JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 7, pp. 2055-2065, Jul. 2026

Check for updates

공공데이터 연계의 구조적 조건: 컬럼 유형과 참조체계를 중심으로

이 정 윤[1] · 김 학 래[2*]
[1]중앙대학교 문헌정보학과 박사과정 [2]중앙대학교 문헌정보학과 교수

Structural Conditions for Public Data Linkage: Column Types and Reference Systems

JeongYun Lee[1] · Haklae Kim[2*]

[1]Ph.D. Student, Department of Library and Information Science, Chung-Ang University, Seoul, 06974, Korea
[2]Professor, Department of Library and Information Science, Chung-Ang University, Seoul, 06974, Korea

[요 약]

데이터 간 연계는 개별 데이터만으로는 파악하기 어려운 패턴과 관계를 발견하고 새로운 가치를 창출하는 기반이 된다. 한국 공공데이터는 기관 내부 데이터베이스의 메타데이터와 컬럼명 중심의 표준화에 머물러 있어, 실제 개방·제공되는 데이터 간 연계에 한계가 있다. 본 연구는 정형 데이터의 연계 가능성을 검토하기 위해 공공데이터 제공표준 250건을 대상으로 컬럼과 인스턴스 계층의 표현 이질성을 분석하고, 개선 방안을 탐색한다. 분석 결과, 공통표준용어 준수율은 전체 컬럼명 기준 67.0%였으나 고유 컬럼명 기준으로는 15.5%에 그쳐 컬럼명 표준화의 효과가 일부 빈출 컬럼명에 국한되었다. 또한 동일 컬럼명을 공유하는 컬럼군의 93.4%는 동일 유형으로 분류되었으나, 동일 유형이 다양한 컬럼명으로 표현되고 동일 유형 내에서도 참조체계가 데이터셋마다 달라 컬럼과 인스턴스 전반에서 표현 이질성이 확인되었다. 이러한 결과는 현재의 컬럼명 중심 표준화가 데이터 연계의 실효성 확보에 한계가 있음을 시사한다. 데이터의 실질적 연계를 위해서는 제공 시점에 컬럼별 유형과 참조체계가 함께 명시되어야 하며, 이를 위한 정책적·제도적 보완이 필요하다.

[Abstract]

Data linkage enables the discovery of patterns and relationships that cannot be identified from a single dataset. In South Korea, public data standardization has focused largely on metadata and column names at the database level, offering limited support for linking publicly released datasets. This study examines the linkage potential of structured data by analyzing representation heterogeneity at the column and instance levels across 250 provision-standard public datasets and explores possible improvements. Compliance with common standard terms reached 67.0% of all column occurrences but only 15.5% of unique column names, suggesting that standardization benefits only a small set of frequently used names. Of the column groups sharing an identical name, 93.4% fell into the same type. Identical types were expressed through diverse column names, and same-type columns relied on different reference systems across datasets, revealing substantial heterogeneity at both layers. These findings indicate that column-name standardization alone cannot ensure effective data linkage. Standardization must therefore extend to column types and reference systems and be supported by policy and institutional measures for their provision and use.

색인어 : 공공데이터, 데이터 연계, 표현 이질성, 표준화, 정형 데이터
Keyword : Public Data, Data Linkage, Representation Heterogeneity, Standardization, Structured Data

http://dx.doi.org/10.9728/dcs.2026.27.7.2055

Received 04 June 2026; **Revised** 29 June 2026
Accepted 07 July 2026

***Corresponding Author; Haklae Kim**

Tel: +82-2-820-5561
E-mail: haklaekim@cau.ac.kr

Ⅰ. 서 론

서로 다른 형태와 출처의 데이터를 하나의 맥락에서 결합할 때, 개별로는 포착하기 어려운 패턴과 관계를 발견할 수 있다. 데이터 연계의 가치는 최근 AI 에이전트(AI Agents)의 부상으로 더욱 강조되고 있다. 자율적으로 정보를 탐색·추론하며 작업을 수행하는 AI 에이전트는 다양한 출처를 가로질러 필요한 정보를 식별하고 결합할 수 있다. 이때 상호 연결 가능한 구조로 준비된 데이터는 에이전트의 효율적인 작동을 위한 핵심 기반이 된다.

정형 데이터(structured data)를 연계하기 위한 조건은 크게 기술적 메타데이터(descriptive metadata), 컬럼(column), 인스턴스(instance) 계층으로 구분할 수 있다. 이러한 계층 간의 정합성은 데이터 연계를 가능하게 하는 핵심적인 구조적 조건(structural conditions)이다. 기술적 메타데이터는 데이터셋의 발견과 식별을, 컬럼은 데이터셋 간 구조적 연계의 출발점을, 인스턴스는 실세계 개체(entity) 단위의 연결을 담당한다. 따라서 데이터 연계와 통합 활용을 위해서는 세 계층의 정합성이 확보되어야 한다.

데이터 연계의 조건은 공공 영역에도 동일하게 적용된다. 행정·복지·교통·보건 등 각 분야에 분산되어 축적된 공공데이터는 그 자체로도 가치가 있으나, 영역 간 연계를 통해 사회현상에 대한 보다 다층적인 이해와 정책 수립의 근거를 제공할 수 있다. 이러한 잠재적 효용은 데이터가 연계 가능한 형태로 준비되어 있을 때 실현 가능하다. 한국의 공공데이터 정책은 메타데이터와 컬럼명의 형식적 표준화를 위한 제도적·운영적 노력을 지속해 왔으나, 데이터 간 실질적 연계는 여전히 용이하지 않은 실정이다. 그 병목은 동일한 대상이 데이터셋마다 다르게 표현되는 표현의 이질성(representation heterogeneity)에서 비롯되며, 크게 컬럼과 인스턴스 두 계층에 걸쳐 나타난다. 컬럼 계층은 의미상 동일한 대상이 서로 다른 컬럼명으로 표현되거나, 동일한 컬럼명이 상이한 의미로 사용되는 문제가 있다. 따라서 컬럼명 일치만으로는 데이터 간 의미적 연계를 보장하지 못한다. 인스턴스 계층에는 컬럼명이 표준화된 이후에도, 동일 개체를 지칭하는 값이 서로 다른 표현 방식이나 코드 체계로 기록되는 양상이 나타나며, 이는 개체 단위 연계를 제한한다. 현재 공공데이터 정책의 컬럼명 표준화는 기관 내부 데이터베이스 중심의 정비에 머물러 실제 개방 데이터셋 수준까지 확장되지 못하고 있으며, 인스턴스 계층의 표준화는 이에 준하는 체계조차 갖추지 못한 상태이다. 이러한 정책적 공백은 한국 공공데이터의 실질적 연계 가능성을 제약하는 핵심 원인으로 작용한다.

본 논문은 이와 같은 문제의식을 바탕으로, 공공데이터포털에서 제공되는 제공표준 데이터 250건을 기반으로 데이터 연계 관점의 컬럼과 인스턴스 계층 정합성을 분석한다. 구체적으로, 다음 세 가지 연구 질문을 설정한다: RQ1. 제공표준 데이터의 컬럼명은 어느 수준으로 표준화되어 있는가? RQ2. 동일한 컬럼명은 데이터셋 간에 동일한 유형으로 사용되고 있는가? RQ3. 동일한 유형으로 분류된 컬럼들은 동일한 참조체계를 따르고 있는가? 분석 결과를 통해 컬럼명 표준화만으로는 해소되지 않는 한계를 드러내고, 데이터 제공 시점에서 컬럼의 유형과 참조체계가 함께 명시되어야 함을 제안한다. 나아가 이를 뒷받침할 제도적 보완 방향을 논의함으로써, 한국 공공데이터의 연계 가능성 제고를 위한 정책 논의에 실증적 근거를 제공하고자 한다.

본 논문의 구성은 다음과 같다. 2장은 데이터 연계 접근법과 한국 공공데이터 표준화 연구의 흐름을 계층적으로 검토하고, 3장은 분석 대상의 수집과 분석 방법을 기술한다. 4장은 컬럼과 인스턴스 계층의 표현 이질성 실태를 실증하고, 5장은 그 결과를 논의한다. 6장은 결론과 향후 연구 방향을 제시한다.

Ⅱ. 선행 연구

2-1 계층적 데이터 연계 접근법

데이터의 자동 연계와 통합에 관한 연구는 오랫동안 여러 층위에서 전개되어 왔다. 메타데이터 수준의 표준화 연구는 데이터셋 자체를 기술하는 정보(명칭, 제공기관, 분류체계, 갱신주기 등)를 공통 어휘 체계로 표현하려는 시도에서 출발한다. DCAT(Data Catalog Vocabulary)[1]은 웹에 게시된 데이터 카탈로그 간 상호운용성 확보와 통합 검색 지원을 목적으로 설계된 W3C 표준 어휘로, 데이터셋과 서비스를 기계가 읽을 수 있는 일관된 방식으로 기술하도록 한다. 유럽연합은 이를 확장한 DCAT-AP(DCAT Application Profile for data portals in Europe)[2]를 통해 메타데이터 항목별 제약 수준과 통제어휘 적용 요건을 구체적으로 규정하여 회원국 간 공공데이터 통합 검색과 데이터 포털 간 메타데이터 교환 기반을 마련하였다. 그러나 메타데이터 수준의 기술은 데이터셋의 발견과 접근성 제고의 기반은 될 수 있으나[3], 실제 값 단위의 데이터 연계와 통합으로까지 이어지지는 못한다.

발견한 데이터셋들을 실제로 결합하는 단계의 근본적인 어려움은 표현의 이질성이다. 표현의 이질성은 동일한 대상을 독립적인 주체들이 각자의 맥락과 목적에 따라 상이하게 표현함으로써 발생하는 현상을 의미한다[4]. 이 문제를 해소하고 데이터 연계와 통합을 가능하게 하는 핵심 기술로 스키마 매칭·매핑(schema matching·mapping)이 연구되어 왔다[5]. 스키마 매칭은 서로 다른 데이터 소스에서 의미적으로 대응되는 요소를 식별하는 것으로, 스키마 통합(schema integration)을 위한 과정으로 설명된다[6]. 초기 연구는 컬럼명, 데이터 타입, 제약조건 등의 구조적 정보를 기반으로 매칭을 수행하는 스키마 레벨 접근이 주를 이루었으나, 이러한 정보만으로는 충분한 매칭 성능을 확보하기 어렵다는 한계가

지적되었다[7]. 이에 따라 실제 값을 보조 정보로 활용하는 인스턴스 기반 접근이 시도되었으며, 표현 방식의 차이로 인한 오매칭을 완화하고 매칭 정확도를 높일 수 있는 것으로 나타났다[8]. 그러나 자동 매칭은 여전히 휴리스틱에 의존하며, 인간의 개입이 필요한 난제로 남아 있다[9]. 최근에는 이러한 한계를 극복하기 위해 LLM을 활용한 컬럼명 확장[10], RAG·CoT를 결합한 지식그래프 매핑[11], 단계적 파이프라인과[12] 에이전트 기반 유형분류와 정제[13] 등의 연구가 수행되고 있다.

한편, 스키마 수준의 정렬만으로는 실질적인 데이터 연계가 보장되지 않는다는 인식 아래, 데이터 연계 문제는 레코드 간 동일 개체 여부를 판단하는 엔티티 매칭·해소(entity matching·resolution) 문제로 확장되어 왔다[14]. 문자열 유사도나 확률적 모델에 기반한 중복 탐지 연구 등에서 시작하여[15] 사전학습 언어모델을 활용하여 두 레코드의 매칭 여부를 이진 분류 문제로 형식화한 딥러닝 기반 시스템[16]을 거쳐, 범용 LLM에 적절한 프롬프트 설계만으로도 파인튜닝 모델에 준하는 매칭 성능을 달성할 수 있음을 보인 연구[17]에 이르기까지 다양한 방식으로 발전하고 있다. 그러나 기존의 연구 상당수는 후보 쌍이 사전에 주어지는 제한적 매칭 설정에 의존하거나 벤치마크 성능 향상에 초점을 두어 대규모 실세계 데이터 환경으로 일반화하기 어렵다는 문제가 있다. 또한 최신 언어모델도 의미적 이질성을 완전히 파악하는 것은 여전히 한계가 있다[18].

기술적 접근과 함께, 정책적 차원의 데이터 연계 기반을 마련하려는 시도가 이루어져 왔다. 유럽연합은 공공데이터의 상호운용성을 확보하기 위해 EIF(European Interoperability Framework)[19]와 Interoperable Europe Act(2024)[20]를 중심으로 공통 어휘체계와 분류코드의 표준화를 추진하고 관련 법적, 제도적 기반을 강화하는 시도를 하였다. 그러나 프레임워크 수준의 표준화 접근은 실제 데이터 연계에 직접적으로 적용되기 어렵다는 한계가 있다[21].

2-2 한국의 공공데이터 표준화 연구

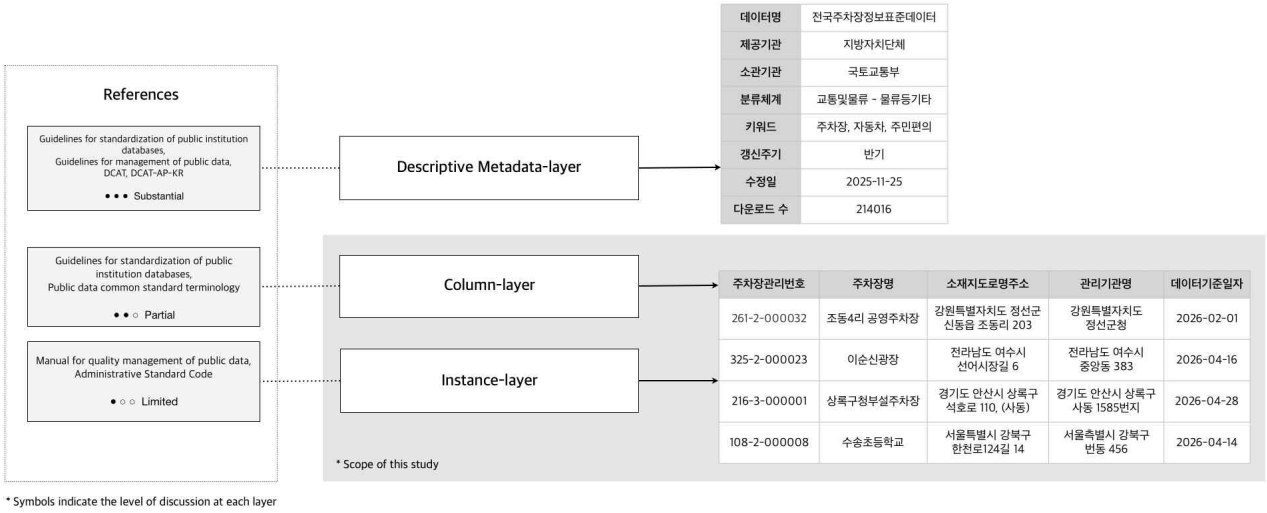
한국의 공공데이터 관리 체계는 초기에 양적 개방 중심으로 추진되었으며, 이는 행정 효율성과 민간 활용도를 높이는 정책 수단으로 평가되었다[22]. 그러나 이러한 접근은 품질 측면에서 한계를 드러냈고, 데이터 정확도·최신성·일관성이 국민 신뢰와 민간 활용에 중대한 영향을 미친다는 인식이 확대되면서[23],[24] 정부는 품질관리에 대한 정책적 강조를 점차 확대해 나갔다.

한국의 공공데이터 표준화 체계는 「공공데이터의 제공 및 이용 활성화에 관한 법률」(이하 공공데이터법)[25]과 그 하위 지침을 기반으로 구성된다. 공공데이터법 제22조와 제23조는 정형 데이터를 중심으로 품질관리와 표준화 원칙을 규정하며, 「공공데이터 관리지침」[26]과 「공공기관의 데이터베이스 표준화 지침」[27]은 이를 구체적으로 이행하기 위한 절차와 기준을 제시한다. 특히 동 표준화 지침 제11조부터 제13조는 공통표준용어, 코드, 도메인의 정의와 관리 원칙을 명시함으로써, 데이터베이스 설계 단계에 적용 가능한 표준화 기반을 마련하고 있다.

이러한 제도적 기반 위에서 수행된 국내 학술 연구는 국제 연구와 유사한 측면이 있다. 데이터셋의 발견과 포털 간 상호운용성을 다루는 연구는 DCAT을 비롯한 국제 표준 기반의 메타데이터 변환 도구 설계[28]와 국내 109개 데이터 포털의 공통 메타데이터를 분석해 한국 환경에 맞게 확장한 DCAT-AP-KR을 제안한 연구가 수행되었다[29]. 이후 컬럼 수준의 정보까지 표현한 지식그래프를 활용하는 통합 검색 프레임워크[30]로 이어지며 점차 정교해지고 있다. 컬럼명 자체의 표준화 문제는 공통표준용어와 제공표준 중심으로 다양한 방식으로 다루어졌다. 제공표준 데이터셋의 컬럼명 간 의미적 연결성을 분석하거나[31], 순환신경망 기반 컬럼명 자동 표준화 모델의 제안[32], 데이터베이스 설계 단계에서부터 표준화를 구현하는 방안을 제시한 연구[33] 등이 그 예이다. 인스턴스 단위의 값을 직접 다룬 연구로는 행정표준코드의 기관코드를 활용하여 부정확한 기관명 데이터를 개선하려는 시도나[34] 링크드 데이터 기반 연계시스템 구축 방안[35]이 제안된 바 있다. 그러나 이러한 연구들은 특정 영역에 국한되어 있거나 구체적인 운영 지침으로 발전되지 못하였으며, 동일 개체 식별 문제를 본격적으로 다루는 연구는 여전히 부족한 상태다.

현재 공공데이터 연계 관련 논의 수준과 적용 범위를 계층별로 정리하면 그림 1과 같다. 기술적 메타데이터는 데이터명·제공기관·갱신주기·분류체계 등 데이터셋의 발견과 식별에 활용되는 정보로 구성되며, 이 계층에 관한 논의는 공공기관의 데이터베이스 표준화 지침, 공공데이터 관리지침과 국제 표준 적용 등 비교적 충실히 다루어져 왔다. 컬럼 계층은 '공공데이터 공통표준용어'가 공통으로 사용되어야 하는 명칭과 도메인을 정의하여 2025년 12월 9차 개정 기준 약 13,159개의 표준용어를 제공한다. 이를 적용한 표준 데이터셋인 '공공데이터 제공표준'은 2026년 3월 20차 개정 기준 300종이 운영되고 있다. 인스턴스 계층은 행정표준코드가 기관·행정구역·일부 분류체계 영역에 한정하여 표준 코드 체계를 제공하고, 공공데이터 품질관리 매뉴얼[36]이 용어의 의미와 형식에 관한 규칙을 두어 정확성과 일관성을 강조하고 있으나, 그 적용 범위는 제한적이다.

종합하면, 국내 공공데이터 연구는 기술적 메타데이터 표준화와 상호운용성 논의를 중심으로 축적되어 왔으며, 컬럼과 인스턴스 계층에 대한 실증적 검토는 제한적이었다. 분석 범위 측면에서, 컬럼명을 다룬 연구는 공통표준용어와 제공표준이라는 제도적 틀이나 데이터베이스 설계 단계에 초점을 두어 실제 개방·제공되는 데이터셋에서의 적용 현황을 검토하지 못하였고, 인스턴스 단위를 다룬 연구는 기관코드 개선이나 링크드 데이터 연계와 같이 특정 영역과 목적에 국한되었다. 특히 기존 연구는 두 계층 중 한쪽만 다루어, 컬럼명 표준



*Examples of public data are presented in Korean (original languages)

그림 1. 공공데이터 연계를 위한 세 가지 준비 계층과 표준화 현황

Fig. 1. Three preparation layers of public data linkage and their standardization status

화가 실제 데이터 연계에 어느 정도 기여하고 어디에서 한계를 드러내는지를 보여주는 진단 결과 제시가 부족하였다. 본 연구는 제공표준 데이터 250건을 대상으로 컬럼과 인스턴스 두 계층의 표현 이질성을 함께 검토한다는 점에서 기존 연구와 구별된다. 이를 통해 컬럼명 표준화만으로는 해소되지 않는 데이터 연계의 한계를 드러내고, 컬럼별 유형과 참조체계의 명시를 포함한 개체 중심 접근의 필요성을 제시한다.

Ⅲ. 연구 설계

3-1 분석 대상과 수집 방법

현행 공공데이터 표준화는 기관 내부의 데이터베이스 구축과 품질관리에 머무르며, 개방·유통을 위한 데이터셋을 가공하는 과정에서 발생하는 정보 손실로 인해 의미적으로 연계 가능한 형태로 준비되어 있음을 보장하지 못한다. 공통표준용어가 존재하지만, 그 자체가 데이터셋 단위의 일관된 적용을 의미하지는 않는다. 따라서 동일 의미가 서로 다른 명칭으로 표현되거나 동일 명칭이 서로 다른 의미로 사용되는 사례가 발생한다. 실제 값 수준에서는 행정표준코드의 적용 범위가 제한적이고 공공데이터 품질관리 매뉴얼 또한 문자열 수준의 유효성 점검에 초점을 두고 있어, 동일 개체를 지칭하는 값이 데이터셋마다 상이한 표기나 코드 체계로 기록되는 표현 이질성은 본격적으로 다루어지지 않는다. 결과적으로 한국 공공데이터 표준화는 메타데이터와 컬럼명 수준에 집중되어 왔으며, 컬럼명의 일관된 적용과 인스턴스 수준의 정합성은 상대적으로 미흡한 상태로 남아 있다.

이러한 한계를 실증적으로 진단하기 위해, 본 연구는 데이터 간 의미적 연결의 핵심이 되는 컬럼 계층과 실제 연계 가능성에 직접 영향을 미치는 인스턴스 계층을 분석 대상으로 설정한다. 공공데이터포털에서 제공되는 데이터셋 대부분은 CSV, XLSX 등 테이블 형식의 정형 데이터임을 고려하여, 분석 대상은 파일 형식으로 제공되는 제공표준 데이터 250건이다. 제공표준 데이터는 공공데이터 관리지침에 따라 표준 제정·준수 의무·적합성 점검·등록 절차가 적용되며 컬럼 목록, 코드값 정의, 참조 행정표준코드와 법령 등을 명시한 별도 문서가 제공된다. 본 연구는 표준화 절차가 공식적으로 적용되는 데이터셋에도 잔존하는 표현 이질성을 검토함으로써 연계의 한계를 진단한다. 수집한 250건은 공공데이터포털 직접 다운로드 176건, 지방행정인허가 통합 다운로드 시스템 70건, 외부 시스템 연계 등 별도 방식 4건으로 구성된다.

3-2 분석 방법

본 연구는 세 가지 연구 질문을 중심으로 분석을 수행하며, 각 연구 질문과 그에 대응하는 분석 방법은 그림 2와 같다. RQ1은 제공표준 데이터셋의 컬럼명 분포(RQ1-M1)와 공통표준용어 준수율(RQ1-M2)을 통해 측정한다. 컬럼명은 전체 컬럼명과 고유 컬럼명 두 기준으로 집계한다. 전체 컬럼명은 동일한 명칭이 여러 데이터셋에서 반복 등장하는 경우를 모두 포함한 값으로 실제 사용 빈도를 반영하고, 고유 컬럼명은 중복을 제거하고 서로 다른 명칭만을 대상으로 한 값으로 어휘적 다양성을 반영한다. 예컨대 '기관명'이 10건의 데이터셋에서 사용된 경우 전체 컬럼명 수는 10개로, 고유 컬럼명 수는 1종으로 산출된다. 공통표준용어 준수율은 2025년 12월 기준 개정·폐기되지 않은 13,159개를 대상으로, 전체 컬럼명과 고유 컬럼명 각각 표준용어와 일치하는 비율로 산출한다. 준수 여부는 공백을 제거한 후 공통표준용어 목록과 문자열이 완전히 일치하는 경우로 한정한다.

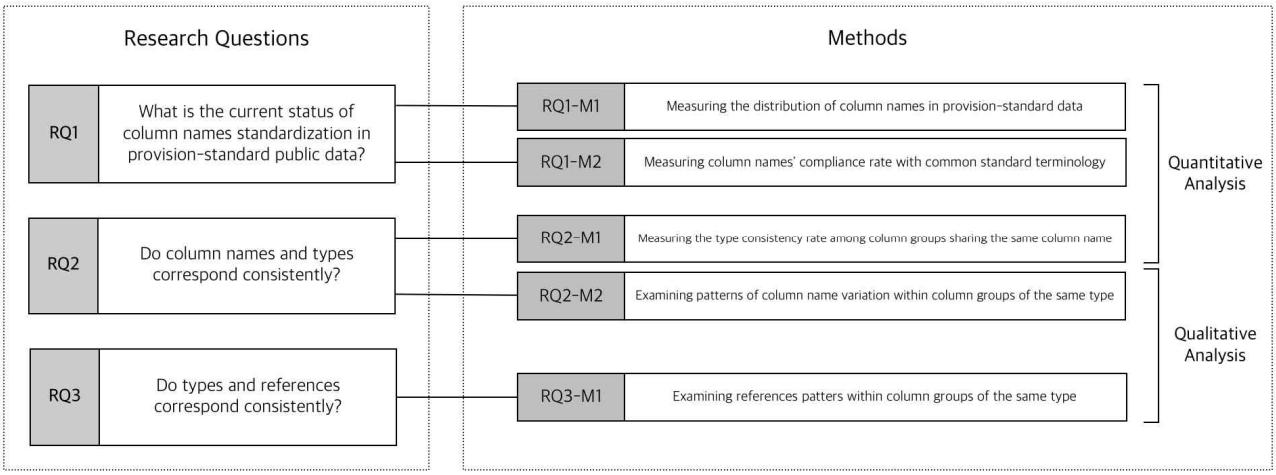


그림 2. 연구질문과 방법
Fig. 2. Research questions and methods

RQ2와 RQ3 분석은 컬럼의 유형 분류를 공통 전제로 한다. 컬럼 유형은 컬럼명과 인스턴스 값의 패턴을 함께 고려하는 위계적 체계로, 표 1과 같이 구성된다. Tier1은 값의 성격에 따라 네 가지 유형으로 구분된다. ID(Identifier, 식별자)는 우편번호, 도로명ID, 학교ID와 같이 특정 대상을 고유하게 식별하는 값이며, CL(Classification, 분류체계)은 한국표준산업분류와 같이 기존의 공식 분류체계에 기반하여 정의된다. CV(Controlled Vocabulary, 통제어휘)는 공식 분류체계에 해당하지는 않으나 일정한 코드 또는 값의 규칙성을 지니는 유형이며, L(Literal, 리터럴)은 앞선 세 유형에 포함되지 않고 상대적으로 자유로운 형태로 표현되는 값을 의미한다. Tier2는 Tier1을 표현 방식에 따라 세분화한 단계이다. ID·CL·CV는 코드형(코드-명 매칭 체계를 통해서만 해석 가능한 값)과 명칭형(텍스트 자체로 의미가 파악되는 값)으로, L은 구조형(날짜, 전화번호, 좌표 등)과 자유형(구조형에 해당하지 않는 비정형 텍스트)으로 구분된다. Tier3는 컬럼이 지칭하는 의미적 대상을 기준으로 식별된 분류 단위이다.

유형 판정은 제공표준 문서의 '설명'과 '허용 값'을 우선 근거로 삼고, 관련 정보가 부족한 경우 인스턴스 값의 구조와 형식적 특성을 함께 고려하는 상향식(bottom-up) 접근을 적용하였다. 분류체계는 연구자가 초기 기준을 적용한 뒤, 판단이 모호한 사례를 반복 검토하면서 Tier3 항목과 판정 기준을 보완하는 방식으로 정교화하였으며, 기준이 안정화된 후 전체 컬럼을 재검토하여 분류의 일관성을 확인하였다. 둘 이상의 유형으로 해석될 수 있는 컬럼은 지칭하는 주된 의미 대상을 기준으로 단일 유형을 부여하였다. 예를 들어 '업태구분명'과 같이 외부 분류체계에서 유래한 것인지 데이터셋 고유의 통제어휘인지 모호한 경우, 허용 값이 공식 분류체계의 항목과 일치하면 CL로, 그렇지 않으면 CV로 판정하였다. 한편, 동일한 의미 대상이 코드형과 명칭형으로 동시에 표기되는 경우에는 모두 단일 Tier3 항목으로 통합된다. 예를 들어, 도시철도 노선의 식별자를 지칭하는 '노선번호'(ID-코드형)와 '노선명'(ID-명칭형)은 '도시철도노선명/노선번호'라는 단일 항목으로 정의된다.

RQ2는 컬럼명과 의미적 유형의 일관성을 양방향으로 검증한다. 동일한 컬럼명을 공유하는 컬럼군이 단일한 의미적 유형으로 수렴하는지 검토한다(RQ2-M1). 유형이 단일하지 않다면, 컬럼명이 의미적 일관성을 보장하지 못함을 의미한다. 반대로, 동일한 의미적 유형으로 분류된 컬럼들의 컬럼명 변이를 정성적으로 검토한다(RQ2-M2). 도메인 맥락과 무관한 표기 변형이나 단순 동의어 사용이 다수 관찰될 경우, 이는 컬럼명 표준화의 부재를 시사한다. 다만 도메인 맥락에 따른 명칭 차이는 자연스러운 현상이므로, 본 분석은 일률적 일치를 전제하지 않는다.

표 1. 수준별 컬럼 유형 분류
Table 1. Tier-based column type classification

Tier1	Tier2	Tier3
ID (Identifier)	ID-code	postal code, school name/ID, road name/ID, etc.
	ID-name	
CL (Classification)	CL-code	road type, administrative standard code (zoning classification name, building use, building purpose), etc.
	CL-name	
CV (Controlled Vocabulary)	CV-code	data update type, road route direction, establishment type name, etc.
	CV-name	
L (Literal)	L-structured	phone number, date/time, year, coordinate-longitude, coordinate-latitude, etc.
	L-free	

RQ3는 RQ2에서 식별한 유형 분류를 한 단계 더 들어가, 동일 유형 내에서 컬럼 값이 의존하는 참조체계가 일관되게 사용되는지 검토한다(RQ3-M1). 참조체계란 컬럼 값이 따르는 외부 표준, 식별자 체계, 코드 체계를 의미하며, 인스턴스 단위의 표현 이질성을 가르는 핵심 요인이다. 예를 들어, '국

가코드'로 명명된 컬럼들은 동일한 유형으로 분류되지만, 같은 국가라도 데이터셋에 따라 ISO 3166-1의 서로 다른 표기(예: '대한민국', 'KOR', 'KR', '410')로 기록되거나 KOICA 국가코드('1424') 혹은 특정 데이터셋에 국한된 체계로 기입된다. 이는 서로 다른 출처의 국가코드 체계가 혼재되어 사용되고 있음을 시사한다.

동일 유형 컬럼군의 참조체계를 일치·불일치로 정량 집계하려면 각 컬럼이 실제로 참조한 체계가 확정될 수 있어야 한다. 그러나 컬럼별 참조체계는 제공표준 문서나 메타데이터에 명시되지 않는 경우가 많고, 공공데이터 전반에 적용되는 참조체계 목록 자체가 정비되어 있지 않으며, 하나의 컬럼에 복수의 체계가 사용된 경우도 존재한다. 이로 인해 대다수 컬럼군에서는 값이 단일 체계를 따르는지 복수 체계가 혼재하는지조차 확정하기 어렵고, 참조체계를 추정할 수 있는 사례도 그 수를 하한값('~개 이상')으로만 제시할 수 있다. 이러한 확정 불가능성은 그 자체로 참조체계가 제공 시점에 명시되지 않는다는 구조적 공백을 드러내는 결과이다. 따라서 본 연구는 컬럼명과 인스턴스 값의 형식을 종합적으로 검토하여 참조체계를 추정하되, 새로운 분류 체계를 구축하거나 정량적 일치도를 측정하기보다 단일 참조체계로 수렴되지 않는 사례를 중심으로 그 양상과 원인을 정성적으로 기술한다.

Ⅳ. 연구 결과

4-1 컬럼명 표준화 현황(RQ1)

RQ1에 대한 분석 결과는 다음과 같다. 250건 제공표준 데이터셋의 전체 컬럼 수는 5,637개이며, 이 중 고유 컬럼명은 1,807종이다. 그림 3은 고유 컬럼명을 출현 빈도 순으로 정렬하였을 때의 순위-빈도 분포를 나타낸 것이다. 소수의 컬럼명이 높은 빈도로 사용되는 반면, 대다수의 컬럼명은 극히 낮은 빈도로 등장하는 롱테일(long-tail) 분포를 보인다. 중앙값은 1회, 평균은 3.1회로 나타났으며, 상위 25%에 해당하는 컬럼명도 대부분 1~2회 수준에 머무르는 반면, 최대 186회

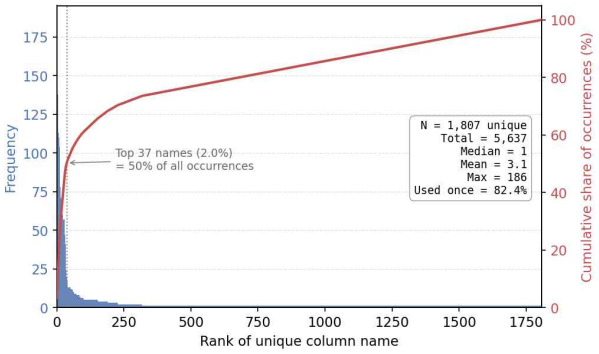


그림 3. 제공표준 데이터 컬럼명 순위-빈도 분포
Fig. 3. Rank-frequency distribution of column names

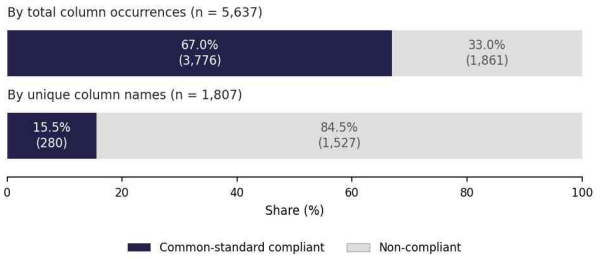


그림 4. 제공표준 데이터 컬럼명의 공통표준용어 준수율
Fig. 4. Compliance rate of column names with common standard terms

까지 반복적으로 등장하는 컬럼명이 존재한다. 전체 1,807종 중 1,488종(82.4%)이 단 한 번만 사용된 반면, 상위 37종(2.0%)의 컬럼명이 전체 출현의 50.3%를 차지하여 소수의 컬럼명 사용이 집중되는 현상이 뚜렷하게 나타난다.

제공표준 데이터의 컬럼명 준수율은 측정 기준에 따라 상이하게 나타났다. 전체 컬럼명 기준으로 5,637개 중 3,776개(67.0%)가 공통표준용어에 등재된 명칭을 사용하고 있다. 반면, 고유한 컬럼명 1,807종 기준으로 보면 280종(15.5%)만 표준을 따르고 있으며, 나머지 1,527종(84.5%)은 비표준 명칭이다. 이러한 차이는 표준 컬럼명이 실제 데이터에서 반복적으로 활용되는 반면, 비표준 컬럼명은 종류는 다양하나 개별 출현 빈도는 낮은 구조적 특성을 반영한다.

4-2 컬럼명과 유형의 일관성 (RQ2)

동일한 컬럼명을 공유하는 컬럼들이 동일한 유형으로 분류되는지 검증하기 위해, 둘 이상의 데이터셋에 출현하는 컬럼명 319종(4,149개)를 분석 대상으로 설정하였다(단일 데이터셋에만 출현하는 컬럼명은 유형 분기를 판별할 수 없어 제외). 분석은 각 컬럼명을 단위로, 그에 속한 컬럼들이 단일 유형으로 수렴하는지를 판정하는 방식으로 수행하였다. 분석 결과, 319종 중 298종(93.4%)은 일관된 유형으로 분류되었고, 21종(6.6%)은 동일 컬럼명임에도 서로 다른 유형으로 분기되었다. 표 2의 Tier 수준별 분기 양상을 살펴보면, Tier1·Tier2는 일치하나 Tier3에서만 분기된 경우가 17종으로 가장 큰 비중을 차지하였다. 예를 들어, '관리번호'는 모든 데이터셋에서 ID-코드형으로 일관되게 분류되었으나, 의미적으로는 데이터관리번호·공원관리번호·택시승강장관리번호 등 서로 다른 9가지 식별 체계를 지칭하는 데 사용되었다. '운영상태' 또한 CV-명칭형으로 수렴하였으나, 값 집합은 '정상운영·휴관·폐쇄', '운영·휴교·폐교' 등 데이터셋마다 상이했다. Tier2까지 분기된 2종은 '도로노선방향'과 '도로종류'로, '상행·하행·양방향' 또는 '일반국도·지방도·군도'와 같은 명칭형 값과 '1·2·3', 'UR·WR·NR'과 같은 코드형 값이 혼재되어 컬럼명만으로는 값의 표현 형식을 예측하기 어려웠다. Tier1까지 분기된 2종은 '가맹점명'(ID-명칭형, L-자유형)과 '설치목

표 2. 동일 컬럼명 컬럼군의 유형 일치율
Table 2. Type consistency rate of column groups sharing an identical column name

Type	Type Agreement by Tier			Unique column names (%)	Column names
	Tier1	Tier2	Tier3		
Consistent	Match	Match	Match	298 (93.4%)	-
Inconsistent	Match	Match	Diverge	17 (5.3%)	관리번호, 시설유형, 등록번호, 구분명, 운영상태, 시설구분, 시설구분명, 학교명, 지정목적, 시설종류, 시도교육청 코드, 설치형태, 도로명, 노선번호, 노선명, 학구분류, 축산물가공 업구분명
	Match	Diverge	Diverge	2 (0.6%)	도로노선방향, 도로종류
	Diverge	Diverge	Diverge	2 (0.6%)	가맹점명, 설치목적
Total				319	-

*Column names are presented in Korean

적'(CV-명칭형, L-자유형)으로, 동일 컬럼명이 서로 다른 의미 범주와 속성을 지칭하여 의미적 이질성이 가장 크게 나타났다.

RQ2-M2에 해당하는 동일 유형 내 컬럼명 변이 양상은 다음과 같다. 도메인 맥락이 반영된 수식어 첨가 형태가 가장 광범위하게 나타났다. 예를 들어 기관기업명 유형의 컬럼이 '관할기관명', '제공기관명', '주최기관명', '운영기관명', '관리기관명'과 같이 상이하게 명명되거나, 전화번호 유형이 '고객센터전화번호', '관리기관전화번호', '관할소방서전화번호' 등으로 분기되는 경우가 이에 해당한다. 이는 지칭 대상을 구체화하기 위해 수식어를 덧붙인 결과로, 의미적 특성을 컬럼명에 반영한 변이로 볼 수 있다. 반면, 도메인 맥락과 무관한 변이도 다수 확인되었다. '설립연도'와 '설립년도'처럼 표기 규칙이 통일되지 않은 경우나 '상세위치', '세부위치', '세부위치 내용'처럼 유사 명칭이 표준화되지 않은 채 혼용되는 경우가 이에 해당한다.

4-3 유형과 참조체계의 정합성 (RQ3)

동일한 유형으로 분류된 컬럼이라도 데이터셋별로 인스턴스가 따르는 참조체계가 상이한 경우가 다수 확인되었다. 이

표 3. 동일 유형 컬럼군 내 참조체계 양상(예시: 업종업태분류)
Table 3. Reference patterns within the same-type column group (example: business category)

Type (Tier3)	Column name	Number of datasets	Number of estimated references	Estimated references
Business Category	업태 구분명	24	6+	식품위생업태유형, 식품위생업종유형, 보건의료기관, 석유판매업구분, KSIC, 법령-사료관리법
	위생 업태명	13	1+	식품위생업태유형
	문화체육 업종명	12	2+	면허세종별, 지방행정인허가 - 업종분류
	문화사업 자구분명	9	2+	면허세종별, 지방행정인허가 - 업종분류
	업종 구분명	5	3+	KSIC, 법령-물환경보전법 시행규칙 별표4 폐수배출시설
	업종명	5	1+	KSIC
	축산업무 구분명	3	1+	법령-축산물 위생관리법
	환경업무 구분명	3	1+	Other
	사업장 업종명	1	1+	건축물용도

*Column names and references are presented in Korean

양상은 동일한 컬럼명이 서로 다른 참조체계의 값을 포함하는 경우와 서로 다른 컬럼명이 같은 참조체계를 공유하는 두 방향에서 나타났으며, 기관명, 행정구역 등 다수 유형에서 공통적으로 관찰되었다. 다만 대부분의 유형에서는 참조체계를 추정하기 어려웠으며, 표 3은 그중 양상이 가장 폭넓게 드러나는 Tier3 '업종업태분류(CL-명칭형)' 사례이다. 제공표준데이터에서 업종업태분류 유형으로 구분되는 컬럼명은 약 9종이며, 컬럼별로 추정된 참조체계의 수는 1개에서 6개 이상까지 분포한다.

동일한 컬럼명 안에서 다양한 참조체계의 값이 사용된 양상이 확인된다. 24건의 데이터셋에서 등장하는 '업태구분명' 컬럼은 최소 6종의 서로 다른 체계에서 유래한 값을 포함한다. 예를 들어, 일반음식점 데이터셋에는 '한식', '일식', '즉석판매제조가공업'과 같은 식품위생업태유형 기반의 값이, 석유판매업소 데이터셋에는 '일반판매소', '주유소', '항공유판매소'와 같은 석유판매업구분 기반의 값이 등장한다. 반대로 서로 다른 컬럼명에서 동일한 참조체계가 사용되는 양상도 나타난다. 한국표준산업분류(KSIC)에서 유래한 것으로 추정되는 값들이 '업태구분명'(예: 배출가스전문정비사), '업종명'(예: 화물자동차운송사업자), '업종구분명'(예: 분뇨수집운반

업체) 등 세 컬럼에 걸쳐 분포한다. 행정표준코드의 면허세종별과 지방행정인허가의 업종분류 역시 '문화체육업종명', '문화사업자구분명' 등의 컬럼에 공통으로 적용되었다.

한편, 추정 참조체계 중에는 이미 폐지된 분류체계도 포함된다. 행정표준코드의 식품위생업태유형과 식품위생업종유형은 현재 폐지된 코드 체계이며, 한국표준산업분류의 과거 차수에서 유래한 값이 현행 데이터에서 여전히 사용되고 있다. 또한 단일 컬럼 안에서 어떤 참조체계를 따랐는지 특정하기 어려운 값이 혼재하는 사례도 다수 확인되었다. 데이터셋이 원 분류체계의 값을 자체적으로 변환하거나 독자적인 분류를 적용한 경우, 기본 분류체계와 일부 상이한 표기법을 사용한 경우가 이에 해당한다. 표 3에서 추정 참조체계 수를 '~개 이상'으로 표기한 것은 이러한 사례를 반영한 결과이며, 실제 참조된 체계의 수는 이보다 많을 수 있다.

Ⅴ. 토 론

5-1 컬럼명 표준화의 효과와 한계 (RQ1)

제공표준 데이터의 공통표준용어 준수율은 전체 컬럼명 기준과 고유 컬럼명 기준 사이에 큰 차이를 보인다(67.0% vs. 15.5%). 이러한 차이는 소수의 표준 컬럼명이 다수의 데이터셋에서 반복적으로 사용되어 전체 기준 준수율을 높인 반면, 비표준 컬럼명은 어휘적으로 다양하게 분산되어 개별 출현 빈도가 낮은 데 기인한다. 이는 컬럼명 표준화의 효과가 빈번하게 사용되는 일부 컬럼명에 국한되며, 롱테일 영역의 컬럼명은 일관되게 정규화되지 않았음을 시사한다.

다만, 이러한 결과가 모든 컬럼명의 일률적 통일이 필요함을 의미하지는 않는다. '도로명주소', '도로명전체주소', '소재지도로명', '소재지주소'와 같이 동일한 의미가 상이하게 표기된 사례는 표준화의 적용을 통해 일관성을 제고할 필요가 있다. 그러나 '종점도로명주소', '기점도로명주소', '역사도로명주소'와 같이 고유한 의미를 반영한 명칭까지 단일 명칭으로 통합하거나 공통표준용어에 일괄 등재하는 것은 실현 가능성이 낮고 표준화 본래의 목적에도 부합하지 않는다. 따라서 컬럼명 표준화의 효용은 일률적 통일이 아닌 도메인 맥락이 작용하지 않는 변이에 선별적으로 적용될 때 확보된다.

5-2 컬럼별 유형 정보 명시 필요성 (RQ2)

동일 컬럼명을 공유한 컬럼군의 유형 일치율이 93.4%로 높게 나타난 것은 컬럼명 표준화가 의미적 일관성 확보에 실질적으로 기여하고 있음을 의미한다. 유형이 분기된 일부 예외(21종)는 컬럼명을 수정하는 등의 보완으로 비교적 용이하게 해소할 수 있다. 그러나 연계 가능성을 충분히 확보하기 위해서는 동일 컬럼명 수준을 넘어, 동일 유형이 서로 다른 컬럼명으로 분산되는 경우까지 식별되어야 한다. RQ2-M2에서 확인되었듯, 동일 유형으로 분류된 컬럼군 내에서 컬럼명 변이는 광범위하게 나타났다. 이러한 변이의 일부는 도메인 맥락을 구체화하기 위한 수식어 부가처럼 의미적 특성을 반영하지만, 표기 규칙의 비일관성이나 유사 컬럼명의 혼용처럼 도메인 맥락과 무관한 변이도 함께 나타난다. 이는 컬럼명만으로는 연계 가능한 컬럼을 식별하기 어려움을 의미한다.

또한 분석에서 제외된 1회 등장 컬럼명이 전체 고유 컬럼명의 80% 이상을 차지하는 만큼, 이들에 유형을 부여한다면 상이한 컬럼명임에도 동일한 유형에 속하거나 사실상 동일한 컬럼임에도 다른 명칭으로 분산되어 있는 사례가 상당수 드러날 가능성이 높다. 결국 데이터셋 간 실질적 연계 가능성을 판단하기 위해서는 컬럼별 유형 정보가 메타데이터 수준 또는 별도의 문서를 통해 명시적으로 제공될 필요가 있다.

5-3 참조체계 정합성 확보를 위한 제도적 요건 (RQ3)

동일 유형 컬럼군이 따르는 참조체계 양상은 컬럼명과 유형 매핑에서 확인된 한계보다 근본적인 문제를 드러낸다. 컬럼 수준에서 동일 유형으로 분류되더라도 데이터셋별로 인스턴스가 의존하는 참조체계가 상이할 경우 값 수준의 연계는 성립하지 않기 때문이다. 유형의 일치는 데이터 연계 가능성의 필요조건일 뿐 충분조건이 되지 못하며, 동일 유형 식별 이후에도 인스턴스 수준의 정합성은 별도로 확보되어야 한다.

인스턴스 수준의 정합성을 확보하기 위해서는 우선 컬럼별로 어떤 참조체계를 따르는지가 명시되어야 한다. 현재는 메타데이터에 컬럼 정보가 기술되지 않으며, 제공표준 문서에도 관련 법령 수준의 언급에 그친다. 개별 컬럼이 어떤 체계를 따르는지 식별되지 않는 한, 데이터셋 간 인스턴스의 매핑 가능성은 판단하기 어렵다. 또한 동일 유형의 컬럼들이 공통으로 참조할 수 있는 명확한 기준 체계가 마련되어야 한다. 단일 컬럼 내에서도 복수의 참조체계가 혼재하는 사례가 다수 확인되는 만큼, 공공데이터 전반에 적용되는 단일 식별자 체계, 분류체계, 코드북이 제공되고 각 데이터셋이 이를 일관되게 참조할 수 있는 구조가 전제되어야 한다. 나아가, 행정표준코드나 한국표준산업분류와 같이 명확한 분류체계가 이미 존재하는 영역에서도 데이터셋이 이를 따르지 않거나 자체적인 표기 변형을 가하는 사례가 많으므로, 마련된 기준의 일관된 적용을 보장하는 제도적 장치가 함께 요구된다.

Ⅵ. 결 론

본 연구는 서로 다른 공공데이터를 연계하기 위해 컬럼과 인스턴스 계층의 이질성을 실증 분석하고, 구조적 조건을 확립하기 위한 제도적 기반을 제안하였다. 공공데이터 제공표준 250건을 대상으로 분석한 결과는 다음과 같다. 첫째, 컬럼명은 높은 어휘적 다양성과 사용 분포가 롱테일 특성을 보이며, 컬럼명 표준화 효과가 제한적으로 나타났다. 둘째, 동일 컬럼명은

대체로 일관된 유형으로 분류되었으나, 동일한 유형이 다양한 명칭으로 표현되어 컬럼명만으로는 연계 가능한 컬럼을 식별하기 어려웠다. 셋째, 동일한 유형 내에서도 참조체계가 다양하게 나타나 값 수준의 데이터 연계에는 한계가 있었다.

이러한 결과는 현행 표준화의 적용 범위가 두 축에서 제한된 데서 비롯된다. 한 축은 표준화가 다루는 계층의 한계로, 표준화가 메타데이터와 컬럼명 계층에 집중되어 연계의 성립을 실제로 결정하는 유형과 참조체계 계층까지 내려가지 못한다. 다른 한 축은 표준화가 작동하는 단계의 한계로, 기관 내부 데이터베이스에 머물러 데이터셋이 가공·등록·개방·활용되는 단계까지 이어지지 못한다. 두 한계가 맞물린 결과, 데이터베이스 단계에서 정의된 코드 체계와 도메인 정보는 개방용 데이터셋으로 가공되는 과정에서 상당 부분 손실되고, 활용 단계의 정합성은 현행 표준화의 적용 범위 밖에 놓인다. 즉, 본 연구가 확인한 표현 이질성은 활용자의 사후 정제로 해소될 문제라기보다, 데이터가 구축·등록·개방되는 단계에 규범적 장치가 부재한 데서 비롯된 구조적 문제이다. 따라서 공공데이터의 연계 가능성을 제고하려면 두 축의 한계에 각각 대응해야 한다. 작동 단계의 한계에 대해서는 표준화 적용 범위를 기관 내부 데이터베이스에서 개방·유통 데이터셋 단계까지 확장해야 한다. 표준화가 다루는 계층의 한계에 대해서는 컬럼별 유형과 참조체계, 허용값의 명시를 데이터 제공 시점에 의무화하고, 동일 의미 영역에 적용되는 단일 분류체계 또는 식별자 체계를 정비하여 각 데이터셋이 이를 일관되게 참조하도록 보장해야 한다. 나아가 이러한 개선이 실효성을 가지려면 개별 기관의 노력을 넘어, 법적·제도적 기반과 함께 기관 간 협력 구조, 연계 현황의 목록화, 품질관리 협의체 운영과 같은 행정적 절차가 함께 갖추어져야 한다.

본 연구는 몇 가지 한계를 지닌다. 분석 대상이 제공표준 데이터 250건으로 한정되어 결과를 공공데이터 전체로 일반화하기 어려우며, 참조체계의 정보가 제공되지 않는 현 상태에서 컬럼명과 인스턴스 값을 단서로 한 추정에 의존하여 결과가 정성적 진단 형태에 머무른다. 다만 제공표준 데이터는 표준화 절차를 공식적으로 적용받는 데이터군임을 고려하면 일반 공공데이터에서는 이질성이 더욱 광범위하게 나타날 가능성이 높다. 향후 연구는 인스턴스 수준의 표현 이질성을 해소하기 위한 기술적 접근으로 확장될 필요가 있다. 컬럼별 유형의 자동 분류 모형 개발, 인스턴스 단위의 자동 정제 모듈 설계, 참조 데이터에 기반한 동일 개체 식별 환경의 구축이 핵심 후속 과제가 될 수 있으며, 이러한 기술적 기반이 제도적 개선과 결합될 때 공공데이터는 AI가 자율적으로 탐색하고 활용할 수 있는 형태로 준비될 수 있다.

감사의 글

이 논문은 2026년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음

참고문헌

[1] R. Albertoni, D. Browning, S. J. D. Cox, A. G. Beltran, A. Perego, P. Winstanley. Data Catalog Vocabulary (DCAT) - Version 3 [Internet]. Available: https://www.w3.org/TR/vocab-dcat-3/.

[2] J. De Cock, E. Dhondt, P. Fragkou, J. Klímek, and A. Sofou. DCAT-AP 3.0.1 [Internet]. Available: https://semiceu.github.io/DCAT-AP/releases/3.0.1/.

[3] W. Yang, R. Fu, M. B. Amin, and B. Kang, “The Impact of Modern AI in Metadata Management,” Human-Centric Intelligent Systems, Vol. 5, No. 3, pp. 323-350, 2025. https://doi.org/10.1007/s44230-025-00106-5

[4] F. Giunchiglia and M. Bagchi, “Representation Heterogeneity,” arXiv:2207.01091, 2022. https://doi.org/10.48550/arXiv.2207.01091

[5] J. Madhavan, P. A. Bernstein, and E. Rahm, “Generic Schema Matching with Cupid,” in *Proceedings of the 27th International Conference on Very Large Data Bases,* Roma, Italy, pp. 49-58, 2001.

[6] E. Rahm and P. A. Bernstein, “A Survey of Approaches to Automatic Schema Matching,” *The VLDB Journal*, Vol. 10, No. 4, pp. 334-350, 2001. https://doi.org/10.1007/s007780100057

[7] Y. Karasneh, H. Ibrahim, M. Othman, and R. Yaakob, “Integrating Schemas of Heterogeneous Relational Databases through Schema Matching,” in *Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, Kuala Lumpur, Malaysia, pp. 209-216, 2009. https://doi.org/10.1145/1806338.1806380

[8] J. He and W. Wang, “An Integrated Heterogeneous Web Service Retrieval via Combination of Instance- and Metadata-Based Schema Matching Method,” *Geo-Spatial Information Science*, Vol. 18, No. 2-3, pp. 111-123, 2015. https://doi.org/10.1080/10095020.2015.1065075

[9] P. A. Bernstein, J. Madhavan, and E. Rahm, “Generic Schema Matching, Ten Years Later,” *Proceedings of the VLDB Endowment,* Vol. 4, No. 11, pp. 695-701, 2011. https://doi.org/10.14778/3402707.3402710

[10] T. Cai, S. Sheen, and A. Doan, “Columbo: Expanding Abbreviated Column Names for Tabular Data Using Large Language Models,” arXiv:2508.09403, 2025. https://doi.org/10.48550/arXiv.2508.09403

[11] N. Vandemoortele, B. Steenwinckel, S. Van Hoecke, and F. Ongenae, “Scalable Table-to-Knowledge Graph Matching from Metadata Using LLMs,” in *Prpceedings of the Semantic Web Challenge on Tabular Data to Knowledge*

Graph Matching 2024, co-located with the 23rd International Semantic Web Conference, Baltimore, pp. 54-68, 2024.

[12] Y. Liu, E. Pena, A. Santos, E. Wu, and J. Freire, "Magneto: Combining Small and Large Language Models for Schema Matching," arXiv:2412.08194, 2024. https://doi.org/10.48550/arXiv.2412.08194

[13] D. Qi, Z. Miao, and J. Wang, "Cleanagent: Automating Data Standardization with LLM-Based Agents," arXiv:2403.08291, 2024. https://doi.org/10.48550/arXiv.2403.08291

[14] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, and K. Stefanidis, "End-to-End Entity Resolution for Big Data: A Survey," arXiv:1905.06397, 2019. https://doi.org/10.48550/arXiv.1905.06397

[15] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 19, No. 1, pp. 1-16, 2007. https://doi.org/10.1109/TKDE.2007.250581

[16] Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, and W.-C. Tan, "Deep Entity Matching: Challenges and Opportunities," *Journal of Data and Information Quality (JDIQ)*, Vol. 13, No. 1, pp. 1-17, 2021. https://doi.org/10.1145/3431816

[17] R. Peeters and C. Bizer, "Using Chatgpt for Entity Matching," in *Proceedings of the ADBIS 2023 Short Papers, Doctoral Consortium and Workshops on New Trends in Database and Information Systems,* Barcelona, Spain, pp. 221-230, 2023. https://doi.org/10.1007/978-3-031-42941-5_20

[18] M. H. Moslemi, A. Mousavi, B. Behkamal, and M. Milani, "Heterogeneity in Entity Matching: A Survey and Experimental Analysis," *Data & Knowledge Engineering*, Vol. 164, 102575, 2026. https://doi.org/10.1016/j.datak.2026.102575

[19] European Commission. New European Interoperability Framework [Internet]. Available: https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf.

[20] L. Hernández Quirós, R. S. Smith, and S. Schade, The Interoperable Europe Act: A Proposed Monitoring Framework − Transitioning Monitoring from the European Interoperability Framework to the Interoperable Europe Act, Publications Office of the European Union, Luxembourg, 2025. https://data.europa.eu/doi/10.2760/9683763

[21] V. Kalogirou, A. Stasis, and Y. Charalabidis, "Adapting National Interoperability Frameworks Beyond EIF 3.0: The Case of Greece," in *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, Athens, Greece, pp. 234-243, 2020. https://doi.org/10.1145/3428502.3428536

[22] Y. W. Hong, "A Study on the Invigorating Strategies for Open Government Data," *Journal of the Korean Data and Information Science Society*, Vol. 25, No. 4, pp. 769-777, 2014. http://doi.org/10.7465/jkdi.2014.25.4.769

[23] H. C. Kim and G. Y. Gim, "A Study on Public Data Quality Factors Affecting the Confidence of the Public Data Open Policy," *Journal of Information Technology Services*, Vol. 14, No. 1, pp. 53-68, 2015. http://doi.org/10.9716/KITS.2015.14.1.053

[24] E. S. Han, "A Study on Public Data Opening Status and Utilization Policy," in *Proceedings of the Korea Contents Association Conference*, pp. 75-76, 2018.

[25] Ministry of the Interior and Safety, Act on Promotion of the Provision and Use of Public Data, Act No. 19408, 2023.

[26] Ministry of the Interior and Safety, Guidelines for Management of Public Data, Notice No. 2021-70, 2021.

[27] Ministry of the Interior and Safety, Guidelines for Standardization of Public Institution Databases, Notice No. 2025-19, 2025.

[28] K. Park, H. S. Won, and K. H. Ryu, "A Design and Implementation of a DCAT-Based Metadata Transformation Tool for Interoperability in Open Data Platforms," *Journal of Digital Contents Society*, Vol. 19, No. 1, pp. 59-65, 2018. http://doi.org/10.9728/dcs.2018.19.1.59

[29] H. Park and H. Kim, "DCAT-AP-KR: Application Profile for Interoperability of Data Portals in Korea," *Journal of Digital Contents Society*, Vol. 23, No. 11, pp. 2249-2258, 2022. https://doi.org/10.9728/dcs.2022.23.11.2249

[30] C. Song and H. Kim, "Development and Evaluation of an Ontology-Based Datamap Search Framework for the Interoperability of Data Portals," *Journal of Digital Contents Society*, Vol. 26, No. 4, pp. 921-930, 2025. https://doi.org/10.9728/dcs.2025.26.4.921

[31] H. Kim, "Metadata Analysis of Open Government Data by Formal Concept Analysis," *Journal of the Korea Contents Association*, Vol. 18, No. 1, pp. 305-313, 2018. https://doi.org/10.5392/JKCA.2018.18.01.305

[32] D. W. Shin, J. Lim, Y. Mun, and H. Jung, "Data Standardization Verification and Transition Model Based on Public Data Common Standard Terminology," *Journal of Knowledge Information Technology and Systems(JKITS)*, Vol. 18, No. 3, pp. 513-524, 2023. http://doi.org/10.34163/jkits.2023.18.3.002

[33] J. Kim and K.-M. Ko, "A Study on Data Quality

Improvement Through the Application of Public Data Standards," *Journal of Korea Multimedia Society*, Vol. 27, No. 12, pp. 1453-1462, 2024. http://doi.org/10.9717/kmms.2024.27.12.1453

[34] H. Kim and H. Kim, "Method for Improving Public Data Quality Applying the Organization Code of Standard Code for Administration," *Journal of Digital Contents Society*, Vol. 23, No. 3, pp. 481-488, 2022. http://doi.org/10.9728/dcs.2022.23.3.481

[35] S.-Y. Yoon, "A Study on National Linking System Implementation Based on Linked Data for Public Data," *Journal of the Korean Society for Information Management*, Vol. 30, No. 1, pp. 259-284, 2013. http://doi.org/10.3743/KOSIM.2013.30.1.259

[36] Ministry of the Interior and Safety. Public Data Quality Management Manual ver. 2.1 [Internet]. Available: https://www.data.go.kr/bbs/rcr/selectRecsroom.do?pageIndex=1&originId=PDS_0000000000000516.

이정윤(JeongYun Lee)

2026년 : 중앙대학교 문헌정보학과 정보학 석사

2020년～2024년: 중앙대학교 문헌정보학과
2024년～2026년: 중앙대학교 문헌정보학과 정보학 석사
2026년～현 재: 중앙대학교 문헌정보학과 정보학 박사과정
※관심분야 : 인공지능, 지식그래프, 공공데이터 등

김학래(Haklae Kim)

2010년 : 아일랜드 국립대학교 (공학박사)

2004년～2009년: Digital Enterprise Research Institute, Ireland
2009년～2016년: 삼성전자
2017년～2019년: 한국과학기술정보연구원
2019년～현 재: 중앙대학교 문헌정보학과 교수
※관심분야 : 지식그래프, 인공지능, 데이터 사이언스 등