 | 논리 데이터 모델 | 주제영역 · 주제영역구성 · 엔터티 · 속성 정의서 |
| DM-DST-003 | 물리 데이터 모델 | 데이터베이스 · 테이블 · 컬럼 정의서 |

산출물의 파일명은 `문서번호_산출물명-대상_V버전_작성일.xlsx` 형식이다. 예를 들어 `DM-DST-002_논리데이터모델-고객관리_V0.9.0_20260818.xlsx`는 고객관리 대상의 논리 데이터 모델 산출물이다. 첫 시트인 표지에는 문서번호·버전·작성일·작성자를 기록하며, 파일명의 버전·작성일과 표지의 값이 일치하여야 한다. 값이 바뀌면 오타 하나를 고쳐도 버전을 올린다. 그래야 "같은 버전 번호는 같은 내용"이 성립하여 버전 번호만으로 어떤 내용을 들고 있는지 구별할 수 있다.

## 16.6.3 논리와 물리를 잇는 기록

제4장은 논리 모델과 물리 모델의 대응을 매핑 정의서로 관리할 것을 권한다. KDSM은 별도의 매핑 정의서를 두지 않고 테이블 정의서의 **관련 엔터티** 항목에 그 테이블이 어느 엔터티에서 구현되었는지를

기록한다. 두 엔터티를 하나의 테이블로 통합한 경우에는 관련 엔터티를 모두 기재한다. 이 항목이 논리 모델과 물리 모델을 연결하는 근거가 되며, 검증 규칙은 이 연결을 따라 엔터티와 테이블, 속성과 컬럼을 대조한다.

식별관계와 비식별관계도 같은 원칙으로 기록한다. 속성 정의서에 기본키 순번과 참조 엔터티·참조 속성을 기재하면, 부모의 기본키가 자식의 기본키 일부가 되는지에 따라 두 관계가 구별된다. ERD는 이 기록에서 생성되므로 관계선을 따로 그리지 않는다.

## 16.7 검토와 승인

### 16.7.1 표준의 확정과 공표

후보 표준은 Phase 6에서 확정되고 공표된다. 확정은 표준 관리 조직이 후보를 검토하여 승인하는 것이고, 공표는 확정된 표준을 표준사전과 표준코드로 등록하여 조직 전체가 참조할 수 있게 하는 것이다. 공표된 뒤의 변경은 반드시 변경관리를 거쳐 사전·모델·정의서에 함께 반영한다.

가이드의 조항은 예외 없이 지켜야 하는 규칙과 특별한 사유가 없는 한 따라야 하는 지침으로 나뉜다. 규칙을 위반한 후보는 반려되고, 지침의 예외는 사유를 기록하여 승인을 받는다. 승인된 예외도 누적하여 관리한다. 같은 예외가 반복하여 승인되면 그 예외를 계속 승인할 것이 아니라 표준의 개정을 발의한다.

### 16.7.2 규칙에 의한 검증

KDSM은 산출물의 표준 준수 여부를 104종의 검증 규칙으로 검사한다. 규칙은 검사 대상에 따라 다음과 같이 나뉜다.

Table 16.9. 검증 규칙의 구성

| 검사 대상 | 규칙 수 | 검사하는 것의 예 |
|---|---|---|
| 데이터 표준(단어·용어·도메인·도메인그룹과 분류·코드) | 38 | 영문약어의 유일성, 물리명과 영문약어 조합의 일치, 분류어 종결 |
| 메타파일 헤더 | 4 | 시트의 헤더 구성 |
| 논리 모델(엔터티·속성·주제영역·주제영역 배정) | 28 | 이름의 표준용어 등재 여부, 참조 대상의 존재 |
| 물리 모델(컬럼·테이블·데이터베이스) | 16 | 컬럼 타입과 표준도메인의 일치, 기본키의 구성 |
| 논리 ↔ 물리 대조 | 16 | 엔터티와 테이블, 속성과 컬럼의 이름·기본키·Null 허용·참조의 일치 |
| 신규 등재분 ↔ 기존 사전 대조 | 2 | 새 항목과 기존 항목의 충돌 |

검증 결과에는 세 단계의 심각도가 있다. ERROR는 반드시 고쳐야 할 위반, WARNING은 검토가 필요한 사항, INFO는 필요한 참조 자료가 없어 판정하지 않았다는 기록이다. INFO를 따로 두는 것은 검사하지 않은 항목을 통과한 것으로 보고하지 않기 위해서이다. 통과한 규칙의 수만으로 안전을 판단해서는 안 된다.

검증에는 기준선(baseline)의 개념도 쓰인다. 기존 사전에 이미 있던 위반과 이번 산출물이 새로 유발한 위반을 구분하는 것이다. 기존 사전만으로 검사한 결과를 기준선으로 삼아 차감하면, 검토자는 이번 변경이 무엇을 어긋나게 하는지만 확인할 수 있다.

### 16.7.3 사전의 변경과 모델의 변경

표준 사전과 데이터 모델은 변경 경로가 다르다. 사전은 조직 전체가 공유하는 기준이므로 변경하려면 제안과 판정을 거친다. 제안 권한이 있는 사용자가 등록·수정·폐기를 제안하고, 표준승인 권한이 있는 사용자가 승인하거나 반려한다. 승인은 사전 반영과 검토 이력, 변경이력의 기록을 하나의 트랜잭션으로 수행하므로 "승인되었으나 반영되지 않은" 중간 상태가 없다.

반면 모델은 그 기준을 사용하는 결과물이므로 모델편집 권한으로 즉시 저장한다. 다만 저장하기 전에 검증 규칙으로 검사하고, 위반이 하나라도 있으면 표 전체를 저장하지 않는다. 일부만 반영되면 화면에 보이는 것과 저장된 것이 달라지기 때문이다.

Table 16.10. 사전과 모델의 변경 경로

| 구분 | 표준 사전 | 데이터 모델 |
|---|---|---|
| 성격 | 조직의 합의를 요구하는 기준 | 기준을 적용한 결과물 |
| 변경 방법 | 제안 → 판정(승인·반려) | 표 편집 후 저장 |
| 요구 권한 | 제안 · 표준승인 | 모델편집 |
| 사전 검사 | 제안이 새로 유발하는 위반만 표시 | 위반이 있으면 저장 거부 |
| 기록 | 검토 이력과 변경이력 | 수정 시각에 의한 동시 편집 판별 |

## 16.8 지원 도구

KDSM 데이터 모델링 방법론의 Support 계층에는 파이썬 라이브러리 `kdsm`과 웹 애플리케이션이 있다. 라이브러리는 커맨드라인 명령으로도 제공되며, 웹 애플리케이션은 판단을 따로 구현하지 않고 라이브러리를 호출한다. 두 도구가 다루는 것은 데이터 자체가 아니라 데이터의 구조와 이름에 관한 정보, 즉 메타데이터이다.

### 16.8.1 라이브러리와 커맨드라인

Table 16.11. kdsm 명령과 프로세스 단계

| 명령 | 하는 일 | 지원하는 프로세스 단계 |
|---|---|---|
| `suggest` | 한글 용어명을 표준단어 단위로 분해하여 표준용어 후보를 제시한다. 사전에 없는 낱말은 LLM에게 등재 초안을 받을 수 있다 | 표준단어·표준용어 정의 |
| `validate` | 산출물과 사전의 표준 준수 여부를 검증 규칙으로 검사한다 | 정의 단계 전체, 표준 준수 여부 점검 |
| `ddl` | 물리 데이터 모델 산출물에서 DDL을 생성한다 | 물리 데이터 모델링 |
| `erd` | 산출물·메타테이블·데이터베이스 카탈로그에서 ERD를 생성한다 | 논리·물리 데이터 모델링 |
| `load` · `apply` | 메타파일과 산출물을 메타테이블에 적재한다 | 표준화 공표, 표준화 적용 |
| `export` | 메타테이블을 읽어 산출물을 다시 생성한다 | 변경관리 |
| `drift` | 정의서와 운영 데이터베이스의 차이를 대조한다 | 표준 준수 여부 점검 |
| `diagnose` | 운영 데이터베이스의 표준 준수 수준을 비율과 등급으로 측정한다 | 표준 준수 여부 점검 |

라이브러리는 두 방향으로 쓰인다. 신규 구축 시점에는 **문서에서 데이터베이스로**, 즉 산출물을 검증하고 DDL을 생성하여 스키마를 만들고 메타데이터를 적재한다. 운영이 시작된 뒤에는 **데이터베이스에서 문서로**, 즉 운영 데이터베이스를 정의서와 대조하고 표준 준수 수준을 진단하며 산출물과 ERD를 다시 생성한다. 운영 전환 이후에는 기준이 문서에서 데이터베이스로 바뀌므로, 문서를 손으로 맞추지 않고 메타테이블에서 다시 생성한다. 이렇게 생성한 산출물은 버전 `0.0.0`, 작성자 공란으로 기록되어 아직 사람이 검토하지 않았음을 나타낸다.

검증과 진단은 서로 다른 질문에 답한다. 검증은 산출물을 작성하는 시점에 통과와 실패를 판정하고, 대조는 데이터베이스에 반영한 뒤 정의서와의 차이를 찾으며, 진단은 이미 운영 중인 데이터베이스가 표준을 얼마나 따르는지를 비율로 산출한다. 표준 사전이 없는 스키마도 진단할 수 있으며, 이때는 스키마 내부에서 서로 일치하지 않는 부분을 찾는다.

### 16.8.2 웹 애플리케이션

웹 애플리케이션은 같은 기능을 조직의 여러 사용자가 화면에서 쓰도록 제공한다. 표준 사전 조회, 용어 추천과 제안, 제안의 판정과 변경이력, 주제영역·논리 모델·물리 모델의 표 편집, 논리 모델에서 물리 모델의 생성, DDL 생성, ERD, 산출물의 생성과 적재, 검증·진단·대조가 화면으로 제공된다. 기능마다 조회·제안·모델편집·표준승인·시스템관리의 다섯 권한 가운데 필요한 것을 요구하며, 권한 사이에는 상하 관계가 없다.

라이브러리와 웹 애플리케이션의 사용법은 각각 「데이터 표준화 라이브러리 사용자 매뉴얼」(유충현 (2026a))과 「데이터 표준화 애플리케이션 사용자 매뉴얼」(유충현 (2026b))이 다룬다.

> ⚠ 주의: 도구가 판단을 대신하지 않는다
>
> `suggest`가 제시하는 후보는 사전의 단어로 분해할 수 있는 조합일 뿐, 그 조합이 업무상 적절한지는 판단하지 않는다. LLM이 만든 등재 초안도 상태는 언제나 후보이다. 도구의 출력은 검토의 재료이며, 등재와 확정은 사람이 정한다.

## 16.9 사례 연구

### 16.9.1 사례: KDSM 자신에게 KDSM을 적용하다

KDSM의 라이브러리와 웹 애플리케이션은 그 자체로 데이터베이스를 사용하는 시스템이다. 이 두 시스템의 데이터 구조는 KDSM 데이터 모델링 방법론으로 설계되었다. 방법론이 스스로를 설명할 수 있는지를 확인하는 가장 직접적인 방법이 방법론을 자기 자신에게 적용하는 것이기 때문이다.

#### 16.9.1.1 배경

KDSM에는 두 개의 스키마가 있다. `kdsm_meta`는 표준 사전과 데이터 모델 정의를 담는 메타테이블의 스키마이고, `kdsm_app`은 웹 애플리케이션이 사용자·제안·작업처럼 화면에서 만들어지는 데이터를 담는 스키마이다. 두 스키마 모두 정의서를 먼저 작성하고, 정의서에서 DDL을 생성하여 만든다.

#### 16.9.1.2 접근 방법

**메타테이블.** `kdsm_meta`의 14개 메타테이블은 표준단어·표준용어·표준도메인(그룹·분류 포함)·표준코드(마스터·상세)의 데이터 표준 영역, 주제영역·주제영역구성·엔터티·속성의 논리 모델 영역, 데이터베이스·테이블·컬럼의 물리 모델 영역으로 나뉜다. 이 구조는 시스템 관리자가 작성하는 DM-LIB 산출물 6종에 정의되어 있다. DM-LIB은 세 영역을 논리 정의서와 물리 정의서로 한 번 더 나누어 001~003과 004~006이 짝을 이루며, 같은 사실을 두 문서에 기록하므로 서로 대조한다.

**애플리케이션 테이블.** `kdsm_app`은 사용자가 작성하는 것과 같은 DM-DST 산출물 3종으로 정의되었다. 논리 모델에는 조직·사용자·사용자권한·제안·제안검토·변경이력·작업·작업산출물·접속대상의 9개 엔터티가 「kdsm 애플리케이션」이라는 하나의 주제영역에 배정되어 있고, 물리 모델에는 이에 대응하는 9개 테이블이 있다. 권한·제안대상·제안유형·제안상태·검토결과·작업종류·작업상태의 7개 표준코드는 데이터 표준 정의 산출물에 정의되어 있다.

두 스키마의 테이블은 각각 `MT_`와 `AP_` 접두어를 단다. 이 접두어는 앞서 설명한 대로 표준용어가 아니라 모델링 규약이므로 사전에 담지 않는다. 산출물을 검증하고 적재할 때에는 허용할 접두어를 검증 규칙에 따로 알려 준다.

#### 16.9.1.3 얻은 교훈

자기 적용 과정에서 확인된 사항은 일반 조직에도 그대로 적용된다.

1. **접두어는 사전 밖에 둔다.** 접두어를 알려 주지 않고 애플리케이션 산출물을 검증하면, 규칙은 표준용어에서 계산한 물리명(예: `OGNZ`)을 기대하므로 `AP_OGNZ` 같은 엔터티 영문명을 위반으로 보고한다. 접두어를 표준용어에 넣어 해결하면 "물리명은 영문약어의 조합"이라는 원칙이 무너진다. 규약은 규칙의 설정으로, 표준은 사전으로 관리하여야 둘이 섞이지 않는다.
2. **적재는 검증을 통과한 것만 반영한다.** 적재는 반영하기 전에 검증을 수행하고, 위반이 하나라도 있으면 아무것도 반영하지 않는다. 적재는 하나의 트랜잭션으로 수행되므로 일부만 반영된 상태가 남지 않는다.
3. **애플리케이션이 참조하는 코드는 확정 상태여야 한다.** 웹 애플리케이션은 표준코드 가운데 확정 상태의 코드만 읽는다. 후보 상태로 적재된 코드는 제외되므로, 필요한 코드명이 없으면 서버가 기동 시점에 중단한다. 후보와 확정의 구분이 문서상의 표시가 아니라 시스템의 동작을 결정한다.

4. **코드값이 아니라 코드명을 기준점으로 삼는다.** 애플리케이션은 코드값이 아니라 코드명으로 코드를 찾는다. 코드값을 다시 부여하여도 애플리케이션을 수정할 필요가 없고, 코드명을 바꾸면 기동 시점에 즉시 드러난다. 어긋남이 운영 중에 조용히 남지 않게 하는 설계이다.
5. **하나의 메타테이블에 여러 모델을 담는다.** 메타테이블에는 KDSM 자신의 모델과 애플리케이션의 모델이 함께 적재된다. ERD를 생성할 때 스키마를 선택하여 한쪽만 그리면 되므로, 모델마다 별도의 저장소를 둘 필요가 없다.

#### 16.9.1.4 성과

두 스키마의 구조는 정의서에서 출발하여 검증 → DDL 생성 → 적재의 경로로 만들어지며, 적재된 정의는 다시 ERD와 산출물로 생성된다. 정의서와 데이터베이스가 같은 사실을 각각 기록하지만, 그 사이를 사람이 아니라 검증 규칙과 생성기가 잇는다. 이것이 KDSM 데이터 모델링 방법론이 일반 조직에 제안하는 작업 방식과 같다.

### 요약

1. 이름과 형식을 매번 다시 정하지 않는다
   - 사람은 표준단어와 표준도메인을 정하고, 물리명·영문명·데이터 타입과 길이는 그 결정에서 도출된다. 한 번 정한 것을 여러 곳에 반복하여 기록하지 않는다.
2. 7 Phase · 14 Activity · 18 Task
   - 요구사항분석 → 데이터 표준화 정의 → 확정 → 적용의 흐름이며, 정의 단계의 산출물은 후보이다. 실제 작업은 논리 모델에서 시작되어 필요한 용어를 그때 등재한다.
3. 네 사전과 용어 문법
   - 표준단어·표준용어·표준도메인·표준코드가 이름과 형식을 결정한다. 분류어는 용어의 맨 뒤에 하나만 오며 값의 성격과 도메인을 결정한다.
4. 세 가지 정의 산출물
   - 데이터 표준(DM-DST-001), 논리 모델(DM-DST-002), 물리 모델(DM-DST-003)로 나뉘며, 테이블 정의서의 관련 엔터티가 논리와 물리를 잇는다.
5. 규칙에 의한 검증과 이원화된 변경 경로
   - 104종의 규칙이 산출물과 사전을 검사한다. 사전은 제안과 판정을 거쳐, 모델은 검증을 통과한 저장으로 바뀐다.

## 다음 학습 가이드

### 16.9.2 관련 본서 참조 장

- 제 4장 데이터 모델링 표준 — 데이터 모델링의 이론적 기반
- 제 14장 국제 표준 심화 — ISO/IEC 11179 명명 규칙
- 제 15장 디플래닉스 방법론 개요 — KDSM 전체 개관
- 제 17장 KDSM 메타데이터 관리 — 모델링 메타데이터 관리

### 16.9.3 KDSM 자료

- KANDLE 데이터 표준화 방법론 포털의 데이터 모델링 방법론 — 프로세스, 표준 및 절차 매뉴얼, 산출물 템플릿, 가이드, 테크니컬 노트의 정본
- 데이터 표준화 라이브러리 사용자 매뉴얼(유충현 (2026a)) — 네 사전의 작성 기준, 검증 규칙, `kdsm` 명령의 사용법
- 데이터 표준화 애플리케이션 사용자 매뉴얼(유충현 (2026b)) — 화면에서 수행하는 제안·판정·모델 편집·점검

## 연습 문제

1. 여러분 조직의 핵심 엔터티 하나를 선택하여, 그 엔터티의 속성 다섯 개에 KDSM 명명 규칙을 적용하시오. 각 속성의 구성 단어와 분류어, 적용할 표준도메인을 밝히고, 사전에 새로 등재하여야 할 표준단어가 있으면 함께 제시하시오.
2. KDSM 프로세스의 표에서는 표준단어 정의가 논리 데이터 모델링보다 앞서지만, 실제 작업은 논리 모델에서 시작된다고 하였다. 두 설명이 모순되지 않는 이유를 "논리적 차례"와 "시간적 차례"의 구분으로 설명하시오.
3. 다음 두 이름의 문제점을 KDSM의 분류어 선택 기준에 따라 지적하고 고쳐 쓰시오: (a) 승인·미승인·승인대기의 세 값을 갖는 `승인여부`, (b) 코드 체계와 연결되지 않은 자유 입력 문자열 `고객구분`.
4. KDSM은 표준 사전의 변경에는 제안과 판정을 요구하면서 데이터 모델의 변경은 모델편집 권한으로 즉시 저장하게 한다. 이렇게 경로를 나눈 근거를 설명하고, 모델 변경에도 승인 절차를 두어야 하는 조직이 있다면 그 조건은 무엇인지 논하시오.
5. 사례 연구에서 테이블 접두어를 표준용어에 넣지 않고 검증 규칙의 설정으로 관리하였다. 여러분 조직에서 사전에 담아야 할 것과 규칙의 설정으로 관리하여야 할 것을 각각 두 가지씩 들고 그 기준을 제시하시오.

CHAPTER 17

# KDSM 메타데이터 관리

> **i** 집필 예정 챕터
>
> 본 챕터는 디플래닉스 KDSM의 메타데이터 관리 방법론을 다루는 챕터입니다. 상세 내용은 추후 집필 예정입니다.

## 17.1 개요

KDSM 메타데이터 관리는 국내 조직이 메타데이터를 효과적으로 수집·관리·활용하기 위한 실무 방법론이다. 제 5장 메타데이터 관리의 이론과 ISO/IEC (2023b) 국제 표준을 기반으로 국내 조직 환경에 맞는 접근을 제공한다.

## 17.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- KDSM 메타데이터의 계층 구조를 이해한다
- ISO/IEC 11179 기반 메타데이터 등록소를 설계할 수 있다
- 메타데이터 수집·관리 프로세스를 운영할 수 있다
- 메타데이터 활용을 통한 데이터 거버넌스를 실현할 수 있다

## 17.3 KDSM 메타데이터 원칙

## 17.4 KDSM 메타데이터 등록소

## 17.5 KDSM 메타데이터 수집·관리 프로세스

## 17.6 KDSM 메타데이터 활용

## 17.7 도구와 인프라

## 17.8 사례 연구

### 요약

### 다음 학습 가이드

#### 17.8.1 관련 본서 참조 장

- 제 5장 메타데이터 관리 — 메타데이터 관리의 이론적 기반
- 제 14장 국제 표준 심화 — ISO/IEC 11179 심층
- 제 15장 디플래닉스 KDSM 개요 — KDSM 전체 개관
- 제 16장 KDSM 데이터 모델링 — 모델과 메타데이터의 관계
- 제 18장 KDSM 데이터 품질 — 품질 메타데이터

### 연습 문제

1. (예정) 여러분 조직의 핵심 데이터 요소 5개에 대해 ISO/IEC 11179 기반 KDSM 메타데이터 명세서를 작성하시오.

CHAPTER 18

# KDSM 데이터 품질 관리

> **i** 집필 예정 챕터
>
> 본 챕터는 디플래닉스 KDSM의 데이터 품질 관리 방법론을 다루는 챕터입니다. 상세 내용은 추후 집필 예정입니다.

## 18.1 개요

KDSM 데이터 품질 관리는 국내 조직이 데이터 품질을 체계적으로 측정·관리·개선하기 위한 실무 방법론이다. 제 6장 데이터 품질 관리의 이론과 ISO (2022b), ISO/IEC (2008) 국제 표준을 기반으로 국내 조직 환경에 맞는 접근을 제공한다.

## 18.2 본 장의 학습 목표

본 장을 학습한 후 독자는 다음을 수행할 수 있어야 한다.

- KDSM 데이터 품질 프레임워크를 이해한다
- 조직에 적합한 품질 차원과 KPI를 설계할 수 있다
- 데이터 품질 측정·모니터링 체계를 구축할 수 있다
- 품질 개선 프로세스를 운영할 수 있다

## 18.3 KDSM 품질 관리 원칙

## 18.4 KDSM 품질 차원

## 18.5 KDSM 품질 측정 체계

## 18.6 KDSM 품질 관리 프로세스 (PDCA)

## 18.7 데이터 품질 이슈 대응

## 18.8 도구와 인프라

## 18.9 사례 연구

### 요약

### 다음 학습 가이드

#### 18.9.1 관련 본서 참조 장

- 제 6장 데이터 품질 관리 — 데이터 품질의 이론적 기반
- 제 14장 국제 표준 심화 — ISO 8000, ISO/IEC 25012 심층
- 제 15장 디플래닉스 KDSM 개요 — KDSM 전체 개관
- 제 16장 KDSM 데이터 모델링 — 모델링 단계의 품질
- 제 17장 KDSM 메타데이터 — 품질 메타데이터 관리

### 연습 문제

1. (예정) 여러분 조직의 핵심 데이터셋에 대해 KDSM 품질 KPI를 설계하시오.

# 용어 사전

이 부록은 본서에서 사용된 데이터 표준화 관련 주요 용어를 가나다순으로 정리한 종합 사전이다. 각 용어는 실무에서 자주 혼동되거나 정확한 이해가 필요한 것들 중심으로 선정되었으며, 관련 챕터를 병기하여 심화 학습이 가능하도록 하였다.

용어는 다음 원칙에 따라 정리되었다:

- **한국어 표기가 널리 사용되는 경우**: 한국어를 표제어로, 영문을 병기
- **영문 약어가 더 널리 사용되는 경우**: 영문 약어를 표제어로 배치 (별도 섹션)
- **동의어·유사어**: 대표어에 정의하고 별칭에 참조 표시
- **국제 표준의 공식 정의가 있는 경우**: 해당 표준을 인용

## 가

**가용성 (Availability)**
데이터가 필요할 때 접근·이용 가능한 정도. ISO/IEC (2008) 의 시스템 종속 특성 중 하나. (→ 제 14장)

**개념 도메인 (Conceptual Domain)**
ISO/IEC (2023c) 메타모델에서 값의 의미적 범위를 정의하는 요소. 예: '성인 나이 범위'는 개념 도메인이며, '18~150 정수'는 이를 실현한 값 도메인(Value Domain). (→ 제 14장)

**개념 모델 (Conceptual Data Model)**
데이터의 비즈니스 의미를 추상적으로 표현하는 최상위 모델. 엔티티와 관계를 정의하되 기술적 세부사항은 배제. (→ 제 4장)

**개념 드리프트 (Concept Drift)**
시간에 따라 입력 데이터와 타겟 변수 사이의 관계 자체가 변화하는 현상. 데이터 드리프트의 세 유형(공변량·개념·라벨) 중 가장 심각. (→ 제 12장)

**개인정보보호법 (Personal Information Protection Act, PIPA)**
한국의 개인정보 보호에 관한 기본법. 2020년 개정으로 가명·익명처리 표준, 정보주체 권리 강화. (→ 제 1장, 제 12장)

**거버넌스 (Governance)**
조직의 자산·활동을 효과적으로 관리하기 위한 정책·프로세스·역할·책임의 체계. 데이터 거버넌스는 데이터 자산에 대한 거버넌스이다. (→ 제 9장)

**결정적 매칭 (Deterministic Matching)**
완전 일치 규칙에 기반한 엔티티 해소 방법. 예: SSN 동일 시 동일 인물로 판정. 정확하나 커버리지가 낮음. (→ 제 7장)

**계보 (→ 데이터 계보)**

**골든 레코드 (Golden Record)**
특정 실체(예: 고객 홍길동)에 대해 조직이 참조하는 유일하고 신뢰할 수 있는 표준 표현. MDM의 핵심 산출물. 식별→통합→정제→풍부화의 4단계로 생성. (→ 제 7장)

**공변량 드리프트 (Covariate Drift)**
입력 데이터의 분포가 변화하지만 입력-출력 관계는 유지되는 데이터 드리프트 유형. PSI로 측정. (→ 제 12장)

**공정성 (Fairness)**
AI 모델 출력이 보호 속성(성별·인종·연령 등)에 따라 차별적이지 않은지의 평가. 통계적 동등성, 기회 균등, 예측 동등성, Calibration 등 여러 정의가 존재하며, Kleinberg et al. (2017) 의 '불가능성 정리'에 따라 이들 정의를 동시에 만족할 수 없음. (→ 제 12장)

**공존 방식 (→ Coexistence)**

**관측성 (→ 데이터 관측성)**

**규제 (Regulation)**
정부가 부과하는 법적 요구사항. 데이터 관련 주요 규제로 GDPR, CCPA, 개인정보보호법, Basel Committee on Banking Supervision (2013), European Parliament (2024) 등. (→ 제 1장)

**균형성 (Balance)**
AI 학습 데이터에서 클래스·세그먼트별 적절한 비율을 유지하는 특성. 학습 데이터 5대 표준화 차원 중 하나. (→ 제 12장)

**기밀성 (Confidentiality)**
권한 있는 사용자만 데이터에 접근할 수 있도록 하는 특성. ISO/IEC (2008) 품질 특성 중 하나. (→ 제 14장)

## 나

**날짜 형식 표준 (Date Format Standard)**
날짜·시간 데이터의 통일된 표현 형식. 국제 표준 ISO 8601 (YYYY-MM-DD) 이 대표적. (→ 제 3장)

**내부 감사 (Internal Audit)**
조직이 자체적으로 관리 시스템의 실효성을 평가하는 활동. ISO/IEC (2023a) 등 인증 대상 표준의 필수 요구사항. (→ 제 9장, 제 14장)

## 다

**다중 클라우드 (Multi-Cloud)**
하나의 조직이 여러 클라우드 서비스 제공자(AWS·Azure·GCP 등)를 동시에 사용하는 전략. 엔터프라이즈의 89%가 채택하고 있으며, 데이터 표준화의 새 과제. (→ 제 13장)

**다크 데이터 (Dark Data)**
조직이 수집·저장하고 있으나 다시 활용되지 못하는 데이터. IBM 조사에서 조직 전체 데이터의 약 80% 차지. 표준화 부재가 주된 원인. (→ 제 1장)

**대표성 (Representativeness)**
학습 데이터가 타겟 모집단의 통계적 분포를 반영하는 정도. AI 학습 데이터의 5대 표준화 차원 중 하나. (→ 제 12장)

**데이터 (Data)**
사실·측정·관찰의 원시적 표현. DIKW 계층의 최하위 요소. (→ 제 2장)

**데이터 거버넌스 (Data Governance)**
데이터를 조직의 전략적 자산으로 관리하기 위한 정책·프로세스·조직 체계의 총체. Ladley (2019)는 이를 데이터 관리의 '헌법'에 비유한다. (→ 제 9장)

**데이터 계보 (Data Lineage)**
데이터가 원천에서 최종 사용까지 이동하는 경로와 변환 이력. 개념적/논리적/물리적 계보의 3유형이 있으며, LF AI & Data Foundation (2024) 이 오픈 표준을 제공. (→ 제 5장, 제 13장)

**데이터 계약 (Data Contract)**
데이터 생산자와 소비자 간의 명시적 약속. 스키마·의미·SLA·품질 기준·호환성 정책·오너십의 6개 요소로 구성. 마이크로서비스 API 계약과 유사한 개념. (→ 제 2장, 제 13장)

**데이터 계층 (Data Tier)**
클라우드 스토리지의 접근 빈도별 계층(Hot/Warm/Cold). FinOps 실무의 핵심. (→ 제 13장)

**데이터 관측성 (Data Observability)**
데이터 파이프라인의 상태·품질·성능을 지속적으로 모니터링하는 접근. 이상 탐지·SLA 관리 도구가 이 영역에 속함. (→ 제 2장, 제 12장)

**데이터 도메인 (Data Domain)**
(1) 특정 데이터 요소가 가질 수 있는 허용 가능한 값의 집합. (2) 조직 내 데이터의 주요 관심 영역(고객·제품·재무 등). (→ 제 2장, 제 7장)

**데이터 딕셔너리 (Data Dictionary)**
데이터베이스·정보 시스템에서 사용되는 데이터 요소들의 정의·형식·관계를 문서화한 참조 자료. (→ 제 5장)

**데이터 라이프사이클 (Data Lifecycle)**
데이터의 생성(Create) → 저장(Store) → 사용(Use) → 공유(Share) → 아카이빙(Archive) → 폐기(Destroy)의 전체 수명 주기. (→ 제 2장)

**데이터 레이크 (Data Lake)**
비정형·반정형·정형 데이터를 원시 형태로 저장하는 저장소. 2010년대 초 등장한 2세대 데이터 아키텍처. (→ 제 13장)

**데이터 레이크하우스 (Data Lakehouse)**
데이터 웨어하우스와 데이터 레이크의 장점을 결합한 아키텍처. 저비용 오브젝트 스토리지 위에 ACID 트랜잭션·스키마 관리·성능 최적화 제공. Armbrust et al. (2021) 이 창시. (→ 제 13장)

**데이터 마이그레이션 (Data Migration)**
데이터를 한 시스템·형식·환경에서 다른 곳으로 이전하는 활동. (→ 제 8장)

**데이터 마트 (Data Mart)**
특정 부서·주제에 특화된 소규모 데이터 웨어하우스. (→ 제 4장)

**데이터 메시 (Data Mesh)**
Dehghani (2022) 가 제안한 도메인 중심 분산 데이터 관리 패러다임. 4대 원칙(도메인 오너십·Data as Product·셀프서비스 인프라·연합 컴퓨팅 거버넌스)에 기반. (→ 제 2장, 제 13장)

**데이터 모델 (Data Model)**
데이터의 구조·관계·제약 조건을 형식적으로 표현한 추상적 모델. 개념·논리·물리의 3계층으로 구분. (→ 제 4장)

**데이터 모델링 (Data Modeling)**
비즈니스 요구사항을 데이터 구조로 변환하는 활동. Chen (1976) 의 ER 모델이 이론적 기반. (→ 제 4장)

**데이터 사전 (→ 데이터 딕셔너리)**

**데이터 스튜어드 (Data Steward)**
특정 데이터 도메인의 일상적 품질 관리·표준 적용을 담당하는 실무 책임자. Business·Technical·Operational 3유형으로 구분. (→ 제 7장, 제 9장)

**데이터 스튜어드십 (Data Stewardship)**
Plotkin (2020) 은 이를 '조직이 데이터를 자산으로 관리하기 위해 데이터 사용자들이 수행하는 활동'이라 정의. (→ 제 9장)

**데이터 오너 (Data Owner)**
특정 데이터 도메인의 비즈니스 결정 권한을 가진 부서장/임원. '무엇을(what)'을 결정. (→ 제 9장)

**데이터 요소 (Data Element)**
ISO/IEC (2023b) 의 표준에 따라 정의된 데이터의 기본 단위. Object Class + Property + Representation의 3부분 구조. (→ 제 5장, 제 14장)

**데이터 요소 개념 (Data Element Concept, DEC)**
ISO/IEC (2023c) 메타모델의 핵심 요소. 표현 방식(정수·문자열·코드 등)과 독립된 데이터의 의미 단위. 시맨틱 상호운용성의 이론적 기반. (→ 제 5장, 제 14장)

**데이터 웨어하우스 (Data Warehouse, DW)**
통합·주제 지향·비휘발성·시간 종속적 데이터 저장소. Inmon (2005) 와 Kimball and Ross (2013) 이 서로 다른 접근법을 제시. (→ 제 4장)

**데이터 자산 (Data Asset)**
조직에 가치를 창출하는 데이터. 자산으로서 관리·평가·투자 대상이 됨. (→ 제 1장)

**데이터 자산 분류 (Data Classification)**
데이터를 민감도·중요도·규제 요건에 따라 등급화(공개/내부/기밀/극비)하는 활동. (→ 제 9장)

**데이터 주권 (Data Sovereignty)**
데이터가 저장·처리되는 국가·리전의 법률이 적용된다는 원칙. GDPR, PIPL 등이 대표적. 클라우드 아키텍처 결정에 직접 영향. (→ 제 13장)

**데이터 준수 (Data Compliance)**
조직의 데이터 활동이 관련 규제·표준·정책을 준수하는 상태. GDPR·개인정보보호법 등에 대한 대응이 핵심. (→ 제 9장, 제 14장)

**데이터 카스토디안 (Data Custodian)**
데이터의 기술적 관리(저장·보안·백업)를 담당하는 IT 실무자. '어디에(where)'를 담당. (→ 제 9장)

**데이터 카탈로그 (Data Catalog)**
조직 내 데이터 자산의 메타데이터를 중앙에서 관리·검색·거버넌스하는 시스템. Alation, Atlan, Collibra, DataHub 등이 대표. (→ 제 5장, 제 10장)

**데이터 통합 (Data Integration)**
여러 원천의 데이터를 통합된 뷰로 결합하는 활동. ETL, ELT, 페더레이션 등 여러 접근이 있음. (→ 제 4장)

**데이터 패브릭 (Data Fabric)**
Gartner가 제창한 분산 데이터 자산의 지능적 통합 아키텍처. AI/ML 기반 자동화 도구를 활용. (→ 제 13장)

**데이터 표준화 (Data Standardization)**
조직 내·외부에서 데이터를 공통의 정의·구조·형식·값으로 일관되게 관리하기 위한 정책·프로세스·기술의 총체. 데이터의 '공통 언어' 확립. (→ 제 1장)

**데이터 품질 (Data Quality)**
데이터가 의도된 목적에 적합한 정도. 6대 차원(완전성·정확성·일관성·유효성·유일성·적시성)으로 평가. (→ 제 6장)

**데이터셋 카드 (Dataset Card, Datasheets for Datasets)**
Gebru et al. (2021) 이 제안한 AI 학습 데이터셋의 표준 문서 양식. 수집 동기·구성·프로세스·편향·유지 등 9개 섹션. (→ 제 12장)

**데이터셋 시프트 (→ 데이터 드리프트)**

**데이터 드리프트 (Data Drift)**
시간 흐름에 따라 데이터 분포가 학습 시점과 달라지는 현상. AI 모델 성능 저하의 주요 원인. 공변량·개념·라벨 드리프트로 구분. (→ 제 12장)

**DIKW 계층 (Data-Information-Knowledge-Wisdom)**
Ackoff (1989) 이 제안한 데이터 → 정보 → 지식 → 지혜의 계층 모델. 데이터 관리 이론의 기초. (→ 제 2장)

## 라

**라벨 일관성 (Label Consistency)**
AI 학습 데이터의 라벨이 어노테이터·시점에 무관하게 일관된 정도. IAA(Inter-Annotator Agreement)로 측정. (→ 제 12장)

**라이프사이클 (→ 데이터 라이프사이클)**

**로우 데이터 (→ 원시 데이터)**

**리스크 데이터 (Risk Data)**
금융 기관이 리스크 관리·보고에 사용하는 데이터. Basel Committee on Banking Supervision (2013) 가 표준을 규정. (→ 제 3장, 제 11장)

## 마

**마스터 데이터 (Master Data)**
조직의 비즈니스를 구성하는 핵심 실체(고객·제품·조직·위치·계정 등)에 대한 데이터. 여러 시스템에서 반복적으로 참조되는 안정적 정보. (→ 제 7장)

**마스터 데이터 관리 (Master Data Management, MDM)**
마스터 데이터를 통합·표준화·관리하는 활동. Registry/Consolidation/Coexistence/Centralized의 4대 아키텍처 스타일. Loshin (2010a) 이 표준 저서. (→ 제 7장)

**마이데이터 (MyData)**
개인정보 자기결정권에 기반한 데이터 이동권 제도. 한국은 금융·의료·통신 분야로 확대. (→ 제 11장)

**매치 (Match)**
MDM에서 동일 실체 후보 레코드를 식별하는 규칙. 이름·주소·연락처 등의 유사도 기반. Precision과 Recall의 균형이 관건. (→ 제 7장)

**매치·머지 (Match-Merge)**
MDM의 핵심 프로세스. 매칭된 레코드를 병합(Merge)하여 골든 레코드를 생성하고 서바이버십 규칙으로 속성값을 결정. (→ 제 7장)

**매칭 (→ 엔티티 해소, 매치)**

**메타데이터 (Metadata)**
데이터에 관한 데이터. 데이터의 구조·의미·출처·품질 등을 설명하는 정보. 기술적(Technical)·비즈니스(Business)·운영(Operational) 3계층으로 구분. (→ 제 5장)

**메타데이터 등록소 (Metadata Registry, MDR)**
표준화된 메타데이터를 등록·관리하는 시스템. ISO/IEC (2023b) 이 국제 표준을 규정. (→ 제 5장, 제 14장)

**명명 규칙 (Naming Convention)**
데이터 요소의 이름을 부여하는 표준 규칙. ISO/IEC (2023b) Part 5가 4원칙(유일성·필요성·정확성·간결성)과 3부분 구조를 규정. (→ 제 14장)

**모델 붕괴 (Model Collapse)**
AI가 생성한 데이터가 다시 학습 데이터로 사용되어 모델 성능이 저하되는 현상. 생성형 AI 시대의 새로운 표준화 이슈. (→ 제 12장)

**모델 카드 (Model Card)**
Mitchell et al. (2019) 이 FAT* 2019 컨퍼런스에서 제안한 AI 모델 문서화의 표준 양식. 모델 상세·의도된 용도·학습 데이터·평가·성능·윤리적 고려사항·주의사항의 7개 섹션. Google, Microsoft, Meta, Hugging Face 등이 채택. (→ 제 12장)

**모던 데이터 스택 (Modern Data Stack)**
여러 전문 도구를 표준 인터페이스로 조합한 엔드투엔드 데이터 파이프라인 접근. Fivetran·Snowflake·dbt·Atlan·Monte Carlo 등의 조합이 전형적. (→ 제 10장)

**미켈란젤로 (Michelangelo)**
Uber가 2017년 발표한 ML 플랫폼. 피처 스토어 개념의 원조. (→ 제 12장)

## 바

**바젤 위원회 (Basel Committee on Banking Supervision, BCBS)**
은행 감독의 국제 표준을 개발하는 기구. Basel Committee on Banking Supervision (2013) (효과적 리스크 데이터 집계 및 리스크 보고 원칙)이 대표 산출물. (→ 제 3장)

**벤더 락인 (Vendor Lock-in)**
특정 벤더의 제품·서비스에 종속되어 다른 대안으로 전환이 어려운 상태. 오픈 표준·오픈 데이터 형식으로 완화 가능. (→ 제 10장, 제 13장)

**변화 관리 (Change Management)**
조직 변화에 대한 저항을 관리하고 정착을 촉진하는 활동. Kotter의 8단계 모델이 대표. (→ 제 8장)

**분류 체계 (Classification Scheme)**
데이터·자산을 체계적으로 분류하는 계층 구조. ISO/IEC (2023b) 메타모델의 구성 요소 중 하나. (→ 제 5장)

**비식별화 (De-identification)**
개인정보를 식별할 수 없도록 처리하는 활동. 가명처리·익명처리·차분 프라이버시 등의 기법이 있음. (→ 제 12장)

**BCBS 239**
바젤 위원회가 2013년 발표한 '효과적인 리스크 데이터 집계 및 리스크 보고 원칙'. G-SIB(글로벌 시스템적 중요 은행)에 14개 원칙 준수 요구. (→ 제 3장, 제 11장)

## 사

**산업 표준 (Industry Standard)**
특정 산업의 규범적 표준. 금융(Basel Committee on Banking Supervision (2013), ISO (2022c)), 의료(HL7 FHIR, CDISC), 제조(ISA-95) 등. 규제 준수와 상호운용성에 직결. (→ 제 3장)

**서바이버십 (Survivorship)**
MDM에서 여러 원천의 속성 중 어느 값을 골든 레코드로 채택할지 결정하는 규칙. Source-based/Recency/Completeness/Frequency/Trust score 등의 유형. (→ 제 7장)

**서브젝트 영역 (Subject Area)**
조직 데이터의 주요 관심 영역을 그룹화한 개념. 데이터 도메인과 유사하나 조직 관점에서 넓은 범위를 지칭. (→ 제 2장)

**성숙도 모델 (Maturity Model)**
조직의 특정 능력 수준을 단계별로 평가·개선하는 프레임워크. DCAM 6단계, CMMI-DMM 5단계가 대표. (→ 제 3장)

**셀프서비스 (Self-Service Data)**
비즈니스 사용자가 IT 개입 없이 데이터를 직접 활용하는 접근. 데이터 민주화의 실현 방식. (→ 제 13장)

**스키마 (Schema)**
데이터의 구조·형식·제약 조건을 정의하는 명세. 관계형 DB, JSON, Avro 등 여러 표기법이 있음. (→ 제 4장)

**스키마 레지스트리 (Schema Registry)**
스트리밍 데이터의 스키마 버전 관리·호환성 검증을 제공하는 시스템. Kafka Schema Registry 등이 대표. (→ 제 13장)

**스튜어드 (→ 데이터 스튜어드)**

**시맨틱 상호운용성 (Semantic Interoperability)**
서로 다른 시스템이 데이터의 의미를 동일하게 이해할 수 있는 능력. ISO/IEC (2023b) 의 DEC(Data Element Concept)가 이를 위한 이론적 기반 제공. (→ 제 14장)

**시맨틱 레이어 (Semantic Layer)**
기술적 데이터 구조와 비즈니스 의미를 매개하는 추상화 계층. Cube, dbt Semantic Layer 등이 대표. (→ 제 2장)

**시맨틱 인코딩 (Semantic Encoding)**
표준 용어·코드를 사용하여 데이터의 의미를 부호화하는 접근. ISO 8000-110 마스터 데이터 품질의 4대 특성 중 하나. (→ 제 14장)

**신선도 (Freshness)**
데이터가 현재 시점의 현실을 반영하는 정도. AI 학습 데이터의 5대 표준화 차원 중 하나. (→ 제 12장)

**실용 품질 (Pragmatic Quality)**
사용 목적 대비 데이터의 적합성. ISO 8000-8이 정의한 3대 품질 유형 중 하나(구문·의미·실용). (→ 제 14장)

## 아

**아키텍처 스타일 (→ MDM 아키텍처 스타일)**

**아파치 아이스버그 (Apache Iceberg)**
Netflix가 2017년 창시한 오픈 테이블 포맷. 숨겨진 파티셔닝·스키마 진화·시간 여행 기능 제공. Blue (2018) 이 소개. Snowflake, Databricks, AWS 등이 채택 확대. (→ 제 13장)

**어노테이션 (Annotation)**
AI 학습 데이터에 라벨을 부여하는 활동. 어노테이터 간 합의도(IAA, Inter-Annotator Agreement)가 품질 지표. (→ 제 12장)

**엔티티 (Entity)**
데이터 모델에서 관심 대상이 되는 실체. Chen (1976) ER 모델의 기본 요소. (→ 제 4장)

**엔티티 관계 모델 (Entity-Relationship Model, ER Model)**
Chen (1976) 이 제안한 데이터 모델링의 이론적 기반. 개체(Entity)와 관계(Relationship)로 도메인을 표현. (→ 제 4장)

**엔티티 해소 (Entity Resolution)**
여러 원천의 레코드 중 동일 실체를 가리키는 것을 식별하는 문제. 결정적/확률적/ML 기반의 3대 접근법. Christen (2012) 가 표준 저서. Fellegi-Sunter 모델이 확률적 매칭의 이론적 기반. (→ 제 7장)

**연방 학습 (Federated Learning)**
분산된 데이터를 그대로 유지하며 모델을 학습하는 접근. 개인정보 보호·데이터 주권 대응. (→ 제 12장)

**연방 피처 스토어 (Federated Feature Store)**
데이터는 각 위치에 유지하고 피처 정의·로직만 중앙화하는 피처 스토어 접근. 데이터 주권 규제 환경에서 활용. (→ 제 12장)

**오너 (→ 데이터 오너)**

**오너십 (→ 데이터 오너십)**

**오브젝트 클래스 (Object Class)**
ISO/IEC (2023c) 메타모델에서 관심 대상을 정의하는 요소. 예: '사람', '조직'. 데이터 요소 3부분 명명의 첫 요소. (→ 제 14장)

**오픈 테이블 포맷 (Open Table Format)**
오브젝트 스토리지 위에서 데이터베이스와 유사한 기능을 제공하는 오픈 표준. Apache Iceberg·Delta Lake·Apache Hudi가 3대 대표. 벤더 락인 방지의 핵심 전략. (→ 제 13장)

**옴니채널 (Omnichannel)**
여러 채널에서 일관된 고객 경험을 제공하는 접근. 통합 고객 뷰(MDM)가 필수 전제. (→ 제 7장)

**용어사전 (→ 데이터 딕셔너리, 비즈니스 용어집)**

**원시 데이터 (Raw Data)**
가공되지 않은 원본 데이터. 원시 형태 그대로 저장하는 것이 데이터 레이크의 특징. (→ 제 13장)

**유효성 (Validity)**
데이터가 정의된 형식·범위·규칙을 준수하는 정도. 데이터 품질 6대 차원 중 하나. (→ 제 6장)

**유일성 (Uniqueness)**
데이터가 중복 없이 하나만 존재하는 정도. 데이터 품질 6대 차원 중 하나이며 MDM의 핵심 KPI. (→ 제 6장, 제 7장)

**의미 상호운용성 (→ 시맨틱 상호운용성)**

**이상 탐지 (Anomaly Detection)**
통계·ML 기반으로 비정상 데이터를 자동 발견하는 접근. Monte Carlo, Anomalo, Lightup 등이 대표 도구. (→ 제 10장, 제 12장)

**이커머스 (E-commerce)**
인터넷 기반 상거래. PIM(Product Information Management)이 중요. (→ 제 7장)

**인증 (Certification)**
공인 기관이 조직의 특정 표준 준수를 공식 인정하는 절차. ISO 9001, ISO/IEC 27001, ISO/IEC (2023a) 등이 인증 대상 표준. (→ 제 14장)

**일관성 (Consistency)**
데이터가 관련된 다른 데이터와 모순이 없는 정도. 데이터 품질 6대 차원 중 하나. (→ 제 6장)

## 자

**자산 (→ 데이터 자산)**

**자연어 처리 (Natural Language Processing, NLP)**
인간 언어를 처리하는 AI 기술. LLM 시대에 사내 문서 표준화가 중요해짐. (→ 제 12장)

**재현성 (Reproducibility)**
AI 모델을 동일한 데이터·코드·환경으로 재실행하면 동일 결과가 나오는 특성. 데이터 버전·코드 버전·환경 정의 3요소 필요. (→ 제 12장)

**적시성 (Timeliness)**
데이터가 필요한 시점에 이용 가능한 정도. 데이터 품질 6대 차원 중 하나. (→ 제 6장)

**정규화 (Normalization)**
(1) DB 이론: 데이터 중복을 최소화하기 위한 관계 스키마 분해. Codd (1970) 의 1~5NF. (2) 표준화 문맥: 데이터 값을 표준 형식으로 변환. (→ 제 2장, 제 4장)

**정보 (Information)**
처리·구조화된 데이터. DIKW 계층에서 데이터의 상위 요소. (→ 제 2장)

**정확성 (Accuracy)**
데이터가 실제 세계와 일치하는 정도. 데이터 품질 6대 차원 중 하나. (→ 제 6장)

**조합 폭발 (Combinatorial Explosion)**
시스템 수 N에 대해 통합 경로가 N(N-1)/2로 증가하는 현상. 표준화 없이는 관리 불가. (→ 제 1장)

**조화 (Harmonization)**
서로 다른 표준·시스템 간의 조율을 통해 상호 호환성을 확보하는 활동. 통일(unification)과는 다른 접근. (→ 제 2장)

**중복률 (Duplication Rate)**
동일 실체가 중복 표현된 비율. MDM의 핵심 KPI. (→ 제 7장)

**지식 (Knowledge)**
정보에 맥락·경험이 결합되어 실행 가능해진 상태. DIKW 계층에서 정보의 상위 요소. (→ 제 2장)

**지혜 (Wisdom)**
지식에 판단·가치가 결합된 최상위 상태. DIKW 계층의 최상위. (→ 제 2장)

## 차

**차분 프라이버시 (Differential Privacy)**
통계적으로 개인을 식별할 수 없도록 하는 프라이버시 보호 기법. Dwork and Roth (2014) 가 이론적 기반 제공. (→ 제 12장)

**차원 모델링 (Dimensional Modeling)**
Kimball and Ross (2013) 이 제안한 데이터 웨어하우스 모델링 접근. Fact와 Dimension 테이블로 구성. (→ 제 4장)

**참조 데이터 (Reference Data)**
코드·분류·표준값 집합. 국가 코드, 통화, 성별 등. 매우 안정적이며 외부 표준을 준수. (→ 제 2장, 제 7장)

## 카

**카탈로그 (→ 데이터 카탈로그)**

**컴플라이언스 (Compliance)**
규제·표준·정책 준수. ISO/IEC (2008) 품질 특성 중 하나이자 거버넌스의 4대 활동 영역 중 하나. (→ 제 9장, 제 14장)

**클라우드 네이티브 (Cloud-Native)**
클라우드 환경에 최적화된 설계·개발·운영 접근. 자동 확장·관리형 서비스·API 우선 등의 특징. (→ 제 10장, 제 13장)

## 타

**태깅 (Tagging)**
데이터 자산에 메타데이터·분류를 부여하는 활동. AI 기반 자동 태깅이 최근 확산. (→ 제 5장, 제 10장)

**통합 관리 시스템 (Integrated Management System, IMS)**
여러 ISO 관리 시스템 표준을 하나의 정책·문서 체계로 통합 운영하는 접근. HLS(High-Level Structure) 공유로 가능. (→ 제 14장)

**트랜잭션 데이터 (Transaction Data)**
비즈니스 이벤트·활동 기록. 주문·결제·재고 이동 등. 빈번한 생성, 시간 종속적. (→ 제 2장, 제 7장)

## 파

**파케이 (Parquet)**
열 지향 데이터 파일 형식. Apache Iceberg·Delta Lake 등 오픈 테이블 포맷의 기반. (→ 제 13장)

**파티션 프루닝 (Partition Pruning)**
불필요한 파티션 스캔을 회피하여 쿼리 비용을 절감하는 기법. FinOps 실무. (→ 제 13장)

**편향 (Bias)**
데이터·모델의 체계적 오류. 역사적·표본·측정·라벨·집계 편향의 5유형. AI 시대의 새로운 품질 차원. (→ 제 12장)

**표준 (Standard)**
조직·산업·국가·국제 차원의 공식적 합의. 공식 국제 표준(ISO)·산업 표준·기업 표준으로 분류. (→ 제 2장)

**표준화 (Standardization)**
표준을 개발·적용·유지하는 활동. (→ 제 1장, 제 2장)

**표준화 프레임워크 (Standardization Framework)**
데이터 표준화를 조직 차원에서 체계적으로 추진하기 위한 뼈대. DAMA-DMBOK2, ISO 표준 시리즈 등. (→ 제 3장)

**품질 (→ 데이터 품질)**

**품질 차원 (Quality Dimensions)**
데이터 품질을 측정·관리하는 관점. DAMA 6대 차원(완전성·정확성·일관성·유효성·유일성·적시성), ISO/IEC (2008) 15개 특성 등. (→ 제 6장)

**프로파일링 (→ 데이터 프로파일링)**

**피처 (Feature)**
원시 데이터로부터 모델이 학습 가능한 형태로 변환된 입력 변수. AI 시대의 표준화 대상. (→ 제 12장)

**피처 스토어 (Feature Store)**
피처의 표준화·재사용·거버넌스를 지원하는 인프라. Feast, Tecton, Databricks Feature Store 등이 대표. Uber의 Michelangelo가 원조. (→ 제 12장)

## 하

**하이브리드 클라우드 (Hybrid Cloud)**
온프레미스와 퍼블릭 클라우드를 결합한 IT 인프라. 데이터 주권·규제 대응에 활용. (→ 제 13장)

**학습 데이터 (Training Data)**
AI 모델 학습에 사용되는 데이터. 5대 표준화 차원(대표성·균형성·신선도·라벨 일관성·출처 추적성). (→ 제 12장)

**합의도 (Inter-Annotator Agreement, IAA)**
여러 어노테이터가 동일 데이터에 대해 얼마나 일치된 라벨을 부여하는지의 지표. Cohen's Kappa, Fleiss' Kappa 등으로 측정. (→ 제 12장)

**확률적 매칭 (Probabilistic Matching)**
통계적 가중치에 기반한 엔티티 해소 방법. Fellegi and Sunter (1969) 의 Fellegi-Sunter 모델이 이론적 기반. Splink, Zingg 등이 현대 구현체. (→ 제 7장)

**환영 (Hallucination)**
LLM이 사실이 아닌 정보를 생성하는 현상. 학습 데이터 표준화와 RAG로 완화. (→ 제 12장)

**High-Level Structure (HLS)**
ISO 관리 시스템 표준의 공통 10개 조항 구조. ISO/IEC (2024) 로 정의. ISO 9001, 27001, 42001 등이 공유. 통합 관리 시스템(IMS)의 기반. (→ 제 14장)

# 영문 약어 사전

**AI Act**
EU Artificial Intelligence Act. 2024년 발효된 유럽연합의 AI 규제법. 고위험 AI 시스템에 학습 데이터 품질·거버넌스·문서화 요구. (→ 제 1장, 제 12장)

**AIMS (AI Management System)**
AI 관리 시스템. ISO/IEC (2023a) 이 국제 표준. (→ 제 14장)

**Annex SL**
ISO/IEC Directives Part 1의 부속서. 관리 시스템 표준의 공통 구조(HLS) 정의. (→ 제 14장)

**API (Application Programming Interface)**
시스템 간 상호작용을 위한 인터페이스. REST API가 대표. (→ 제 10장)

**AWS (Amazon Web Services)**
아마존의 클라우드 서비스. S3, Glue, SageMaker 등의 데이터 서비스 제공. (→ 제 13장)

**Azure**
마이크로소프트의 클라우드 서비스. Purview, Synapse, Fabric 등의 데이터 서비스 제공. (→ 제 13장)

**BCBS 239**
→ 바젤 위원회 (사전 본문 참조)

**BI (Business Intelligence)**
데이터 기반 비즈니스 의사결정 지원 도구·프로세스. Tableau·Power BI·Looker 등이 대표. (→ 제 10장)

**CCPA (California Consumer Privacy Act)**
미국 캘리포니아주의 소비자 개인정보 보호법. 2020년 발효. (→ 제 1장)

**CDISC (Clinical Data Interchange Standards Consortium)**
임상시험 데이터 표준화 기구. CDASH·SDTM·ADaM·SEND 표준. FDA가 신약 승인 시 준수 요구. (→ 제 3장)

**CDO (Chief Data Officer)**
최고 데이터 책임자. 전사 데이터 전략을 총괄하는 임원. 포춘 1000대 기업의 82%가 임명. (→ 제 9장)

**CI/CD (Continuous Integration/Continuous Deployment)**
소프트웨어 개발의 자동 통합·배포 파이프라인. DataOps에도 적용. (→ 제 10장)

**CMMI-DMM (Data Management Maturity Model)**
CMMI Institute의 데이터 관리 성숙도 모델. 5단계 성숙도(Performed→Optimized). (→ 제 3장)

**DAMA (Data Management Association International)**
데이터 관리 전문가 국제 협회. 1985년 설립. DMBOK 발간, CDMP 자격증 운영. (→ 제 3장)

**DAMA-DMBOK2**
DAMA Data Management Body of Knowledge 2판. 데이터 관리의 사실상 국제 표준. 11개 지식 영역 정의. (→ 제 3장)

**DataOps**
데이터 파이프라인 자동화·통합·오케스트레이션 접근. Cataldo and Foote (2022) 가 표준 저서. (→ 제 10장)

dbt (data build tool)
: SQL 기반 데이터 변환 도구. 모던 데이터 스택의 핵심. (→ 제 10장, 제 13장)

DCAM (Data Management Capability Assessment Model)
: EDM Council의 데이터 관리 역량 평가 모델. 6단계 성숙도. 금융 산업에서 널리 활용. (→ 제 3장)

DEC (Data Element Concept)
: ISO/IEC (2023c) 메타모델의 핵심. 표현 방식과 독립된 의미 단위. 시맨틱 상호운용성의 기반. (→ 제 5장, 제 14장)

Delta Lake
: Databricks가 창시한 오픈 테이블 포맷. Linux Foundation(Delta)로 이관. (→ 제 13장)

DMBOK (→ DAMA-DMBOK2)

ELT (Extract, Load, Transform)
: 원천에서 추출·적재 후 대상에서 변환하는 데이터 통합 접근. 클라우드 데이터 웨어하우스 시대의 표준. (→ 제 10장)

ER Model (→ 엔티티 관계 모델)

ETL (Extract, Transform, Load)
: 원천에서 추출·변환하여 대상에 적재하는 전통적 데이터 통합 접근. (→ 제 4장)

EU AI Act (→ AI Act)

FHIR (Fast Healthcare Interoperability Resources)
: HL7의 최신 의료 정보 교환 표준. REST API·JSON 기반. 미국 21st Century Cures Act, EU EHDS가 채택. (→ 제 3장)

FinOps (Financial Operations)
: 클라우드 재무 관리 실무. 6원칙 프레임워크 제공. (→ 제 13장)

GCP (Google Cloud Platform)
: 구글의 클라우드 서비스. BigQuery, Data Catalog, Vertex AI 등의 데이터 서비스 제공. (→ 제 13장)

GDPR (General Data Protection Regulation)
: EU 개인정보 보호 규정. 2018년 발효. 매출의 최대 4% 벌금. (→ 제 1장)

G-SIB (Globally Systemically Important Bank)
: 글로벌 시스템적 중요 은행. Basel Committee on Banking Supervision (2013) 준수 대상. (→ 제 11장)

HL7 (Health Level Seven)
: 의료 정보 교환 표준화 기구. FHIR이 최신 표준. (→ 제 3장)

HLS (High-Level Structure)
: ISO 관리 시스템 표준의 공통 구조. ISO/IEC (2024) 이 정의. 통합 관리 시스템의 기반. (→ 제 14장)

IAA (Inter-Annotator Agreement)
: 어노테이터 간 합의도. AI 학습 데이터의 라벨 일관성 측정 지표. (→ 제 12장)

IAF (International Accreditation Forum)
: 국제인증포럼. ISO/IEC 42001 등 관리 시스템 표준 인증의 지침 제공. (→ 제 14장)

IEC (International Electrotechnical Commission)
: 국제전기기술위원회. ISO와 함께 정보기술 표준 발간. (→ 제 14장)

IMS (Integrated Management System)
: 통합 관리 시스템. 여러 ISO 관리 시스템 표준을 HLS 기반으로 통합 운영. (→ 제 14장)

IoT (Internet of Things)
: 사물 인터넷. 대량의 스트리밍 데이터를 생성하며 데이터 관리의 새 도전. (→ 제 3장)

ISA-95
: 기업 정보 시스템과 제어 시스템 통합의 국제 표준. IEC 62264로도 발간. 스마트 팩토리 기반. (→ 제 3장)

ISO (International Organization for Standardization)
: 국제표준화기구. 전 세계 표준의 최상위 개발 기구. (→ 제 14장)

ISO 8000
: 데이터 품질 국제 표준 시리즈. 30개 이상의 파트로 구성. (→ 제 6장, 제 14장)

ISO 8000-8
: 정보·데이터 품질 개념. 구문·의미·실용 품질의 3대 유형 정의. (→ 제 14장)

ISO 8000-61
: 데이터 품질 관리 프로세스 표준. Deming PDCA 사이클을 적용. 4대 활동 (Planning·Control·Assurance·Improvement). (→ 제 14장)

ISO 8000-110
: 마스터 데이터 품질 요구사항. 4대 특성(구문 준수·의미 인코딩·데이터 출처·정확성). (→ 제 14장)

ISO 20022
: 금융 메시지 국제 표준. SWIFT MT의 후속. XML/JSON 기반. (→ 제 3장)

ISO/IEC 11179
: 메타데이터 등록소(MDR) 국제 표준. 6개 파트로 구성. 2023년 최신 판. (→ 제 5장, 제 14장)

ISO/IEC 25012
: 데이터 품질 모델 국제 표준. SQuaRE 시리즈의 일부. 15개 품질 특성 정의(고유 5·고유+시스템 종속 7·시스템 종속 3). (→ 제 6장, 제 14장)

ISO/IEC 25024
: 데이터 품질 측정 국제 표준. ISO/IEC (2008) 특성의 정량적 측정 지표 제공. (→ 제 14장)

ISO/IEC 27001
: 정보 보안 관리 시스템 국제 표준. (→ 제 9장, 제 14장)

ISO/IEC 42001
: AI 관리 시스템(AIMS) 국제 표준. 2023년 발효. HLS 구조 + 39개 Annex A 통제(9범주). (→ 제 12장, 제 14장)

JTC 1 (Joint Technical Committee 1)
: ISO/IEC 정보기술 공동 위원회. SC 32(데이터), SC 40(IT 거버넌스), SC 42(AI)가 데이터 관리 관련 표준을 개발. (→ 제 14장)

KATS (Korean Agency for Technology and Standards)
: 국가기술표준원. 대한민국 산업통상자원부 산하 표준 기관. (→ 제 14장)

KPI (Key Performance Indicator)
: 핵심 성과 지표. 거버넌스·품질·MDM의 성과 관리에 활용. (→ 제 9장)

KS (Korean Industrial Standards)
: 한국 산업 표준. KATS가 개발·관리. (→ 제 14장)

**LLM (Large Language Model)**
대규모 언어 모델. ChatGPT, Claude 등. 학습 데이터 표준화와 RAG의 중요성 부각. (→ 제 12장)

**MDM (→ 마스터 데이터 관리)**

**MDM Architecture Styles**
MDM 시스템의 4대 아키텍처 스타일. (1) Registry — 원천 유지, 인덱스만 관리. (2) Consolidation — 허브에 통합, 원천 피드백 없음. (3) Coexistence — 양방향 동기화. (4) Centralized — 허브에서만 CUD. (→ 제 7장)

**MDR (→ 메타데이터 등록소)**

**ML (Machine Learning)**
머신러닝. 데이터로부터 학습하는 AI의 하위 분야. (→ 제 12장)

**MLOps (Machine Learning Operations)**
ML 시스템의 개발·배포·운영을 자동화하는 실무. DevOps의 ML 확장. (→ 제 12장)

**MVS (Minimum Viable Standard)**
최소 실용 표준. 완벽한 표준을 처음부터 만들지 않고 최소한으로 출발하여 실전에서 진화시키는 접근. (→ 제 12장, 제 8장)

**NIA (National Information Society Agency)**
한국지능정보사회진흥원. 국내 데이터·AI 정책 지원 기관. (→ 제 11장)

**NIST AI RMF (AI Risk Management Framework)**
미국 국립표준기술연구소의 AI 리스크 관리 프레임워크. (→ 제 12장)

**OpenLineage**
LF AI & Data Foundation의 데이터 계보 오픈 표준. Job·Run·Dataset·Facet 4요소로 구성. Airflow·dbt·Spark 등 주요 도구가 채택. (→ 제 13장)

**PDCA (Plan-Do-Check-Act)**
Deming의 지속적 개선 사이클. ISO (2016) 데이터 품질 관리에 적용. (→ 제 6장, 제 14장)

**PIM (Product Information Management)**
제품 정보 관리. MDM의 특화 영역. Stibo Systems, Riversand, Pimcore 등이 대표 도구. (→ 제 7장)

**PIPL (Personal Information Protection Law)**
중국의 개인정보 보호법. 데이터 국외 반출 통제. (→ 제 13장)

**PoC (Proof of Concept)**
개념 증명. 도구 선정 전 실제 환경에서의 성능·적합성 검증. (→ 제 10장)

**Provenance**
데이터의 출처와 변형 이력. AI 학습 데이터의 5대 표준화 차원 중 하나. ISO 8000-110 마스터 데이터 품질의 4대 특성 중 하나. (→ 제 12장, 제 14장)

**PSI (Population Stability Index)**
데이터 분포 안정성 지표. PSI > 0.2 시 드리프트 경고. (→ 제 12장)

**RACI (Responsible, Accountable, Consulted, Informed)**
역할·책임 매트릭스. 데이터 거버넌스의 R&R 명확화 도구. (→ 제 9장)

**RAG (Retrieval-Augmented Generation)**
검색 증강 생성. LLM이 외부 지식을 참조하여 답변하는 기법. (→ 제 12장)

ROI (Return on Investment)
: 투자 수익률. 데이터 표준화 프로젝트의 경영진 소통에 필수. (→ 제 1장)

RPO/RTO (Recovery Point/Time Objective)
: 재해 복구의 데이터 손실 허용 시점/복구 소요 시간 목표. (→ 제 10장)

SaaS (Software as a Service)
: 소프트웨어 서비스형. 조직당 평균 130~200개 사용. 표준화 도전의 주요 원인. (→ 제 1장)

SC 32 / 40 / 42
: ISO/IEC JTC 1의 소위원회. SC 32(Data management and interchange), SC 40(IT service management and governance), SC 42(Artificial intelligence). 데이터 관리 관련 국제 표준을 개발. (→ 제 14장)

SDTM (Study Data Tabulation Model)
: CDISC의 임상시험 데이터 표준 모델. FDA 신약 승인 필수. (→ 제 3장)

Schrems II
: EU-US 개인정보 이전에 관한 2020년 EU 사법재판소 판결. EDPB가 후속 가이드라인을 발간. (→ 제 13장)

SLA (Service Level Agreement)
: 서비스 수준 협약. 데이터 계약의 핵심 요소. (→ 제 13장)

SME (Subject Matter Expert)
: 주제 전문가. 데이터 정의·품질 규칙 수립 시 자문. (→ 제 5장)

SQuaRE (Systems and software Quality Requirements and Evaluation)
: ISO/IEC 25000 시리즈. ISO/IEC (2008) (품질 모델), ISO/IEC (2015b) (측정) 등 포함. (→ 제 14장)

SR (Schema Registry)
: 스키마 레지스트리. 스트리밍 데이터의 스키마 버전·호환성 관리 시스템. Confluent Schema Registry가 대표. (→ 제 13장)

Value Domain
: ISO/IEC (2023c) 메타모델에서 허용되는 값의 집합. 예: '0~150 정수'. 데이터 요소의 실현 형태. (→ 제 14장)

W3C (World Wide Web Consortium)
: 웹 표준화 기구. RDF, SPARQL, DCAT 등 시맨틱 웹 표준 개발. (→ 제 14장)

---

## 참고 사항

이 용어사전은 본서에서 사용된 용어의 참조 사전으로, 각 용어의 최신 상세 정의는 다음 출처를 확인하기 바란다:

- **국제 표준 원문**: ISO/IEC (2023b), ISO (2022b), ISO/IEC (2008), ISO/IEC (2015b), ISO/IEC (2023a), ISO/IEC (2024) 등 ISO 표준 원문
- **DAMA 용어집**: DAMA International (2020). The DAMA Dictionary of Data Management (2nd ed.)
- **한국 표준**: e나라표준인증 (https://www.standard.go.kr) — 국내 KS 표준 검색
- **국내 정책 자료**: 국가기술표준원(KATS), 한국지능정보사회진흥원(NIA), 한국데이터산업진흥원(K-DATA)의 발간 자료

## 용어사전 활용 가이드

1. **본서 학습 시**: 각 용어의 정의 뒤에 병기된 관련 챕터로 이동하여 심화 학습이 가능하다.
2. **인용 자료 확인 시**: `@` 로 시작하는 인용 키는 참고문헌에서 상세 출처를 확인할 수 있다.
3. **조직 내 용어집 구축 시**: 본 사전을 조직 표준 용어의 기초 자료로 활용할 수 있으며, 조직 고유의 정의를 추가하여 확장 가능하다.

용어의 정의는 시간에 따라 진화하며, 특히 AI·클라우드 관련 용어는 급속히 확장되고 있다. 조직 내 용어집을 구축·유지할 때 이러한 진화를 지속적으로 반영해야 한다. 특히 다음 영역은 향후 5년간 크게 확장될 것으로 예상된다:

- **AI·ML 관련**: 학습 데이터 표준, Model Card 확장, Foundation Model 거버넌스
- **클라우드 네이티브**: 오픈 테이블 포맷, 데이터 계약, FinOps
- **규제**: EU AI Act 세부 지침, 각국의 AI 관리 시스템 인증 확산
- **국제 표준**: ISO/IEC 5259 시리즈(AI 학습 데이터 품질), ISO/IEC 23894(AI 리스크 관리) 등의 확산

# 표준화 템플릿 모음

이 부록은 데이터 표준화 실무에서 바로 활용할 수 있는 템플릿을 제공합니다. 각 템플릿은 조직의 상황에 맞게 항목을 가감하여 사용할 수 있습니다.

> 템플릿 활용 안내
>
> 아래 템플릿의 표와 양식은 그대로 복사하여 조직의 표준 문서로 활용할 수 있습니다. 괄호 `( )` 안의 내용은 작성 가이드이므로 실제 작성 시 해당 내용으로 대체하세요.

## 데이터 요소 정의서 템플릿

데이터 요소 하나하나의 표준 정의를 기록하는 양식입니다. ISO/IEC 11179 메타데이터 등록소의 기본 속성을 반영합니다.

| 항목 | 내용 |
|---|---|
| **데이터 요소명** | (한글명) |
| **영문명** | (영문 식별자, snake_case) |
| **정의** | (비즈니스 관점의 명확한 정의) |
| **데이터 타입** | (VARCHAR, INTEGER, DATE 등) |
| **길이/정밀도** | (예: VARCHAR(100), NUMBER(10,2)) |
| **허용값 범위** | (도메인 또는 참조 코드) |
| **필수 여부** | (필수 / 선택) |
| **기본값** | (해당 시) |
| **데이터 소유자** | (담당 부서 및 담당자) |
| **출처 시스템** | (원천 시스템명) |
| **관련 표준** | (관련 ISO 표준 등) |

| 항목 | 내용 |
| --- | --- |
| 변경 이력 | (버전, 날짜, 변경 내용) |

## 코드 정의서 템플릿

코드 그룹과 각 코드값의 의미를 표준화하여 기록하는 양식입니다.

**코드 그룹 정보**

- 코드 그룹 ID: (예: CD_GENDER)
- 코드 그룹명: (예: 성별 코드)
- 설명: (코드 그룹의 용도와 의미)
- 사용 시스템: (이 코드를 참조하는 시스템 목록)

**코드값 목록**

| 코드값 | 코드명 | 영문명 | 설명 | 유효여부 |
| --- | --- | --- | --- | --- |
| (값) | (한글명) | (영문명) | (의미) | Y/N |

## 데이터 품질 규칙 정의서 템플릿

데이터 품질을 측정·관리하기 위한 규칙을 정의하는 양식입니다.

| 항목 | 내용 |
| --- | --- |
| 규칙 ID | DQR-[시스템]-[번호] |
| 규칙명 | (규칙의 이름) |
| 대상 테이블/컬럼 | (적용 대상) |
| 품질 차원 | 정확성 / 완전성 / 일관성 / 적시성 |
| 규칙 내용 | (SQL 또는 자연어로 기술) |
| 심각도 | 높음 / 중간 / 낮음 |
| 측정 주기 | 실시간 / 일 / 주 / 월 |
| 담당자 | (담당 부서 및 담당자) |

## 표준화 추진 계획서 템플릿

조직 차원의 데이터 표준화 프로젝트를 기획할 때 사용하는 계획서 양식입니다. 아래 목차 구조를 바탕으로 각 절의 내용을 작성합니다.

1. **추진 배경 및 목적** — 표준화가 필요한 배경과 프로젝트의 목표를 기술합니다.
2. **현황 분석**
   - 2.1 현행 데이터 현황 — 현재 데이터 관리 상태를 진단합니다.
   - 2.2 문제점 분석 — 표준화 부재로 인한 문제점을 도출합니다.
3. **표준화 범위**
   - 3.1 대상 시스템 — 표준화를 적용할 시스템을 정의합니다.
   - 3.2 대상 데이터 — 표준화 대상 데이터의 범위를 정의합니다.

### 4. 추진 방안

- 4.1 표준화 원칙 — 표준화의 기본 원칙을 수립합니다.
- 4.2 단계별 추진 계획 — 실행 단계를 구체화합니다.

### 5. 추진 일정 — 아래 표에 단계별 일정을 기록합니다.

| 단계 | 작업 내용 | 기간 | 담당 |
|---|---|---|---|
| (단계) | (작업 내용) | (기간) | (담당) |

### 6. 기대 효과 — 표준화를 통해 얻을 정량적·정성적 효과를 기술합니다.

### 7. 리스크 및 대응 방안 — 예상 리스크와 그 대응 방안을 정리합니다.

# 실무 체크리스트

## 데이터 표준화 준비도 진단 체크리스트

| 번호 | 진단 항목 | 1(미흡) | 2 | 3(보통) | 4 | 5(우수) |
|---|---|---|---|---|---|---|
| 1 | 데이터 표준화 정책 및 지침이 문서화되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | 데이터 표준 관리 조직(데이터 오너, 스튜어드)이 지정되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | 전사 데이터 용어집(Business Glossary)이 관리되고 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | 데이터 요소에 대한 표준 정의서가 존재한다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | 코드 표준이 정의되고 중앙에서 관리된다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | 날짜/시간 형식 표준이 전사적으로 적용된다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | 데이터 품질 측정 지표(KPI)가 정의되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | 정기적인 데이터 품질 측정이 이루어진다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | 마스터 데이터 관리 체계가 수립되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | 메타데이터가 중앙 카탈로그에서 관리된다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | 데이터 계보(Lineage)가 추적된다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | 데이터 표준 변경 관리 프로세스가 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | 시스템 간 데이터 연계 표준이 정의되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | 개인정보 관련 데이터 표준이 수립되어 있다 | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | 표준화 성과를 정기적으로 보고한다 | ☐ | ☐ | ☐ | ☐ | ☐ |

**총점 계산**: ____ / 75점

Table 0.1. 준비도 점수 해석

| 점수 범위 | 성숙도 수준 | 권고 사항 |
|---|---|---|
| 60~75 | 최적화 단계 | 지속적 개선 및 혁신 |
| 45~59 | 관리 단계 | 프로세스 강화 |
| 30~44 | 정의 단계 | 표준화 체계 수립 |
| 15~29 | 초기 단계 | 기반 구축 필요 |
| 15 미만 | 부재 단계 | 즉각적인 조치 필요 |

# 데이터 표준화 구현 체크리스트

## 1단계: 현황 분석

- ☐ 현행 데이터 목록 수집 완료
- ☐ 데이터 소유자 및 관련자 식별 완료
- ☐ 현행 표준화 수준 진단 완료
- ☐ 문제점 및 개선 기회 도출 완료

## 2단계: 표준화 설계

- ☐ 표준화 원칙 및 정책 수립
- ☐ 데이터 요소 표준 정의 완료
- ☐ 코드 표준 정의 완료
- ☐ 품질 규칙 정의 완료
- ☐ 메타데이터 표준 정의 완료

## 3단계: 시스템 적용

- ☐ 메타데이터 레지스트리 구축
- ☐ 표준화 내용 시스템 반영
- ☐ 데이터 품질 측정 도구 적용
- ☐ 테스트 및 검증 완료

## 4단계: 거버넌스 수립

- ☐ 데이터 거버넌스 조직 구성
- ☐ 표준 변경 관리 프로세스 수립
- ☐ 정기 점검 체계 수립
- ☐ 교육 및 홍보 계획 수립

## 5단계: 운영 및 개선

- ☐ 표준화 성과 측정 지표 모니터링
- ☐ 정기 표준 검토 및 업데이트
- ☐ 이해관계자 만족도 조사
- ☐ 지속적 개선 사항 도출

# 참고문헌


Ackoff, Russell L. 1989. “From Data to Wisdom.” Journal of Applied Systems Analysis 16: 3–9.

Andreessen Horowitz. 2023. Emerging Architectures for Modern Data Infrastructure. https://a16z.com/emerging-architectures-modern-data-infrastructure/.

ANSI/X3/SPARC. 1975. “Interim Report: The ANSI/SPARC Three-Schema Architecture.” ACM SIGMOD Record 7 (2).

Armbrust, Michael, Ali Ghodsi, Reynold Xin, and Matei Zaharia. 2021. “Lakehouse: A New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics.” Conference on Innovative Data Systems Research (CIDR). https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf.

AWS. 2024. Multi-Cloud Data Strategy Whitepaper. Amazon Web Services. https://aws.amazon.com/whitepapers/multicloud-data-strategy.

Basel Committee on Banking Supervision. 2013. Principles for Effective Risk Data Aggregation and Risk Reporting. BCBS 239. Bank for International Settlements.

Berson, Alex, and Larry Dubov. 2011. Master Data Management and Data Governance. 2nd ed. McGraw-Hill.

Bilgen, Mehmet Tugay. 2024. Building a Modern Data Catalog: From Theory to Implementation. Packt Publishing.

Blue, Ryan. 2018. Introducing Apache Iceberg. Netflix Technology Blog. https://netflixtechblog.com/introducing-apache-iceberg.

Brackett, Michael H. 2012. Data Resource Simplexity. Technics Publications.

BSI Group. 2024. Case Study: Integrated Management System in Global Healthcare. British Standards Institution. https://www.bsigroup.com/case-studies.

Cataldo, Mike, and Kurt Foote. 2022. DataOps: The Authoritative Edition. O’Reilly Media.

CDISC. 2023a. SDTM Implementation Cases and Best Practices. Clinical Data Interchange Standards Consortium. https://www.cdisc.org/standards/foundational/sdtm.

CDISC. 2023b. Study Data Tabulation Model (SDTM) Implementation Guide V3.4. Clinical Data Interchange Standards Consortium.

Cervo, Dalton, and Mark Allen. 2011. Master Data Management in Practice. Wiley.

Chen, Peter Pin-Shan. 1976. "The Entity-Relationship Model – Toward a Unified View of Data." ACM Transactions on Database Systems 1 (1): 9–36. https://doi.org/10.1145/320434.320440.

Christen, Peter. 2012. Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.

CMMI Institute. 2019. Data Management Maturity (DMM) Model, Version 1.0. ISACA. https://cmmiinstitute.com/data-management-maturity.

Codd, Edgar F. 1970. "A Relational Model of Data for Large Shared Data Banks." Communications of the ACM 13 (6): 377–87. https://doi.org/10.1145/362384.362685.

Confluent. 2024. The Data Streaming Report 2024. https://www.confluent.io/state-of-data-streaming.

Cube Dev. 2024. What Is a Semantic Layer? Cube Documentation. https://cube.dev/blog/what-is-a-semantic-layer.

DAMA International. 2017. DAMA-DMBOK: Data Management Body of Knowledge. 2nd ed. Technics Publications.

DAMA International. 2020. The DAMA Dictionary of Data Management. 2nd ed. Technics Publications.

Data Mesh Learning. 2023. Data Contracts Specification. https://datacontract.com/.

Databricks. 2024a. The Big Book of Lakehouse Architecture. https://www.databricks.com/resources/ebook/big-book-of-lakehouse.

Databricks. 2024b. The Big Book of MLOps. 2nd ed. https://www.databricks.com/resources/ebook/the-big-book-of-mlops.

Dehghani, Zhamak. 2022. Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly Media.

Deloitte. 2023. IFRS 17 Implementation: Lessons from Global Insurance Groups. https://www2.deloitte.com/insurance/ifrs17.

Drucker, Peter F. 1954. The Practice of Management. Harper Business.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." Foundations and Trends in Theoretical Computer Science 9 (3-4): 211–407. https://doi.org/10.1561/0400000042.

EDM Council. 2020. Data Management Capability Assessment Model (DCAM), Version 2.2. https://edmcouncil.org/frameworks/dcam.

English, Larry P. 1999. Improving Data Warehouse and Business Information Quality. Wiley.

Eryurek, Evren, Uri Gilad, Valliappa Lakshmanan, Anita Kibunguchy-Grant, and Jessi Ashdown. 2021. Data Governance: The Definitive Guide. O'Reilly Media.

European Commission. 2020. A European Strategy for Data. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066.

European Data Protection Board. 2020a. Guidelines 07/2020 on the Concepts of Controller and Processor in the GDPR. https://edpb.europa.eu/our-work-tools/documents.

European Data Protection Board. 2020b. Recommendations 01/2020 on Measures That Supplement Transfer Tools. https://edpb.europa.eu/our-work-tools/documents.

European Data Protection Board. 2023. GDPR Enforcement and Data Quality: Industry Cases. https://edpb.europa.eu.

European Parliament. 2024. EU AI Act: Regulation on Artificial Intelligence. European Union.

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." Journal of the American Statistical Association 64 (328): 1183–210. https://doi.org/10.1080/01621459.1969.10501049.

FinOps Foundation. 2024. FinOps Framework. https://www.finops.org/framework.

Forrester Consulting. 2022. Global Telecom Customer Data Master Case Study. https://www.forrester.com/report/telecom-mdm-tei.

Forrester Research. 2023a. The Forrester Wave: Data Governance Solutions, Q3 2023. https://www.forrester.com/report/the-forrester-wave-data-governance.

Forrester Research. 2023b. The Reality of Data Standardization Projects: Why 65% Fall Short. https://www.forrester.com/report/the-reality-of-data-standardization.

Forrester Research. 2023c. The Total Economic Impact of Modern Data Catalogs.

Gartner. 2021. How to Improve Your Data Quality.

Gartner. 2023a. Emerging Tech: Data Fabric Enables the Data-Driven Enterprise. https://www.gartner.com/en/documents/data-fabric.

Gartner. 2023b. Magic Quadrant for Metadata Management Solutions.

Gartner. 2023c. The State of Data and Analytics Governance. https://www.gartner.com/en/documents/data-analytics-governance.

Gartner. 2024a. Magic Quadrant for Data and Analytics Governance Platforms. https://www.gartner.com/en/documents/data-governance-platforms.

Gartner. 2024b. Magic Quadrant for Data Quality Solutions.

Gartner. 2024c. Magic Quadrant for Master Data Management Solutions. https://www.gartner.com/en/documents/mdm-magic-quadrant.

Gartner. 2024d. Market Guide for Data Management Solutions. https://www.gartner.com/en/documents/data-management-solutions.

Gartner. 2024e. Market Guide for Data Modeling Tools.

Gartner. 2024f. Market Guide for Master Data Management Solutions. https://www.gartner.com/en/documents/master-data-management.

Gebru, Timnit et al. 2021. "Datasheets for Datasets." Communications of the ACM 64 (12): 86–92. https://doi.org/10.1145/3458723.

Harvard Business Review. 2023. Why Data Initiatives Fail: A Special Report. Harvard Business Review, July-August 2023. https://hbr.org/2023/07/why-data-initiatives-fail.

Henderson, Deborah et al. 2017. DAMA-DMBOK2 Framework. DAMA International.

HL7 International. 2023. HL7 FHIR Release 5 Specification. https://hl7.org/fhir/R5/.

Hoberman, Steve. 2014. Data Modeling for the Business. Technics Publications.

Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren, eds. 2019. Automated Machine Learning: Methods, Systems, Challenges. Springer. https://www.automl.org/book/.

Huyen, Chip. 2022. Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications. O'Reilly Media.

IBM. 2016. The Cost of Bad Data Is $3.1 Trillion.

IBM. 2022. The State of Dark Data Report.

IBM Research. 2023. Data Quality for AI: New Dimensions for a New Era. https://research.ibm.com/publications/data-quality-for-ai.

IDC. 2023. Global DataSphere Forecast, 2023-2027: The Growth and Evolution of the World's Data. https://www.idc.com/getdoc.jsp?containerId=datasphere-forecast.

IDC. 2024a. Worldwide Data Management Software Forecast, 2024-2028. https://www.idc.com/getdoc.jsp?containerId=data-management-forecast.

IDC. 2024b. Worldwide Multicloud Data Management Forecast, 2024-2028. https://www.idc.com/getdoc.jsp?containerId=multicloud-forecast.

Inmon, Bill, and Ranjeet Srivastava. 2020. Building the Data Lakehouse. Technics Publications.

Inmon, W. H. 2005. Building the Data Warehouse. 4th ed. Wiley.

International Accreditation Forum. 2024. Guidance on ISO/IEC 42001 Certification. https://iaf.nu/en/home/.

International Society of Automation. 2018. Enterprise-Control System Integration. ANSI/ISA-95.00.01-2010. International Society of Automation. https://www.isa.org/products/ansi-isa-95-00-01-2010-iec-62264-1-mod-enterprise.

ISA. 2018a. ANSI/ISA-95 Case Studies: Manufacturing Data Standardization. International Society of Automation. https://www.isa.org/standards-and-publications/isa-standards/isa-95.

ISA. 2018b. ANSI/ISA-95 Enterprise-Control System Integration. International Society of Automation. https://www.isa.org/standards.

ISO. 2016. Data Quality – Part 61: Data Quality Management: Process Reference Model. ISO 8000-61:2016. International Organization for Standardization. https://www.iso.org/standard/63086.html.

ISO. 2022a. Data Quality – Part 1: Overview. ISO 8000-1:2022. International Organization for Standardization. https://www.iso.org/standard/81745.html.

ISO. 2022b. Data Quality (Multiple Parts). ISO 8000. International Organization for Standardization.

ISO. 2022c. Financial Services – Universal Financial Industry Message Scheme. ISO 20022. International Organization for Standardization.

ISO. 2024. ISO Data Management Standards Portfolio. https://www.iso.org/committee/data-management.html.

ISO Central Secretariat. 2024. ISO in Brief: International Standards for a Sustainable World. https://www.iso.org/publication/PUB100007.html.

ISO/IEC. 2004a. Information Technology – Metadata Registries (MDR) – Part 4: Formulation of Data Definitions. ISO/IEC 11179-4:2004. International Organization for Standardization. https://www.iso.org/standard/35346.html.

ISO/IEC. 2004b. ISO/IEC Guide 2:2004 – Standardization and Related Activities – General Vocabulary. International Organization for Standardization. https://www.iso.org/standard/39976.html.

ISO/IEC. 2008. Software Engineering – SQuaRE – Data Quality Model. ISO/IEC 25012:2008. International Organization for Standardization.

ISO/IEC. 2015a. Information Technology – Metadata Registries (MDR) – Part 5: Naming and Identification Principles. ISO/IEC 11179-5:2015. International Organization for Standardization. https://www.iso.org/standard/60341.html.

ISO/IEC. 2015b. Systems and Software Engineering – SQuaRE – Measurement of Data Quality. ISO/IEC 25024:2015. International Organization for Standardization. https://www.iso.org/standard/35749.html.

ISO/IEC. 2022. Information Security, Cybersecurity and Privacy Protection – Information Security Management Systems. ISO/IEC 27001:2022. International Organization for Standardization.

ISO/IEC. 2023a. Information Technology – Artificial Intelligence – Management System. ISO/IEC 42001:2023. International Organization for Standardization.

ISO/IEC. 2023b. Information Technology – Metadata Registries (MDR) – Part 1: Framework. ISO/IEC 11179-1:2023. International Organization for Standardization. https://www.iso.org/standard/78914.html.

ISO/IEC. 2023c. Information Technology – Metadata Registries (MDR) – Part 3: Registry Metamodel and Basic Attributes. ISO/IEC 11179-3:2023. International Organization for Standardization. https://www.iso.org/standard/78916.html.

ISO/IEC. 2024. Annex SL: High-Level Structure for Management System Standards. https://www.iso.org/sites/directives/current/consolidated/index.xhtml.

ISO/IEC JTC 1. 2024. ISO/IEC JTC 1 Business Plan. https://www.iso.org/committee/45020.html.

Kimball, Ralph, and Margy Ross. 2013. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. 3rd ed. Wiley.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” Proceedings of ITCS 2017. https://arxiv.org/abs/1609.05807.

Kotter, John P. 2012. Leading Change. Harvard Business Review Press.

Ladley, John. 2019. Data Governance: How to Design, Deploy, and Sustain an Effective Program. 2nd ed. Academic Press.

LF AI & Data Foundation. 2024. OpenLineage Specification. https://openlineage.io/docs/spec/.

Loshin, David. 2010a. Master Data Management. Morgan Kaufmann.

Loshin, David. 2010b. The Practitioner's Guide to Data Quality Improvement. Morgan Kaufmann.

Makary, Martin A., and Michael Daniel. 2016. "Medical Error – the Third Leading Cause of Death in the US." BMJ 353: i2139. https://doi.org/10.1136/bmj.i2139.

McKinsey Analytics. 2023. Global Financial Services AI Case Studies. https://www.mckinsey.com/capabilities/quantumblack.

McKinsey Digital. 2023. Multi-Cloud Data Transformation in Financial Services. https://www.mckinsey.com/capabilities/digital.

McKinsey Global Institute. 2023. The Data-Driven Enterprise of 2025. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-data-driven-enterprise-of-2025.

MIT Sloan CISR. 2021. Case Study: Building Data Governance at a Fortune 100 Retailer. https://cisr.mit.edu.

MIT Sloan CISR. 2023. Building a Modern Data Stack: Lessons from Fortune 500 Retailers. https://cisr.mit.edu.

Mitchell, Margaret et al. 2019. "Model Cards for Model Reporting." Proceedings of FAT* '19. https://doi.org/10.1145/3287560.3287596.

Moses, Barr, Lior Gavish, and Molly Vorwerck. 2022. Data Quality Fundamentals: A Practitioner's Guide to Building Trustworthy Data Pipelines. O'Reilly Media.

NewVantage Partners. 2023. Data and AI Leadership Executive Survey 2023. https://www.newvantage.com/thoughtleadership.

Ng, Andrew. 2021. A Chat with Andrew on MLOps: From Model-Centric to Data-Centric AI. DeepLearning.AI. https://www.deeplearning.ai/the-batch/data-centric-ai.

NIST. 2023. AI Risk Management Framework (AI RMF 1.0). National Institute of Standards; Technology.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Olson, Jack E. 2003. Data Quality: The Accuracy Dimension. Morgan Kaufmann.

Plotkin, David. 2020. Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance. 2nd ed. Academic Press.

Project Management Institute. 2021. A Guide to the Project Management Body of Knowledge (PMBOK Guide). 7th ed. Project Management Institute.

Rabanser, Stephan, Stephan Günnemann, and Zachary Lipton. 2019. "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift." Advances in Neural Information Processing Systems (NeurIPS 2019). https://arxiv.org/abs/1810.11953.

Redman, Thomas C. 1996. Data Quality for the Information Age. Artech House.

Redman, Thomas C. 2016. Getting in Front on Data: Who Does What. Technics Publications.

Reis, Joe, and Matt Housley. 2022. Fundamentals of Data Engineering. O'Reilly Media.

Riley, Jenn. 2017. Understanding Metadata: What Is Metadata, and What Is It for? NISO Primer Series. National Information Standards Organization. https://www.niso.org/publications/understanding-metadata.

Rowley, Jennifer. 2007. "The Wisdom Hierarchy: Representations of the DIKW Hierarchy." Journal of Information Science 33 (2): 163–80. https://doi.org/10.1177/0165551506070706.

Sambasivan, Nithya et al. 2021. "Everyone Wants to Do the Model Work, Not the Data Work: Data Cascades in High-Stakes AI." Proceedings of CHI '21. https://doi.org/10.1145/3411764.3445518.

Sculley, David et al. 2015. "Hidden Technical Debt in Machine Learning Systems." Advances in Neural Information Processing Systems (NIPS 2015). https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.

Sebastian-Coleman, Laura. 2018. Navigating the Labyrinth: An Executive Guide to Data Management. Technics Publications.

Seiner, Robert S. 2014. Non-Invasive Data Governance. Technics Publications.

Silverston, Len. 2008. The Data Model Resource Book, Volume 1. Revised. Wiley.

Sinek, Simon. 2009. Start with Why: How Great Leaders Inspire Everyone to Take Action. Portfolio.

Singh, Pawan, and Anita Sharma. 2022. "Data Lineage: A Comprehensive Survey of Techniques and Tools." ACM Computing Surveys 55 (7): 1–38. https://doi.org/10.1145/3514232.

Snow, John. 1855. On the Mode of Communication of Cholera. 2nd ed. John Churchill.

Stibo Systems. 2023. Customer Success Stories: Global Manufacturer PIM Transformation. https://www.stibosystems.com/customers.

TDWI. 2023a. Best Practices in Data Governance Tool Selection. https://tdwi.org/research.

TDWI. 2023b. State of Data Modeling Survey.

Uber Engineering. 2017. Meet Michelangelo: Uber's Machine Learning Platform. https://www.uber.com/blog/michelangelo-machine-learning-platform/.

U.S. Census Bureau. 2020. Statistical Metadata Management at the u.s. Census Bureau.

U.S. Federal Government. 2020. Federal Data Strategy 2020 Action Plan. Office of Management; Budget. https://strategy.data.gov/.

국가기술표준원. 2024. KS 표준 개발 및 채택 지침. 대한민국 산업통상자원부. https://www.kats.go.kr.

금융결제원. 2023. 금융 마이데이터 API 표준 및 사업자 사례집. https://www.kftc.or.kr.

금융보안원. 2023. 금융권 데이터 거버넌스 및 도구 활용 실무 가이드. https://www.fsec.or.kr.

유충현. 2026a. 데이터 표준화 라이브러리 사용자 매뉴얼: 기준과 방법, 그리고 자동화를 통한 효율성, 일관성, 품질 향상 제시. 디플래닉스.

유충현. 2026b. 데이터 표준화 애플리케이션 사용자 매뉴얼: KDSM 웹 애플리케이션의 사용과 구축·운영. 디플래닉스.

한국데이터산업진흥원. 2022. 데이터 거버넌스 도입 우수 사례집. https://www.kdata.or.kr.

한국데이터산업진흥원. 2023a. 국내 기업 데이터 레이크하우스 도입 사례. https://www.kdata.or.kr.

한국데이터산업진흥원. 2023b. 중견기업 오픈소스 데이터 관리 도입 사례집. https://www.kdata.or.kr.

한국지능정보사회진흥원. 2023. 공공기관 데이터 거버넌스 우수 사례집. https://www.nia.or.kr.

한국지능정보사회진흥원. 2024a. ISO/IEC 42001 국내 도입 사례 및 인증 가이드. https://www.nia.or.kr.

한국지능정보사회진흥원. 2024b. 기업 LLM 도입 사례 및 데이터 표준화 가이드. https://www.nia.or.kr.

행정안전부. 2023a. 공공데이터 표준화 백서. https://www.data.go.kr.

행정안전부. 2023b. 광역자치단체 데이터 기반 행정 우수 사례. https://www.mois.go.kr.

행정안전부 and 한국지능정보사회진흥원. 2023. 행정정보 데이터 모델 표준화 가이드.