Journal of KIIT. Vol. 24, No. 8, pp. 243-258, Aug. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.8.243

공공행정정보시스템 메타데이터의 구조 표준성과 거버넌스 불일치 분석

임수혁*, 서영균**

Analysis of Structural Standardization and Governance Inconsistency in Public Administration Information System Metadata

Soohyuk Im*, Young-Kyoon Suh**

본 논문은 교육부 및 한국연구재단의 4단계 BK21 사업(경북대학교 컴퓨터학부 지능융합 소프트웨어 교육연구단)으로 지원된 연구임(41202420214871).

요 약

공공 행정정보시스템은 동일 시스템을 여러 지자체가 분산 운영하며, 행정안전부의 범정부 메타데이터 관리 시스템이 이를 통합 관리한다. 본 연구는 이 중 17개 광역 지자체의 데이터베이스(DB, Database) 메타데이터 전수를 분석한다. 기존 공공데이터 연구는 주로 포털 단계의 품질 평가에 치중하여 원천 DB의 공개통제 비일관성을 포착하지 못했고, 스키마 매칭 연구 역시 동일 시스템의 다기관 운영 환경을 충분히 고려하지 못했다. 이러한 한계를 극복하기 위해 구조적 표준성과 거버넌스 불일치를 분리 측정하는 2계층 6단계 파이프라인을 제안한다. 분석 결과, 구조적 유사도는 0.9920으로 매우 높았으나 충돌의 93.61%가 공개여부 불일치로 나타나, 구조와 거버넌스가 독립적으로 변동함을 실증하였다. 이는 AI 시대 공공데이터 준비도 확보를 위해 원천 DB 수준의 거버넌스 일관성 개선이 시급함을 시사한다.

Abstract

Public administration information systems in South Korea operate under a decentralized structure, managed independently by local governments and centrally integrated via the Ministry of the Interior and Safety's metadata system. This study analyzes census-level Database (DB) metadata from 17 metropolitan local governments. While previous public data research focuses on open data portal quality—failing to capture openness inconsistencies at the source DB level—schema matching studies overlook identical systems deployed across multiple organizations. To address these gaps, we propose a two-tier, six-stage analytical pipeline evaluating structural standardization and governance inconsistency separately. The analysis reveals that while structural similarity is remarkably high (0.9920), 93.61% of conflicts stem from data openness discrepancies, proving that structure and governance vary independently. These findings highlight that enhancing governance consistency at the source DB level is critical for public data readiness in the AI era.

Keywords

public administration information system, data governance, metadata consistency, structural standardization, AI readiness

* 한국지능정보사회진흥원 책임연구원,
경북대학교 정보과학과 박사수료
- ORCID: https://orcid.org/0009-0008-2323-5618
** 경북대학교 컴퓨터학부 및 정보과학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-3124-2566

· Received: May 11, 2026, Revised: Jun. 19, 2026, Accepted: Jun. 22, 2026

· Corresponding Author: Young-Kyoon Suh
Rm. 520, IT-5, 80 Daehak-ro, Buk-gu, Daegu, South Korea
Tel.: +82-53-950-6372, Email: yksuh@knu.ac.kr

Ⅰ. 서 론

공공 행정정보시스템은 중앙정부가 표준 시스템을 개발·배포하고, 지방자치단체 등 개별 기관이 이를 독립적으로 운영하는 분산 배포 구조를 가진다. 표준지방인사정보시스템, 시도행정시스템, 표준기록관리시스템 등은 동일하거나 유사한 업무 기능을 기반으로 여러 기관에서 운영되며, 각 기관은 행정안전부의 데이터베이스 표준화 지침과 공통표준용어 체계에 따라 데이터베이스 메타데이터를 등록·관리한다[1]. 그러나 동일 시스템을 사용하더라도 실제 운영 과정에서는 컬럼명, 데이터타입, 기본키 여부, 공개여부 등 세부 메타데이터 속성이 기관별로 다르게 설정될 수 있다.

"공공 행정정보시스템의 메타데이터 일관성은 데이터 품질과 스키마 일관성 확립, 공공데이터 표준화 및 개방정책, 공공부문 AI 거버넌스 구축을 위해 선결해야 할 핵심 과제다. 데이터 품질 연구는 일관성을 핵심 품질 차원 중 하나로 제시하였고[2]-[4], 스키마 매칭 연구는 이기종 데이터베이스 간 테이블·컬럼 대응 관계를 발견하기 위한 방법론을 발전시켜 왔다[5]-[7]. 공공데이터 분야 선행 연구들은 공개 포털 메타데이터 품질, 공통표준용어 기반 표준화, 공공데이터 플랫폼 성과, 데이터 품질관리, 행정정보 데이터셋 품질평가 등을 활발히 다루었다[8]-[16]. 또한 공공데이터 개방정책과 데이터 기반행정 연구는 제도적 개선과 기관 간 데이터 공동활용의 필요성을 제기했으며[17]-[21], 공공부문 AI 연구는 AI 도입을 위한 데이터 품질과 거버넌스 준비도의 중요성을 강조했다[22]-[27].

그러나 이들 연구는 주로 데이터 품질의 개념적 차원, 이기종 DB 간 스키마 대응, 공개 포털의 데이터셋 메타데이터, 기관 단위 품질관리, 제도·정책 수준의 데이터 거버넌스에 초점을 두었다. 이에 따라 동일 행정정보시스템을 복수 기관이 운영하는 상황에서 원천 DB 메타데이터의 컬럼 집합 구조와 공개여부 등 속성 수준 거버넌스가 기관 간에 얼마나 일관되게 관리되는지에 대한 분석은 부족했다. 이에 이 공백을 보완하기 위해 동일 시스템의 기관 간 메타데이터 일관성을 집합 수준 구조 표준성과 속성 수준 비단일성으로 분리하여 측정한다.

이러한 공백은 AI 기반 행정 서비스와 데이터기반행정의 확산에 따라 더욱 중요해졌다. AI 모델 학습, 행정 데이터 연계, 검색증강생성(RAG, Retrieval-Augmented Generation), 자동화된 데이터 분석 파이프라인은 데이터 값뿐 아니라 데이터를 설명하는 메타데이터의 일관성에 의존한다. 동일한 행정 개념을 나타내는 컬럼이 기관별로 다르게 명명되거나, 동일 컬럼의 공개여부가 기관마다 다르게 설정되어 있다면 데이터 통합 과정에서 수동 검토 비용이 증가하고, AI 학습 데이터 구축 과정에서는 레이블 충돌이나 접근권한 판단 오류가 발생할 수 있다. 따라서 AI 시대의 공공데이터 준비도는 공개 포털 수준의 품질뿐 아니라 원천 데이터베이스 수준의 메타데이터 일관성까지 함께 평가해야 한다.

본 연구는 17개 광역 지자체 행정정보시스템 데이터베이스 메타데이터를 대상으로 기관 간 구조 표준성과 거버넌스 편차를 분리 측정하는 분석 프레임워크를 제안한다. 구체적으로 동일 정보시스템명과 동일 한글 테이블명을 기준으로 비교가능 테이블 그룹(CTG, Comparable Table Group)을 구성하고, CTG 내 기관 간 컬럼 집합의 일치도는 Jaccard 계수로 측정한다. 이어서 동일 한글 컬럼명을 공유하는 기관들 사이에서 영문 컬럼명, 데이터타입, 기본키 여부, 공개여부가 단일하게 유지되는지를 속성 수준에서 점검한다. 이를 통해 집합 수준의 구조 표준성과 컬럼 속성 수준의 비단일성을 독립된 계층으로 분석한다.

주요 기여는 세 가지다. 첫째, 동일 시스템을 복수 기관이 운영하는 공공 행정정보시스템 환경에서 기관 간 메타데이터 일관성 문제를 CTG 기반 비교 문제로 정식화하였다. 둘째, 컬럼 집합 수준의 구조 표준성과 컬럼 속성 수준의 비단일성을 분리 측정하는 2계층 분석 파이프라인을 제안하였다. 셋째, 17개 광역 지자체의 대규모 원천 DB 메타데이터를 대상으로 제안 방법을 적용하여, 공공데이터 표준화 정책이 실제 운영 메타데이터 수준에서 어떻게 구현되고 있는지를 실증적으로 분석하였다. 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 검토하고, 3장에서는 분석 데이터와 방법론을 설명한다. 4

장에서는 실험 결과를 제시하며, 5장과 6장에서는 연구의 시사점과 향후 과제를 논의한다.

II. 관련 연구

2.1 데이터 품질과 스키마 일관성 측정

Wang과 Strong[2]은 일관성(Consistency)을 포함한 4범주 15차원의 데이터 품질 프레임워크를 제시하였다. ISO/IEC 25012:2008[3]은 데이터 품질 모델에서 일관성을 주요 품질 특성으로 정의하였으며, Batini et al.[4]은 데이터 품질 측정이 스키마, 인스턴스, 표현 수준에서 서로 다른 방식으로 수행되어야 함을 정리하였다. 본 연구는 이 중 스키마 수준의 일관성을 주요 측정 대상으로 삼는다.

스키마 매칭 연구는 서로 다른 데이터베이스 간 테이블 또는 컬럼 대응 관계를 탐지하기 위한 방법론을 발전시켰다. Rahm과 Bernstein[5]은 스키마 매칭 접근법을 언어적, 구조적, 제약 기반 방법으로 분류하였다. Koutras et al.[6]은 Valentine 벤치마크를 통해 데이터셋 발견 및 스키마 매칭 기법을 비교하였으며, Jaccard 기반 매처가 스키마 유사도 측정에서 유용한 기준선으로 활용될 수 있음을 보였다. Dharavath와 Singh[7]은 이기종 분산 데이터베이스의 개체 해소 과정에서 Jaccard 유사도를 결합한 방법을 제안하였다. 다만 이러한 연구들은 주로 이기종 DB 간 1:1 대응 관계 발견을 목적으로 하며, 동일 시스템을 운영하는 복수 기관 간 컬럼 집합과 속성값의 일관성 편차를 측정하는 문제와는 분석 목적이 다르다.

2.2 공공행정 DB 표준화와 메타데이터 품질

공공데이터 메타데이터 품질 연구는 주로 공개 포털에 등록된 데이터셋의 완전성, 발견성, 이용 가능성 등을 평가해 왔다. Neumaier 등[8]은 복수의 오픈데이터 포털을 대상으로 메타데이터 품질을 자동 평가하는 방법을 제시하였고, Quarati[9]는 오픈데이터의 이용 추세와 메타데이터 품질 간의 관계를 분석하였다. 이들 연구는 공개 포털 수준의 메타데이터 품질 평가에 기여하였으나, 원천 행정정보시스템의 DB 메타데이터를 직접 분석하지는 않았다.

국내 공공행정 DB 표준화는 행정안전부의 「공공기관의 데이터베이스 표준화 지침」에 근거한다. 해당 지침은 공공기관이 데이터베이스를 구축·운영할 때 공통표준용어를 준용하도록 요구한다[1]. Shin et al.[10]은 비표준화 공공데이터를 공통표준용어로 자동 변환하는 seq2seq 딥러닝 모델을 제안하여 표준화 자동화의 가능성을 실증하였다. 김종윤·박세환[11]은 국내 공공데이터 플랫폼의 개방 성과와 평가 모델을 실증적으로 분석하였으나, 원천 시스템 메타데이터 수준의 관리 문제는 별도로 다루지 않았다.

공공데이터 품질관리 연구에서도 메타데이터 일관성과 표준화의 중요성이 논의되어 왔다. 김현철·김광용[12]은 공공데이터의 정확성, 신뢰성, 접근성, 표현 일관성 등이 개방정책 신뢰에 미치는 영향을 실증하였다. 김용환 등[13]은 공공기관 데이터 품질관리 실태를 분석하여 다수 기관에서 체계적 품질관리의 개선이 필요함을 보고하였다. Song과 Yim[14]은 행정정보 데이터셋의 품질평가 체계를 설계·적용하여 ISP/ISMP 단계부터 예방적 품질관리가 필요함을 제안하였다. Kim et al.[15]은 공공데이터 활용성 제고를 위해 표준화가 필요함을 논의하고, 데이터 선정부터 표준화 방향 설정까지 공공데이터 관리자가 참고할 수 있는 실무적 기준을 제시하였다. 행정안전부의 공공데이터 품질관리 수준평가 지침[16] 역시 공공데이터 품질관리의 제도적 기준을 제공한다. 그러나 이들 연구와 지침은 주로 공개 데이터셋, 공통표준용어, 기관 단위 품질관리, 포털 성과에 초점을 두고 있으며, 동일 행정정보시스템을 복수 기관이 운영하는 상황에서 원천 DB 메타데이터의 컬럼 집합과 속성값이 기관 간에 얼마나 일관되게 유지되는지는 충분히 검증하지 못하였다.

2.3 공공데이터 개방정책과 데이터기반행정

공공데이터 개방 정책은 「공공데이터의 제공 및 이용 활성화에 관한 법률」 시행과 함께 원칙적 공개와 예외적 비공개를 제도화하였다. 그러나 제도와 현실 사이의 괴리는 다양한 연구에서 지속적으로

보고되어 왔다. Yoon 과 Hyun[17]은 공공데이터 포털의 국가중점데이터를 분석하여 데이터 양은 증가하였으나 실질적 활용 가능성은 낮으며 수요자 중심 전환이 필요함을 제언하였다. Kim[18]은 공공기관의 공공데이터 제공거부 사례를 분석하여 법적 의무와 현장 관행 간의 괴리를 실증하였다. Bang[19]은 「데이터기반행정 활성화에 관한 법률」 시행 이후에도 데이터 수집·관리 거버넌스의 실질적 구축이 미흡함을 비교법 관점에서 분석하였다.

데이터기반행정의 실행 수준에 관해서는 Wang et al.[20]이 텍스트 마이닝을 활용하여 17개 광역 지자체의 데이터기반행정 활성화 시행계획을 분석하고 기관별 실행 수준의 편차를 확인하였다. Kim[21]은 데이터기반행정법 시행 이후에도 기관 간 데이터 공동활용이 저해되는 '데이터 칸막이' 현상이 지속됨을 법제도적으로 분석하고, 데이터 공동활용 영향평가 제도 도입을 제안하였다. 이러한 연구들은 공공데이터 개방과 데이터 공동활용의 제도적 과제를 설명하지만, 공개여부 판단이 원천 DB 컬럼 단위에서 기관별로 어떻게 등록·관리되고 있는지를 실증적으로 측정하지는 못하였다. 본 연구의 공개여부 비단일성 분석은 이러한 제도적 논의를 원천 메타데이터 수준에서 검증한다는 점에서 차별성을 가진다.

2.4 공공부문 AI 거버넌스와 데이터 준비도

AI 기반 행정 서비스의 고도화가 국가적 과제로 부상하면서, 공공부문 AI 도입의 선결 조건으로 데이터 품질과 거버넌스 준비도에 대한 논의가 확대되고 있다. 국가인공지능전략위원회[22]는 공공부문 AI 활용과 데이터 기반 행정 고도화를 주요 정책 방향으로 제시하였다. Yoon[23]은 공공부문 데이터 거버넌스의 조직·제도·기술 구성요소 간 구조적 관계를 실증하여, 조직 요소가 제도와 기술 간 매개 효과를 나타내며 데이터기반행정에 유의한 영향을 미침을 확인하였다. Roh[24]는 공공기관 AI 성숙도 모형(PAIMM, Public Agency AI Maturity Model)을 개발하고 실제 기관에 적용하여 공공기관의 AI 활용 수준을 진단하였다.

공공부문 AI 활용과 학습데이터 품질에 관한 연구도 진행되어 왔다. Kim과 Kim[25]은 공공부문 AI 적용 사례를 전수 분석하여 기술 수준은 고도화되고 있으나 활용 영역이 챗봇 등 일부 서비스에 편중되어 있음을 보였다. Park et al.[26]은 공공데이터를 AI 학습데이터로 활용할 때 결측 필드, 값 범위 불명확, 기관 간 불일치 등이 모델 품질을 저해할 수 있음을 사례 기반으로 제시하였다. Lee[27]는 공공 행정에서 ChatGPT 활용을 중심으로 AI 기술에 대한 중요도와 만족도를 분석하였다. 이들 연구는 공공부문 AI 준비도의 조직적, 제도적, 기술적 측면을 다루었으나, AI 학습과 행정 데이터 연계의 기반이 되는 원천 DB 메타데이터의 기관 간 일관성을 직접 측정하지는 않았다. 따라서 AI 준비도 논의는 원천 데이터베이스 수준의 메타데이터 품질과 거버넌스 일관성 진단으로 확장될 필요가 있다.

2.5 선행연구와의 차별성

이상의 선행연구는 데이터 품질, 스키마 일관성, 공공데이터 표준화, 데이터기반행정, 공공부문 AI 거버넌스의 중요성을 제시하였으나, 동일 행정정보시스템을 복수 기관이 운영하는 환경에서 원천 DB 메타데이터의 기관 간 일관성을 대규모로 실증 측정한 연구는 확인하기 어렵다. 표 1에서는 선행 연구와의 차별점을 다루며, 선행 연구가 이 분석 수준을 다루기 어려웠던 이유는 크게 두 가지다.

첫째, 원천 DB 메타데이터에 대한 제도적·기술적 접근 장벽이다. 공개 포털 기반 연구는 일반 이용자가 접근 가능한 데이터셋 메타데이터를 분석할 수 있으나, 원천 행정정보시스템의 DB 메타데이터는 기관 내부 시스템 또는 범정부 메타데이터 관리체계에 등록·관리되는 운영 메타데이터에 해당한다. 이 메타데이터에는 시스템명, 테이블명, 컬럼명뿐 아니라 공개여부, 비공개 사유, 개인정보 여부, 기본키 여부 등 행정 운영 및 정보공개 판단과 관련된 속성이 포함될 수 있다. 따라서 연구자가 이를 전수 확보하기 위해서는 기관 간 자료 제공 협조, 보안·개인정보·비공개 대상 여부 검토, 기관별 등록 방식 차이에 대한 정제가 선행되어야 한다. 본 연구의 원시 데이터 역시 한국지능정보사회

진흥원(NIA)으로부터 연구 목적으로 제공받은 비공개 행정 자료이다.

둘째, 기존 스키마 매칭 연구와 본 연구의 분석 목적이 다르다. 스키마 매칭 연구는 이기종 DB 간 1:1 컬럼 대응 관계를 발견하는 데 초점을 둔다. 반면 본 연구의 대상은 동일하거나 유사한 행정정보 시스템을 여러 기관이 독립적으로 운영하는 환경이다. 이 경우 핵심 문제는 서로 다른 스키마를 연결하는 것이 아니라, 이미 비교 가능한 동일 시스템의 테이블과 컬럼이 기관별로 얼마나 일관되게 유지되는지를 측정하는 것이다. 따라서 본 연구는 동일 정보시스템명과 동일 한글 테이블명을 기준으로 CTG를 구성하고, 집합 수준의 구조 표준성과 속성 수준의 비단일성을 분리하여 측정한다는 점에서 기존 연구와 차별성을 가진다.

III. 데이터 일관성 분석 및 충돌 탐지

3.1 분석 설계

제안 파이프라인은 그림 1과 같이 6단계로 구성하였다. 1단계는 원시 메타데이터 3,703,676건에 8개 세부 절차의 전처리를 적용하여 3,262,730건의 정제 데이터를 생성한다. 2단계는 동일 정보시스템명 × 동일 한글 테이블명 기준으로 기관을 묶어 비교가능 테이블 그룹(CTG) 6,466개를 구성한다. 3단계는 CTG 내 모든 기관쌍(총 387,638쌍)에 대해 쌍별 Jaccard 계수를 계산하여 집합 수준 구조 표준성을 측정한다. 4단계는 동일 한글 컬럼명을 공유하는 기관들 사이에서 영문 컬럼명·데이터타입·PK·공개여부 4개 속성의 비단일성을 점검하여 속성 수준 충돌을 탐지한다.

3단계와 4단계의 분리가 이 파이프라인의 핵심 설계 의도다. Jaccard가 '컬럼 집합이 얼마나 같은가'를 측정한다면, 충돌 탐지는 '같은 컬럼 내 속성이 얼마나 다른가'를 측정한다. 두 계층을 동시에 분석하기 때문에 구조 표준성이 높더라도 거버넌스 편차가 독립적으로 존재할 수 있음을 실증적으로 확인할 수 있다.

5단계는 집합 유사도를 과대평가할 수 있는 공통 운영컬럼을 출현 비율 기반으로 탐색·확정하여 불용어 목록(Stoplist)을 구성한다. 6단계는 불용어 목록 적용 전(Raw)과 후(Filtered)를 CTG·기관쌍 두 수준에서 비교하여 파이프라인의 견고성을 사후 검증한다. 표 2는 단계별 주요 분석 알고리즘을 요약하여 정리한 것이다.

표 1. 선행 연구와 본 연구의 차별성
Table 1. Differences of this study from previous work

Reference	Subject of analysis	Unit of analysis	Research objective	Main methodology	Difference from this study
Neumaier et al. [8]	Open data portal metadata	Dataset-level completeness	Portal quality diagnosis	Automated metric evaluation	Portal-level open data quality; source DB not addressed
Quarati [9]	Metadata of 28 open data portals	Dataset metadata	Empirical analysis of usage correlation	Regression analysis	Usage-centric; inter-agency consistency not measured
Koutras et al. [6]	Data lake table pairs	Column correspondence	Schema matching technique evaluation	Jaccard benchmark	1:1 mapping discovery; not same-system consistency
Shin et al. [10]	Non-standardized public data	Column-level	Automatic conversion to standard terms	seq2seq DNN	Conversion model; no inter-agency comparison
Kim et al. [13]	Public institution data quality	Institution-level diagnosis	Quality management practice analysis	Survey/indicator	Intra-agency quality; no inter-agency schema comparison
D. S. Kim [21]	Data-driven administration law and policy	Policy/institutional level	Legal task proposal	Legal analysis	Institutional analysis; actual DB metadata not addressed
This study	DB metadata of 17 metropolitan local governments	CTG × institution pair × column attribute	Inter-agency schema consistency + conflict type measurement	Two-tier set-theoretic analysis + coverage-ratio	

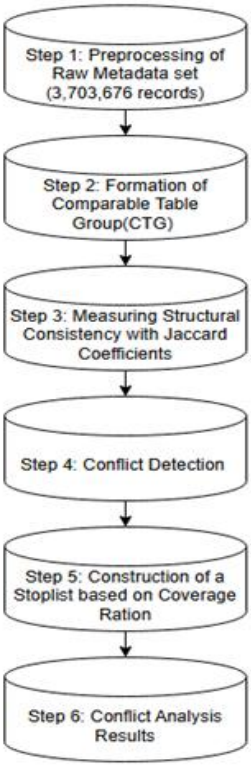

그림 1. 구조 표준성 및 메타데이터 속성 충돌 분석 프로세스

Fig. 1. Hierarchical analysis process for structural standardization and attribute-level conflicts

표 2. 주요 분석 알고리즘 요약

Table 2. Key analysis algorithms summary

Step	Algorithm / technique	Input	Output
① Pre processing	Text normalization, type grouping, duplicate control	Raw records R	Cleaned records R′
② CTG construction	groupby + nunique aggregation	R′	CTG (|G|≥2)
③ Jaccard calculation	Pairwise set intersection/union	Column sets per CTG × institution	avgJ, exact-match ratio
④ Conflict detection	Per-attribute nunique ≥ 2 test	Column metadata within CTG	Conflict column list
⑤ Stoplist	Coverage ratio + regular expression	R′, operational pattern P	불용어 목록 (θ = 0.05)
⑥ Sensitivity	raw/filtered diff + agreement ratio	sim_raw, sim_filtered	Result agreement ratio

3.2 데이터 및 분석단위

본 연구는 행정안전부 범정부 메타데이터 관리시스템의 중앙메타관리시스템으로부터 수집된 17개 시도 광역 지자체 행정정보시스템 메타데이터(2024년 말 기준)를 분석 대상으로 하였다. 분석은 Google Colab Pro(RAM 25 GB) 환경에서 Python 3.10, pandas 2.0, numpy 1.24, openpyxl 3.1을 사용했다. 표 3은 분석환경을 정리한 것이다.

표 3. 분석환경 및 데이터셋 개요

Table 3. Overview of analysis environment and dataset

Category	Description
Analysis language	Python 3.10
Core packages	pandas 2.0, numpy 1.24, openpyxl 3.1
Execution environment	Google Colab Pro (RAM 25 GB) / Single-node batch processing
Data source	Government-wide Metadata Management System, Ministry of the Interior and Safety (as of end-2024)
Target institutions	17 metropolitan local governments (full enumeration)
Raw records	3,703,676
Analyzed attributes	Institution name, information system name, Korean/English table name, Korean/English column name, data type, openness flag, non-disclosure reason, PK indicator (10 attributes)
Records after preprocessing	3,262,730
Final CTG count	6,466 (filtered)
Code repository	https://github.com/s00hyuk/metadata_consistency_analysis

데이터셋은 총 3,703,676건의 레코드(행)와 15개 속성(컬럼)으로 구성되며 표 4는 이를 정리한 것이다. 본 연구의 실험에 사용된 원시 데이터는 한국지능정보사회진흥원을 통해 중앙메타데이터시스템에서 수집하였다. 해당 데이터는 통계적 특성을 보존하면서도 비식별화한 합성데이터와 검증 코드를 GitHub을 통해 공개한다.

합성 데이터(`synthetic_metadata.csv`)는 원본 데이터의 핵심 통계—CTG 수 분포(6,466개), 기관 수 분포(2~17개), 평균 쌍별 Jaccard(0.9920), 완전동일 비율(95.02%), 충돌 비율(~3.2%), 충돌 유형 비율(거버넌스만 93.61%/구조만 2.57%/둘다 3.82%)—를 보존하되 컬럼명·시스템명·기관명을 익명 치환하여 생성하였다. 다음과 같이 합성 데이터를 직접 내려받아

파이프라인을 재현할 수 있다. 합성 데이터 생성 코드(`generate_synthetic_metadata.py`) 역시 동일 저장소에 단독 실행 가능한 형태로 공개되어 있으며, 원시 데이터 없이도 동일한 분포 특성의 데이터를 재생성하고 파이프라인 전 단계를 검증할 수 있다.

표 4. 분석 데이터
Table 4. Analysis data sets

No.	Column name	Type	Note
1	Institution name	Object	
2	Information system name	Object	
3	Information system type	Object	
4	Collection exclusion flag	Object	
5	Korean table name	Object	
6	English table name	Object	
7	Openness flag (Table)	Object	Y/N
8	Non-disclosure reason (Table)	Object	1 - 9*
9	Korean column name	Object	
10	English column name	Object	
11	Openness flag (Column)	Object	Y/N
12	Non-disclosure reason (Column)	Object	1 - 9*
13	Personal information flag	Object	Y/N
14	Data type	Object	
15	PK indicator	Object	Y/N

* Non-disclosure reasons follow article 9, paragraph 1 of the official information disclosure act (national security, decision-making process, personal information, etc.).

분석에는 기관명, 정보시스템명, 한글 테이블명, 한글 컬럼명을 비교 단위 구성용 키로 사용하였고, 영문 컬럼명, 데이터타입, 공개/비공개 여부(컬럼), PK 정보는 충돌 판정 속성으로 사용하였다.

이는 광역 지자체 행정 단위의 전수(全數) 분석이며, 기초자치단체는 동 시스템의 별도 등록 구조를 가지므로 향후 연구 과제로 남긴다. 원시 데이터(3,703,676건)를 17개 기관으로 단순 평균하면 기관당 약 21만 8천 건이나, 실제 기관별 등록 규모는 경기도·서울특별시 등 대규모 기관과 세종·제주 등 소규모 기관 간에 최대 수십 배의 차이가 존재한다. 이러한 불균형에도 불구하고 Jaccard 계수는 크기 차이에 강건(Size-insensitive)하므로, 기관 규모가 비교 결과를 왜곡하지 않는다.

3.3 데이터 전처리

데이터 전처리는 데이터 정제와 분석 신뢰도 확보를 위해 총 8단계의 파이프라인으로 수행했다. 먼저 1단계(결측·공백 처리)에서 필수 식별 키가 결측된 행을 제외한 후, 2단계(한글 항목 정규화)와 3단계(영문 항목 정규화)를 통해 텍스트의 공백을 정제하고 영문은 대문자로 통일하되, 임의의 왜곡을 방지하기 위해 형태소 분석은 배제하고 특수문자(언더스코어, 하이픈)의 차이는 유지하였다. 이어 4단계(공개여부 코드 정규화)와 5단계(PK 여부 정규화)를 통해 관련 코드를 Y/N 값으로 매핑하고 유효하지 않은 오분류 값은 None 처리했으며, 6단계(데이터타입 그룹화)를 통해 DBMS별 세부 타입을 문자·숫자·날짜·논리·기타의 5개 그룹으로 통합하였다. 이때 5·6단계의 정규화 규칙은 예비 분석 과정에서 확인된 오류를 반영하여 최종 보정된 것이다. 다음으로 7단계(플레이스홀더 행 제거)에서는 인천광역시 데이터에서 발견된 수집 불가 일괄 표기 행인 '제외테이블'(15,628행)과 '비표준컬럼'(23,037행)을 전량 제거하였으며, 마지막 8단계(기관 내 중복 통제)에서는 '기관-시스템-테이블-컬럼' 복합 키를 기준으로 중복 데이터 402,281건을 정제하여 대표 행 하나만 유지하였다. 대표 행 선택 방식이 분석 결과에 미치는 영향에 대한 민감도 검증 결과는 4장에서 상세히 기술한다.

3.4 비교가능 테이블 그룹(CTG) 구성 및 구조 표준성 측정

각 CTG G에서 기관 i의 컬럼 집합을 C_i = {한글컬럼명 | 기관=i, 그룹=G}로 정의한다. CTG 내 기관 수가 2 미만인 그룹은 비교 불가능하므로 제외한다.

본 연구가 CTG 구성 기준으로 동일 정보시스템명과 동일 한글 테이블명의 결합을 사용한 것은 동일 시스템을 복수 기관이 운영하는 분산 배포 환경에서 비교 가능한 테이블 단위를 결정론적으로 식별하기 위해서다. 정보시스템명은 중앙에서 배포되거나 동일 업무 기능을 수행하는 시스템 단위를 나타내며, 한글 테이블명은 기관별 DBMS, 영문 약어

규칙, 물리 테이블명 차이에 비해 업무 의미를 비교적 안정적으로 반영한다. 따라서 두 속성을 결합하면 이기종 DB 간 의미 매칭을 추가로 수행하지 않고도 기관 간 비교 가능한 테이블 그룹을 구성할 수 있다. 다만 이 기준은 표면형 일치에 기반하므로 한계를 가진다. 서로 다른 시스템 버전의 테이블이 동일한 한글 테이블명으로 등록된 경우에는 false positive CTG가 생성될 수 있고, 의미적으로 동일한 테이블이 기관별로 다른 한글 테이블명으로 등록된 경우에는 false negative가 발생할 수 있다. 본 연구는 이러한 한계를 줄이기 위해 정보시스템명과 한글 테이블명을 동시에 사용하고, CTG 내 참여 기관 수와 구조 유사도 분포를 사후적으로 점검하였다. 시스템 버전 식별자, 테이블 고유 ID, 의미 기반 테이블 매칭을 활용한 CTG 정교화는 향후 연구 과제로 남긴다.

기관 쌍(i, j)의 Jaccard 계수는 다음 식 (1)과 같이 정의한다.

$$J(i,j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|} \in [0,1] \quad (1)$$

여기서 |C_i ∩ C_j|는 두 기관이 공통으로 보유한 컬럼 수, |C_i ∪ C_j|는 두 기관이 합산하여 보유한 고유 컬럼 수다. J=1.0이면 두 기관의 컬럼 집합이 완전 동일함을, J=0.0이면 공통 컬럼이 전혀 없음을 의미한다.

CTG 내 n개 기관에 대한 그룹 수준 평균 쌍별 Jaccard는 가능한 모든 기관쌍 C(n,2) = n(n-1)/2개의 평균으로 계산한다(식 (2)).

$$avgJ(G) = [2/(n(n-1))] \times \Sigma_{i<j} J(i,j) \quad (2)$$

6,466개 CTG에 걸쳐 총 387,638개 기관쌍이 생성되며, 이 모든 쌍에 대해 Jaccard를 계산한다. 계산 복잡도는 $O(n \cdot k^2)$이며, 여기서 n은 CTG 수(6,466), k는 CTG당 기관 수(≤17, 상수)이므로 실질적으로 O(n)에 준한다.

Jaccard 대신 고려할 수 있는 대안 지표와 비교하면, Dice 계수 D(i,j) = 2|C_i ∩ C_j| / (|C_i| + |C_j|)는 Jaccard보다 교집합에 더 큰 가중치를 부여하나 공개 데이터 공학 문헌에서 Jaccard를 더 광범위하게 사용한다[11]. Overlap 계수 O(i,j) = |C_i ∩ C_j| / min(|C_i|, |C_j|)는 크기 차이가 클 때 과대평가 문제가 있어 본 연구의 맥락에 적합하지 않다. Cosine 유사도는 벡터 기반으로 집합 크기에 민감하여 기관별 컬럼 수가 다른 환경에서 왜곡을 유발할 수 있다. 이러한 비교를 통해 Jaccard가 본 연구의 설계 목적에 가장 부합함을 확인하였다.

CTG 내 기관 수가 1개인 경우는 CTG 구성 조건(|G|≥2)에 의해 사전에 제외한다.

3.5 열 메타데이터 충돌 탐지 및 검증

CTG 내에서 동일 한글 컬럼명을 공유하는 기관들 사이에 표 5의 속성 중 하나라도 단일하지 않으면 충돌로 판정한다. 거버넌스 편차는 법적으로 허용된 기관별 재량과 구별하기 위해 '충돌' 대신 '편차'를 사용하되, 「공공데이터법」의 Negative 방식 개방 원칙 하에서 동일 시스템-동일 컬럼에 대한 공개통제 불일치는 기관 간 조율이 필요한 실질적 편차로 해석한다.

표 5. 메타데이터 일관성 충돌 유형 및 판정기준
Table 5. Metadata consistency conflict types and criteria

Attribute	Conflict type	Decision criterion
English column name	Structural conflict	Distinct values ≥ 2 after normalization
Data type group	Structural conflict	Group categories ≥ 2
PK indicator	Structural conflict	Y/N mixed
Openness flag	Governance discrepancy	Open/Closed mixed

공통운영컬럼은 다수 테이블에 반복 등장하여 집합 유사도를 과대평가할 수 있으므로, 본 연구는 coverage ratio 기반의 불용어 목록(Stoplist)을 구성하였다. 불용어 목록 구성은 후보 탐색과 최종 확정의 두 단계로 수행하였다. 먼저 전체 고유 table-instance 184,489개 중 특정 컬럼이 등장한 비율을 coverage ratio로 정의하고, Θ=0.05 이상 반복 등장하는 컬럼

을 후보로 탐색하였다. 다만 단순 출현 빈도만을 기준으로 할 경우 기관 식별자나 업무 의미를 가진 고빈도 컬럼까지 제거될 수 있으므로, 등록, 수정, 삭제, 사용, 일시, 일자, 비고, 최종, 작성, 처리, 관리, 여부 등 운영관리 성격의 정규표현식 패턴을 함께 적용하였다.

표 6은 Θ=0.05 기준과 운영성 패턴을 동시에 만족한 13개 후보 컬럼과 전문가 검토를 거쳐 확정한 최종 불용어 목록을 정리한 것이다. 후보 중 등록일시, 수정일시, 비고, 최종수정자ID, 등록사용ID, 수정사용ID, 최종수정일시, 사용여부, 삭제여부, 등록일, 등록자ID 등 11개 컬럼은 시스템 운영 또는 이력 관리 목적의 공통운영컬럼으로 판단하여 불용어 목록에 포함하였다. 반면 통합기관코드와 자치단체코드는 coverage ratio와 운영성 패턴 기준을 모두 충족하였으나, 기관 간 비교의 기준점 역할을 수행하는 식별보조컬럼이므로 제외하였다. 이와 같이 본 연구는 반복 출현 정도와 운영관리 의미를 함께 고려하여, 집합 유사도 과대평가 가능성이 높은 컬럼만 제거 대상으로 확정하였다.

출현 비율의 임계값 Θ=0.05는 전체 CTG 중 최소 5% 이상에서 반복 등장하는 컬럼을 공통운영컬럼 후보로 간주하기 위한 기준이다. 단순히 출현 빈도 상위 컬럼을 제거하면 코드, 식별자, 일자 등 실제 업무 의미를 가지는 고빈도 컬럼까지 제거될 수 있으므로, 본 연구는 출현 비율과 운영성 정규표현식 패턴을 함께 적용하였다. 즉, 반복 출현 정도만으로 불용어 목록을 확정하지 않고, 등록, 수정, 삭제, 사용여부, 일시, 비고 등 운영관리 성격을 나타내는 명명 패턴을 동시에 만족하는 컬럼만 후보로 선별하였다. 이후 후보 컬럼에 대해 검토를 수행하여 기관 식별이나 업무 의미 해석에 필요한 컬럼은 제외하고, 집합 유사도를 과대평가할 가능성이 높은 운영관리 컬럼만 최종 불용어 목록에 포함하였다.

Ⅳ. 실험 결과

4.1 구조 표준성 결과

분석 대상이 된 6,468개 CTG의 기관 참여도 분포를 살펴보면, 평균 참여 기관 수는 10.6개, 중앙값은 13개이며 전체의 60.4%(3,908개)가 10개 이상의 기관으로 구성되었다. 만약 매칭 기준의 오류로 인해 서로 다른 시스템의 테이블이 우연히 동일한 명칭을 가져 하나의 그룹으로 묶였다면, 참여 기관 수는 소수에 그쳤을 것이다.

표 6. 최종 불용어 목록 구성 기준과 제거 대상 컬럼 (Θ = 0.05)
Table 6. Final stoplist criteria and removed columns (Θ = 0.05)

Rank	Column name	Coverage ratio*	Operational pattern	Final inclusion	Note
1	Registration datetime	0.1875	✓	Included	–
2	Modification datetime	0.1602	✓	Included	–
3	Remarks	0.1068	✓	Included	–
4	Last modifier ID	0.0939	✓	Included	–
5	Registration user ID	0.093	✓	Included	–
6	Modification user ID	0.093	✓	Included	–
7	Integrated institution code	0.088	✓	Excluded	Institution-identifier auxiliary column
8	Last modification datetime	0.0849	✓	Included	–
9	Use flag	0.07	✓	Included	–
10	Deletion flag	0.0594	✓	Included	–
11	Registration date	0.0578	✓	Included	–
12	Registrant ID	0.0524	✓	Included	–
13	Local Government code	0.0505	✓	Excluded	Institution-identifier auxiliary column

* Coverage ratio = number of unique table-instances in which the column appears / total unique table-instances (184,489). Operational pattern indicates a regular-expression match for operational terms such as registration, modification, deletion, use, date/time, remarks, final, creation, processing, management, and flag.

그러나 절반 이상의 CTG에 10개 이상의 기관이 포함되어 있다는 사실은, CTG로 결합한 테이블들이 실제로 중앙에서 배포되어 각 지자체에서 운영 중인 동일 공통 행정정보시스템의 원천 테이블임을 방증한다. 즉, '동일 정보시스템명 × 동일 한글 테이블명' 기준이 비교 목적에 부합하는 분석 대상을 정확히 식별하고 있음을 뜻한다. 이러한 분석 대상을 바탕으로 도출한 기관 간 구조 표준성 요약 결과는 다음 표 7과 같다.

표 7. 구조 표준성 요약 (공통운영컬럼 제거 전·후)
Table 7. Summary of structural consistency (Before & after removal of common operational columns)

Metric	Raw	Filtered
CTG count	6,468	6,466
Mean pairwise jaccard (avgJ(G))	0.9922	0.9920
Exact-match column set ratio	94.91%	95.02%
Groups with jaccard < 0.95	201	208
Groups with jaccard < 0.50	25	26

정제 후(Filtered) 6,466개 CTG의 avgJ(G) 분포는 고도의 구조적 일관성을 명확히 드러낸다. 기술통계 요약 결과 중앙값(Median)은 1.0000, 사분위범위(IQR, Interquartile Range)는 0.0000(Q1 =1.0, Q3 =1.0)으로 집계되었다. 전체 평균(0.9920)이 중앙값보다 소폭 낮은 것은 Jaccard 계수가 저조한 일부 소수 그룹이 전체 분포를 하방 견인한 결과다. 실제로 왜도(Skewness)는 −7.42로 강한 좌측 편향(Left-skewed)을 나타내며, 이는 대다수 CTG가 완전동일(J=1.0) 상태에 집중된 가운데 소수의 이상치(Outlier) 그룹이 분포의 좌측 꼬리를 형성하고 있음을 확인해 준다. Jaccard<0.95인 208개 그룹(3.2%)이 이 하위 꼬리 구간에 해당한다.

4.2 민감도 분석 결과

커버리지 비율 기반으로 11개 불용어 목록을 구성할 때, 이 필터링이 실제 분석 결과를 과도하게 바꿔버리는 것은 아닌지 확인할 필요가 있다. 만약 불용어 목록에 포함된 컬럼들이 단순 운영관리용이 아니라 기관 간 구조 차이를 반영하는 컬럼이었다면, 이를 제거했을 때 Jaccard 수치와 충돌 건수가 크게 달라졌을 것이다. 이를 확인하기 위해, 불용어 목록 적용 전(Raw)과 후(Filtered)의 Jaccard 계수를 각각 산출하고, CTG 수준(Group)과 기관쌍 수준(Pair) 두 단위로 나누어 두 조건 간의 결과 일치율과 평균 절대 변화량을 비교하였다.

분석 결과, 그룹 수준에서 96.40%의 CTG가, pair 수준에서 99.35%의 기관쌍이 raw와 동일한 결과를 유지하였으며, 평균 절대 변화량도 각각 0.0008과 0.0003에 불과하였다. 즉 불용어 목록을 적용하든 적용하지 않든 CTG 단위와 기관쌍 단위 모두에서 결과의 방향과 크기가 사실상 바뀌지 않았다. 이는 11개 불용어 목록이 분석에서 제거되어야 할 운영관리 컬럼을 적절히 선별한 것이다. 또한 3.5절에서 기술한 커버리지 비율 기반 설계 방식이, 임의적인 상위 N개 제거 방식과 달리, 데이터 분포에 근거한 필터링이라는 점을 사후적으로 확인한 결과이기도 하다.

4.3 메타데이터 충돌 탐지 결과

분석 결과, 필터링(Filtered) 기준으로 총 1,595건의 속성 수준 충돌을 탐지하였다. 전체 고유 비교단위 49,694건 대비 충돌 발생 비율은 약 3.2%이며, 나머지 96.8%의 단위에서는 기관 간 속성값이 일관성을 유지하였다. 충돌 유형별로는 거버넌스 편차만 발생한 경우가 1,493건(93.61%)으로 압도적인 비중을 차지하였고, 구조 속성만 충돌한 경우는 41건(2.57%), 구조 속성과 거버넌스 속성이 동시에 충돌한 경우는 61건(3.82%)을 기록하였다. 이는 17개 기관의 컬럼 집합 구성 자체는 대체로 일치하는 반면, 동일 컬럼에 대한 공개여부 결정 기준은 기관별로 현저하게 다름을 드러낸다.

표 8에서 정제 전(Raw) 조건과 정제 후(Filtered) 조건을 비교하면, 구조 충돌은 115건에서 102건으로 약 11% 감소하였고, 거버넌스 편차는 1,835건에서 1,554건으로 약 15% 감소하였다. 이러한 제한적인 감소 폭은 탐지된 충돌이 특정 공통운영컬럼에 국한되지 않고, 불용어 목록 적용 이후에도 업무 컬럼 전반에 걸쳐 지속되고 있음을 뒷받침한다. 특히 1,595건의 속성 수준 충돌은 단순한 표기 오류를 넘어, 기관 간 데이터 연계 시 실질적인 수동 검토와

조율을 요구하는 최소 분석 단위이다.

표 8. 충돌 층위 비교
Table 8. Comparison of conflict levels

Category	Raw (denom. 1,881)	Filtered (denom. 1,595)
Structural conflict	115 (6.11%)	102 (6.39%)
Governance discrepancy	1,835 (97.55%)	1,554 (97.43%)
Structural only	46 (2.45%)	41 (2.57%)
Governance only	1,766 (93.89%)	1,493 (93.61%)
Both	69 (3.67%)	61 (3.82%)

*Each ratio is calculated against the total number of conflicts under the corresponding condition.

충돌이 발생하는 주요 행정 속성을 파악하고자, 탐지된 충돌 컬럼명을 정규표현식 기반으로 8개 의미범주로 분류하였다. 그림 2는 범주별 충돌 건수와 유형 분포를 나타낸다. 분류 결과, 식별자/코드 범주가 498건(31.22%)으로 가장 많았고, 시간/이력 범주가 327건(20.50%)으로 뒤를 이었다. 두 범주는 전체 충돌의 절반 이상을 차지하였다. 또한 8개 범주 모두에서 거버넌스 편차가 구조 충돌보다 높은 비중을 기록하였으며, 내용/문서 범주에서는 100.00%, 시간/이력 범주에서는 99.39%가 거버넌스 편차에 해당하였다. 이는 공개여부 설정의 기관 간 차이가 특정 속성 유형에 국한되지 않고 행정업무 전반의 컬럼 속성에서 보편적으로 발생함을 시사한다. 한편 수량/값 범주에서는 구조 충돌 비율이 20.83%로 상대적으로 높았는데, 이는 금액·면적 등 수치형 컬럼의 데이터타입을 기관별로 NUMBER, DECIMAL 등 서로 다르게 등록한 데 기인한다. 대표 사례를 살펴보면 거버넌스 편차와 구조 충돌은 서로 다른 방식으로 나타난다. 표준기록관리시스템(RMS, Records Management System)의 공개목록 테이블에 포함된 '결재권자직위직급명' 컬럼은 17개 시도 전체에서 영문 컬럼명(APROV_POS_RANK_NM), 데이터타입(문자형), PK 여부(N)가 동일하였다. 그러나 공개여부는 기관에 따라 '공개' 또는 '비공개'로 다르게 설정되어 있었다. 이는 구조 속성은 표준화되어 있으나 공개통제 기준은 기관별로 다르게 운영되는 명확한 거버넌스 편차 사례이다. 반면 대기환경정보 시스템의 'FAX 전송 여부' 컬럼은 일부 기관에서 영문 컬럼명이 BASIC_FAX_YN으로, 다른 기관에서는 FAX_SEND_YN으로 등록되어 있었다. 이는 동일한 행정 개념을 기관별로 상이한 영문 약어 규칙에 따라 표현한 대표적인 구조 충돌 사례이다.

이어서 탐지된 충돌 1,595건이 정보시스템별로 어떻게 분포하는지를 분석하였다. 충돌이 발생한 정보시스템은 필터링(Filtered) 기준 총 10개였으며, 나머지 시스템에서는 충돌이 탐지되지 않았다. 표 9는 시스템별 충돌 건수, 충돌 유형, 전체 충돌 대비 비중, CTG 수, CTG당 충돌밀도를 정리한 것이다. 여기서 CTG당 충돌밀도는 '충돌 건수 / CTG 수'로 정의되며, 시스템이 보유한 비교가능 테이블 그룹 규모를 고려하여 충돌 집중도를 해석하기 위한 보정 지표다.

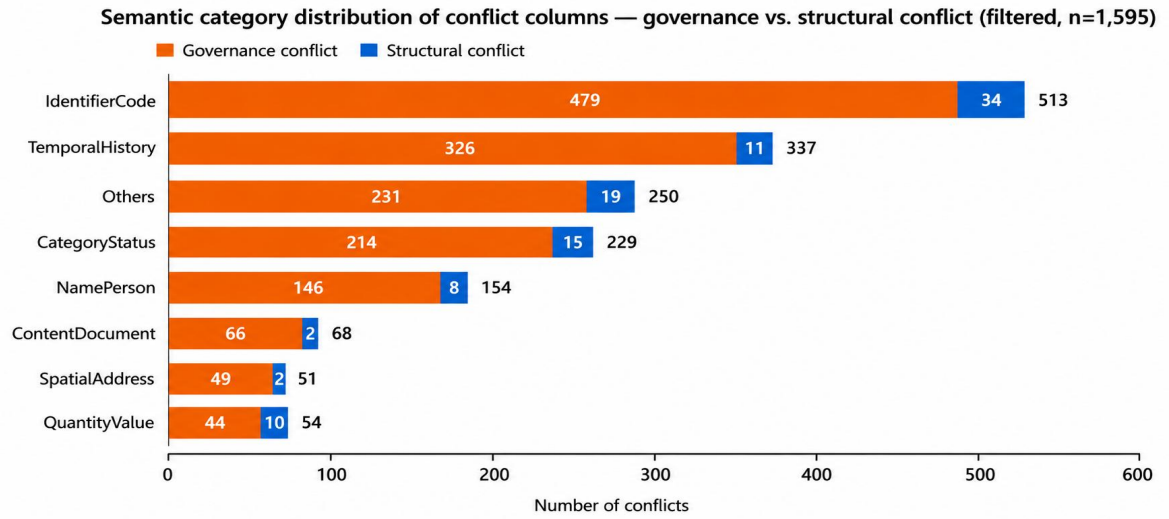

그림 2. 범주별 충돌 건수와 유형 분포 (n=1,595)
Fig. 2. Distribution of conflict counts and types by category (n=1,595)

표 9. 충돌이 발생한 시스템 전체 (filtered 기준, 총 10개)
Table 9. All systems with conflicts (Filtered, N=10)

Information system	Structural only	Governance only	Both	Conflicts	Share of total conflict	CTG count	Conflicts per CTG
Provincial administration system	0	840	6	846	53.04%	1,269	0.667
Standard Records Management System (RMS)	0	312	1	313	19.62%	293	1.068
Comprehensive air quality information system	13	186	45	244	15.30%	31	7.871
Contract information disclosure system	3	77	6	86	5.39%	14	6.143
Research institute information management system	5	61	0	66	4.14%	101	0.653
Urban Planning Information System (UPIS)	3	17	0	20	1.25%	232	0.086
Bus information system	13	0	0	13	0.82%	6	2.167
Fire department homepage	0	0	3	3	0.19%	1	3.000
Local government integrity-e	3	0	0	3	0.19%	467	0.006
Intelligent Transportation System (ITS)	1	0	0	1	0.06%	1	1.000

절대 건수 기준으로는 시도행정시스템, 표준기록관리시스템(RMS), 대기종합정보시스템의 상위 3개 시스템이 전체 충돌의 88.0%를 차지하였다. 그러나 CTG당 충돌밀도로 정규화하면 해석이 달라진다. 시도행정시스템은 충돌 건수 846건으로 가장 많지만 CTG 수가 1,269개로 커서 충돌밀도는 0.667건/CTG로 나타났다. 반면 대기종합정보시스템은 CTG 수가 31개에 불과하지만 충돌 건수가 244건으로, 충돌밀도는 7.871건/CTG에 달하였다. 계약정보공개시스템 역시 14개 CTG에서 86건의 충돌이 발생하여 충돌밀도는 6.143건/CTG로 나타났다. 이는 절대 충돌 건수와 CTG 규모를 보정한 충돌 집중도가 서로 다른 정보를 제공함을 보여준다.

시스템별 충돌 유형 간 차이도 뚜렷하다. 시도행정시스템과 표준기록관리시스템(RMS)은 구조 충돌 대비 거버넌스 편차가 압도적 비율을 차지하여, 중앙 배포 시스템임에도 기관별 공개통제 기준이 불일치함을 입증한다. 반면 버스정보시스템은 13건의 충돌이 모두 구조 충돌에 국한됐고, 거버넌스 편차는 전무했다. 이는 노선·정류장 등 교통 관련 데이터의 영문 컬럼명 약어 규칙을 기관별로 다르게 채택한 결과다. 대기종합정보시스템은 구조 충돌과 거버넌스 편차가 동시에 발생해, 명명 규칙과 공개통제 기준 모두 불균일하게 관리하는 실태를 시사한다.

결과 해석의 엄밀성을 높이기 위해 본 연구는 집합 수준 구조 표준성 지표와 속성 수준 비단일성 지표를 구분하여 해석하였다. 평균 Jaccard 계수는 CTG 내 기관 간 컬럼 집합의 일치도를 나타내는 반면, 속성 수준 충돌은 동일 한글 컬럼명을 공유하는 기관들 사이에서 영문 컬럼명, 데이터타입, PK 여부, 공개여부가 단일하게 유지되는지를 나타낸다. 실제로 집합 수준 평균 Jaccard가 가장 높은 두 시스템－시도행정시스템(평균 Jaccard 0.9976)과 표준기록관리시스템(평균 Jaccard 0.9993)－이 전체 속성 수준 충돌의 72.66%(시도행정 53.04% + 표준기록관리시스템 19.62%)를 차지하였다. 두 계층의 관계를 정량적으로 검증하기 위해, 충돌이 발생한 10개 시스템에 대해 집합 수준 지표(시스템별 평균 Jaccard)와 속성 수준 지표(CTG당 충돌밀도의 로그 변환값) 간 Spearman 순위상관계수를 산출하였다. 그 결과 $\rho = -0.37$ ($p = 0.29$, $n = 10$)로, 두 지표 사이에 통계적으로 유의한 상관이 관찰되지 않았다. 표본 수가 10개로 검정력이 제한적이라는 점을 감안하더라도, 적어도 구조 표준성이 높을수록 속성 일관성도 높아진다는 양의 연관은 본 데이터에서 확인되지 않는다. 이는 컬럼 집합의 구조적 일치도가 높다는 사실만으로 운영 단계의 속성 일관성, 특히 공개여부 거버넌스 일관성을 보장할 수 없음을 의미한다.

따라서 2계층 분석은 단일 유사도 지표만으로는 포착하기 어려운 속성 수준 비단일성을 탐지한다는 점에서 의미가 있다. 집합 수준 Jaccard 계수만 측정하였다면 기관 간 컬럼 집합이 매우 높은 수준으로

일치한다는 결론에 그쳤을 가능성이 높다. 그러나 속성 수준 충돌 탐지를 병행함으로써, 컬럼 집합이 동일하더라도 공개여부나 영문 컬럼명과 같은 세부 속성이 기관별로 다르게 관리되는 사례를 확인할 수 있었다. 이는 공공 행정정보시스템 메타데이터 품질 점검에서 구조 표준성과 속성 일관성을 분리하여 관리할 필요가 있음을 의미한다.

Ⅴ. 결 론

본 논문은 분산 구조로 운영되는 공공 행정정보시스템의 기관 간 DB 메타데이터 일관성을 대규모로 자동 탐지하기 위한 2계층 6단계 파이프라인을 설계하고, 17개 시도 광역 지자체 전수(3,703,676건)에 적용하여 실증하였다. 핵심 설계 원칙은 집합 수준 구조 표준성(쌍별 Jaccard 계수 기반)과 속성 수준 비단일성(컬럼 단위 공개여부·영문명·타입 불일치 기반)을 독립된 계층으로 분리 측정한다는 것이다.

분석 결과, 컬럼 집합 구조의 기관 간 일치도는 매우 높았던 반면 속성 수준에서는 공개여부 거버넌스를 중심으로 한 불일치가 광범위하게 확인되었다. 특히 구조 표준성이 가장 높은 시스템에서 오히려 속성 불일치가 집중되는 역설적 패턴이 나타나, 두 계층이 통계적으로 독립적임을 실증하였다. 이는 집합 수준 단일 지표만으로는 "일관성이 높다"는 결론에 그쳐 속성 수준 거버넌스 편차를 포착할 수 없음을 의미하며, 2계층 분리 측정의 공학적 필요성을 뒷받침한다.

이러한 결과를 바탕으로 본 연구의 공학적 기여는 다음 세 가지로 정리된다. 첫째, 동일 정보시스템명 × 동일 한글 테이블명을 키로 하는 비교가능 테이블 그룹(CTG) 구성 방법론을 제안하여, 이기종 매핑이 아닌 동일 시스템의 기관 간 편차 탐지라는 새로운 분석 문제를 정식화하였다. 둘째, 출현 비율 기반 불용어 목록 자동 탐색 알고리즘으로 공통운영컬럼에 의한 유사도 과대평가를 데이터 분포 근거로 제거하고, 민감도 분석을 통해 그 견고성을 검증하였다. 셋째, LLM 기반 매칭 대신 결정론적 집합 연산을 채택하여 $O(n \cdot k^2) \approx O(n)$ 복잡도로 대규모 전수 처리와 완전한 재현성을 동시에 달성하였다.

실무적으로, 탐지된 속성 불일치는 기관 간 데이터 통합 시 자동 처리가 불가능하여 수동 조율을 요구하는 항목이며, 조율 없이 AI 학습 데이터로 투입될 경우 레이블 충돌을 유발하는 직접적 품질 위험 지점이다. 또한 불일치가 소수 시스템에 집중되는 경향은 전수 수동 점검 대신 이상 집중 시스템을 우선 탐지하는 전략이 품질 개선의 비용 효율성을 높일 수 있음을 시사한다. 제안 파이프라인은 공유 시스템 식별자·비교 가능한 테이블명·컬럼 메타데이터 속성을 갖춘 모든 분산 공공 DB 환경에 이식 가능하며, 코드와 합성 데이터를 공개하여 재현성을 확보하였다.

Ⅵ. 향후 과제

본 연구는 다음과 같은 4가지 연구의 한계를 지닌다.

첫째, CTG 구성이 한글 테이블명 표면형 일치에 의존한다. 서로 다른 시스템 버전이 동일한 한글 테이블명으로 등록된 경우 false positive CTG가 생성될 수 있으며, 의미적으로 동일한 테이블이 다른 이름으로 등록된 경우 false negative가 발생한다. 시스템 버전 식별자나 테이블 고유 ID가 메타데이터에 포함될 경우 이 제약을 추가 통제할 수 있다.

둘째, 속성 비단일성 탐지가 표면형 기반이어서 의미 수준 동의어를 포착하지 못한다. "등록일시"와 "등록일" 처럼 의미적으로 유사하지만 표면형이 다른 컬럼은 별개 CTG로 분리되어 기관 간 비교에서 탐지하지 않는다. 이는 현재 파이프라인의 구조적 false negative 원인이다.

셋째, 2024년 말 단일 시점 스냅샷이어서 속성값 변화 방향을 추적할 수 없다. 탐지된 비단일성이 시간에 따라 수렴하는지 발산하는지를 단일 스냅샷으로는 판단할 수 없으며, 이는 개선 효과 측정의 공학적 한계다.

넷째, 분석 대상을 17개 광역 지자체로 한정한다. 기초자치단체(226개)는 별도 등록 구조를 가지므로 본 연구의 CTG 구성 범위 밖이다. 광역-기초 혼합 데이터셋에서는 CTG 구성 키(동일 시스템명 × 동

일 한글 테이블명)의 재정의가 필요하다.

본 연구의 파이프라인은 다음 네 가지 공학적 방향으로 확장 가능하다.

첫째, 의미 수준 일관성 탐지로의 확장이다. 현재 파이프라인은 한글 컬럼명의 표면형 일치를 CTG 구성 기준으로 사용한다. 이 설계는 "등록일시"와 "등록일"처럼 의미적으로 동일하지만 표면형이 다른 컬럼 쌍을 별개 컬럼으로 처리하여 false negative를 유발할 수 있다. 후속 연구로 공통표준용어 도메인 임베딩(Word embedding) 또는 편집 거리(Edit distance) 기반 퍼지 매칭(Fuzzy matching)을 적용하여 CTG 구성 단계를 의미 수준으로 확장하면, 현재 파이프라인이 포착하지 못하는 동의어 컬럼 간 속성 불일치를 추가 탐지할 수 있다.

둘째, 시계열 스냅샷 기반 속성 변화 탐지 알고리즘 설계다. 단일 시점(2024년 말) 스냅샷에 기반한 현재 분석은 속성 불일치의 발생 시점과 변화 방향을 추적할 수 없다. 연도별 메타데이터 스냅샷을 누적하면 기관 t에서 컬럼 c의 공개여부 속성값이 시점 t_1 에서 t_2 사이에 Y→N 또는 N→Y로 전환된 이벤트를 자동으로 탐지하는 change detection 알고리즘을 구성할 수 있다. 이는 속성 불일치의 원인이 초기 등록 오류인지 사후 변경인지를 구분하는 진단 정밀도를 높인다.

셋째, 파이프라인의 분산 처리 및 스케일아웃 설계다. 현재 구현은 Google Colab 단일 노드(RAM 25 GB) 환경에서 pandas 기반 배치 처리로 수행되었으며, 370만 건 규모에서 O(n) 복잡도로 충분히 처리 가능하였다. 그러나 기초자치단체·중앙부처 포함 시 레코드 수가 수천만 건 이상으로 증가할 경우, Apache Spark 또는 Dask 기반 분산 처리 프레임워크로 CTG 구성(Groupby)과 쌍별 Jaccard 계산(mapPartitions)을 병렬화하면 선형 스케일아웃이 가능하다. 특히 CTG 수 n과 기관 수 k가 독립적으로 증가하는 상황에서 $O(n \cdot k^2)$의 k 축 병렬화가 핵심 최적화 지점이다.

넷째, 속성 불일치 자동 분류 및 이상 탐지 모델 통합이다. 현재 파이프라인은 규칙 기반(nunique ≥ 2) 탐지로 불일치 존재 여부만을 판정한다. 탐지된 불일치를 ① 초기 등록 오류, ② 기관별 재량 적용, ③ 시스템 버전 차이 등 원인 유형으로 자동 분류하는 멀티클래스 분류기를 학습하면 우선순위 기반 개선 권고 생성이 가능해진다. 또한 CTG당 불일치율 분포에서 통계적 이상값(Outlier)을 자동 탐지하는 이상 탐지(Anomaly detection) 모듈을 파이프라인 6단계에 추가하면, 대기종합정보시스템(787.1%)과 같이 집중적 불일치가 발생한 시스템을 입력 전수 검토 없이 자동으로 식별할 수 있다.

References

[1] Ministry of the Interior and Safety, "Guidelines for Database Standardization in Public Institutions", Ministry of the Interior and Safety Notice No. 2025-19, Feb. 2025.

[2] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers", Journal of Management Information Systems, Vol. 12, No. 4, pp. 5-33, Sep. 1996. https://doi.org/10.1080/07421222.1996.11518099.

[3] ISO/IEC, "ISO/IEC 25012:2008 Software engineering — Data quality model", International Organization for Standardization, Dec. 2008.

[4] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement", ACM Computing Surveys, Vol. 41, No. 3, pp. 1-52, Jul. 2009. https://doi.org/10.1145/1541880.1541883.

[5] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching", The VLDB Journal, Vol. 10, No. 4, pp. 334-350, Dec. 2001. https://doi.org/10.1007/s007780100057.

[6] C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati, and A. Katsifodimos, "Valentine: Evaluating matching techniques for dataset discovery", Proc. IEEE International Conference on Data Engineering (ICDE), Online, pp. 468-479, Apr. 2021. https://doi.org/10.1109/ICDE51399.2021.00047.

[7] R. Dharavath and A. K. Singh, "Entity

resolution-based Jaccard similarity coefficient for heterogeneous distributed databases", Advances in Intelligent Systems and Computing, Vol. 379, pp. 497-507, Jan. 2016. https://doi.org/10.1007/978-81-322-2517-1_48.

[8] S. Neumaier, J. Umbrich, and A. Polleres, "Automated quality assessment of metadata across open data portals", ACM Journal of Data and Information Quality, Vol. 8, No. 1, pp. 1-29, Oct. 2016. https://doi.org/10.1145/2964909.

[9] A. Quarati, "Open government data: Usage trends and metadata quality", Journal of Information Science, Vol. 49, No. 4, pp. 887-910, Aug. 2023. https://doi.org/10.1177/01655515211027775.

[10] D. W. Shin, J. Y. Lim, Y. K. Moon, and H. K. Jung, "Data standardization verification and conversion model based on public data common standard terms", Journal of Knowledge Information Technology and Systems, Vol. 18, No. 3, pp. 513-524, Jun. 2023. https://doi.org/10.34163/jkits.2023.18.3.002.

[11] J. Y. Kim and S. H. Park, "A study on the analysis of opening performance and empirical evaluation model of domestic public data platforms", Journal of Knowledge Information Technology and Systems, Vol. 19, No. 1, pp. 105-113, Feb. 2024. https://doi.org/10.34163/jkits.2024.19.1.010.

[12] H. C. Kim and K. Y. Kim, "A study on the effect of public data quality factors on trust in public data open policy", Journal of Information Technology Services, Vol. 14, No. 1, pp. 53-68, Mar. 2015. https://doi.org/10.9716/KITS.2015.14.1.053.

[13] Y. H. Kim, J. H. Lee, Y. J. Yoon, S. I. Lee, and S. J. Kim, "A study on data quality management using analysis of public data quality levels", Journal of the Korea Institute of Information and Communication Engineering, Vol. 27, No. 12, pp. 1465-1472, Dec. 2023. https://doi.org/10.6109/jkiice.2023.27.12.1465.

[14] C. H. Song and J. H. Yim, "A study on data quality assessment of administrative information datasets", The Korean Journal of Archival Studies, No. 71, pp. 237-272, Dec. 2022. https://doi.org/10.20923/kjas.2022.71.237.

[15] E. J. Kim, M. S. Kim, and H. W. Kim, "A study on public data standardization for improving utilization", Knowledge Management Research, Vol. 20, No. 4, pp. 23-38, Dec. 2019. https://doi.org/10.15813/kmr.2019.20.4.002.

[16] Ministry of the Interior and Safety, "Guidelines for Public Data Quality Management Level Assessment", Ministry of the Interior and Safety, May 2023.

[17] S. O. Yoon and J. W. Hyun, "A study on the current status and improvement measures of public data open policy: Focusing on the opening cases of national core data in the public data portal", Korean Public Management Review, Vol. 33, No. 1, pp. 219-247, Mar. 2019. https://doi.org/10.24210/kapm.2019.33.1.010.

[18] E. S. Kim, "A study on legal and institutional improvement measures for promoting the opening and utilization of public data: Focusing on cases of refusal to provide public data", Informatization Policy, Vol. 30, No. 2, pp. 46-67, Jun. 2023. https://doi.org/10.22693/NIAIP.2023.30.2.046.

[19] J. M. Bang, "Data-based administration for the realization of public data utilization", Comparative Law Review, Vol. 21, No. 1, pp. 87-126, Jun. 2021. https://doi.org/10.56006/JCL.2021.21.1.3.

[20] M. H. Wang, H. K. Oh, and M. J. Na, "Analysis of trends in local governments' data-based administration promotion using text mining", Journal of Local Government Studies, Vol. 35, No. 2, pp. 31-50, Jun. 2023. https://doi.org/10.21026/jlgs.2023.35.2.31.

[21] D. S. Kim, "Legal Framework Improvement for Activating Data-Driven Administration", Dankook Law Review, Vol. 48, No. 1, pp. 179-206, Mar.

2024.

[22] National Artificial Intelligence Strategy Committee, "Korea Artificial Intelligence Action Plan: Basic Plan for Artificial Intelligence 2026-2028", National Artificial Intelligence Strategy Committee, Feb. 2026.

[23] K. Yoon, "An empirical study on data governance: Focusing on structural relationships and effects among components", Informatization Policy, Vol. 30, No. 3, pp. 29-48, Sep. 2023. https://doi.org/10.22693/NIAIP.2023.30.3.029.

[24] S. Y. Roh, "From digital public institutions to AI public institutions: Development and application of the Public AI Maturity Model (PAIMM)", Public Institutions and Policy Review, Vol. 3, No. 2, pp. 1-41, Dec. 2025. https://doi.org/10.23412/pipr.3.2.202512.001.

[25] H. S. Kim and B. J. Kim, "An exploratory study on the use of AI in the Korean public sector: Evolution of technology and concentration of use", Korean Public Administration Review, Vol. 57, No. 2, pp. 205-246, Jun. 2023. https://doi.org/10.18333/KPAR.57.2.205.

[26] S. Y. Park, D. Y. Lee, and J. A. Kim, "A case study on public data refinement required for building training data", Asia-Pacific Journal of Convergent Research Interchange, Vol. 9, No. 5, pp. 117-128, May 2023. https://doi.org/10.47116/apjcri.2023.05.10.

[27] I. J. Lee, "Importance-satisfaction analysis of AI technology in public administration: Focusing on the use of ChatGPT", Journal of the Korea Academia-Industrial Cooperation Society, Vol. 25, No. 11, pp. 913-926, Nov. 2024. https://doi.org/10.5762/KAIS.2024.25.11.913.

저자소개

임 수 혁 (Soohyuk Im)

2022년 2월 : 경북대학교 정보과학과(박사수료)
2016년 6월 ~ 현재 : 한국지능정보사회진흥원(NIA) 책임연구원
관심분야 : 국가 데이터 플랫폼, 데이터 거버넌스, 메타데이터 일관성 및 구조적 표준화, AI Readiness

서 영 균 (Young-Kyoon Suh)

2005년 2월 : 한국과학기술원 컴퓨터과학(석사)
2015년 5월 : 美 Univ. of Arizona 컴퓨터과학(박사)
2017년 9월 ~ 현재 : 경북대학교 컴퓨터학부 정보과학과 부교수
관심분야 : 데이터 사이언스, 엣지 컴퓨팅, 응용 AI