DSRI
Data
Research
Brief

Vol. 2026-2
ISSN 2765-1002(Online)

데이터 리서치 브리프

국가데이터처
국가데이터연구원

* 표지 이미지는 생성형 AI 활용 제작

DSRI
Data
Research
Brief

vol. 2026-2
ISSN 2765-1002(Online)

구글리서치의 데이터 품질관리 연구 동향과 시사점

김정민
국가데이터연구원
데이터과학연구팀

- **(검토 목적)** 구글리서치의 **최근 5년간 데이터 품질관리 연구 동향을 심층 분석**해 국내 차원의 AI 데이터 관리 체계 마련을 위한 향후 추진 방향을 모색하였다.
- **(핵심 내용)** 구글의 **데이터 품질 관련 주요 연구 16건**을 총 4개 유형으로 분류하고 유형별 **세부 연구 내용 및 시사점 분석하였다.**
- **(향후 활용)** 본 연구에서 도출된 시사점은 **국가데이터를 'AI 신뢰자산'化**하고, **효용성 있는 AI 데이터 품질관리 기준을 도출**하기 위한 **기초자료로 활용** 가능하다.

연구 배경 및 목적

❍ **(검토 배경)** 최근 **국내외 AI 경쟁력의 축**이 '모델(Model-Centric)'에서 **'데이터(Data-Centric)'로 이동**하면서, 고품질 AI 데이터의 중요성 상승*

* 글로벌 AI 기업은 모델 경쟁이 아닌 데이터 품질 강화로 전략을 선회하고 있으며, 큰 데이터(Bic-Data)보다 고품질 데이터(Good-Data)의 중요성을 강조

❍ **(문제 인식)** 그러나 현실은 여전히 **모델 중심적 관행이 유지***되는 양상

* 데이터 중요성을 인지하고 있으나 AI 데이터에 대한 관리 체계가 모호하고, 고품질 데이터의 정의가 주관적이며, 표준화된 문서화 방법도 정립되지 않은 상황

❍ **(검토 필요성) AX 시대**에는 **데이터 생산 및 관리 시 AI 활용을 전제**해야 하며, 이를 위해서는 관련 연구를 **선도하는 기관*의 선제적 검토 사례를 참고**할 필요

* 구글리서치는 오랜 기간에 걸쳐 AI 데이터 관리의 중요성을 제언해 온 기관

구글리서치 개요 및 검토 전략

❍ **(기관 개요)** 구글의 산하 연구소로서, 데이터 우수성(Data Excellence) 철학 전파와 '데이터 함정' 예방 가이드 배포 등 **데이터 품질 논의를 선도**

❍ **(연구 목표) 구글리서치가 데이터에 주목하는 배경**에는 **세 가지 문제 인식**이 존재

① **알고리즘 포화:** 모델 고도화만으로는 AI 성능에 한계가 있음이 확인

엔듀르 응
(스탠포드大 교수)

*"AI 시스템의 성능은 코드와 데이터의 조합입니다. 알고리즘의 발전 덕분에 이제는 코드보다 **데이터에 더 많은 시간을 쏟아야 할 때**입니다"*

*"AI 성능이 기대처럼 나오지 않을 때, 코드를 수정하는 비중은 줄어들고, **데이터를 정제하는 비중이 늘고 있습니다.**"*

안드레 카파시
(OpenAI 창립맴버)

② **비체계적 데이터 관리:** 만성적인 데이터 업무 저평가 인식 및 문서화 부재

니티야 삼바시반
(前구글 딥마인드 Ph.D)

*"**데이터 작업은 AI의 성패를 좌우하는 결정적 요소**임에도 불구하고, **낮은 지위의 업무로 간주**됩니다. 이는 결국 AI 모델의 실패로 이어집니다."*

*"과거의 **데이터 엔지니어**들이 단순히 파이프라인을 만드는 조연이었다면, 이제 그들은 **AI를 실질적으로 가능하게 하는 핵심 주역입니다.**"*

세미 벤지오
(애플 AI 연구 디렉터)

③ **데이터 절벽:** AI 데이터의 수요가 인간의 생산 속도를 추월해 데이터의 고갈 우려가 현실화

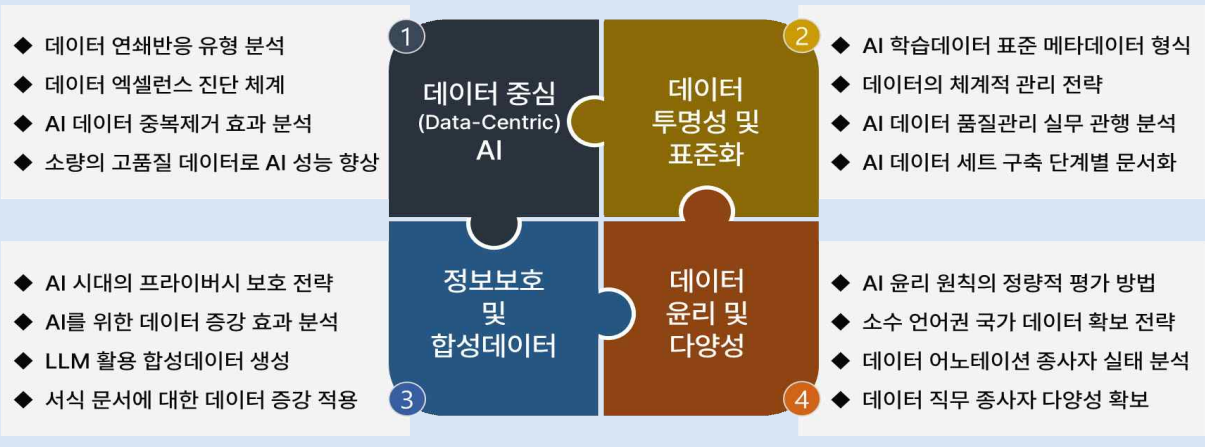
일리야 수츠케버
(GPT-4 개발 총괄)

*"**데이터도 유한한 자원입니다.** 인간이 만든 콘텐츠만으로는 더 이상 모델의 성능을 높이기 어려울 것입니다."*

*"**인간 지식의 총합은 AI 학습 측면에서는 이미 고갈되었습니다.** 사실상 작년(2024년)에 그런 일이 벌어졌습니다."*

일론 머스크
(테슬라, X 대표)

❍ **(검토 전략)** 구글리서치가 **최근 5년간 발행**한 AI 학술 논문 335건 중 서로 다른 관점과 목표를 둔 **데이터 품질 관련 연구 16건을 선별**



검토 결과

❍ **(데이터 중심 AI)** AI의 **성패가 데이터 품질에 좌우**된다는 사실을 **입증·공론화**하는 데 주력(FGI, 전문가 자문, 데이터 전처리 효과 실증 등)

☞ **핵심 용어**

- **데이터 연쇄반응(Data Cascades):** AI 데이터 품질 문제가 원인이 되어 발생하는 연쇄적 부정적 파급효과. 눈에 잘 띄지 않고·지연 발현되며·만연하나, 품질관리로 예방 가능
- **데이터 우수성(Data Excellence):** 책임 있는 수집·저장·사용, 시간 경과에도 유지, 응용 분야 전반 재사용, 경험적·설명력을 갖춘 '우수한' 데이터를 지향하는 철학

ⓐ **데이터 연쇄반응(2021):** AI 데이터 품질 문제로 인한 연쇄적이고 부정적인 파급효과를 정의하고, 전문가 심층 인터뷰를 활용해 발생 유형을 분석*

* AI 평가를 '적합성(fit)' 중심에서 '데이터 품질(data)' 중심으로 전환할 필요성 강조

ⓑ **데이터 엑셀런스(2022):** 100명 이상의 AI 분야 연구자 의견을 종합해, 우수한 데이터의 개념을 정의하고, 평가를 위한 거시적 척도 제안

ⓒ **데이터 중복제거(2022):** 비정형 데이터에 중복제거 기법을 적용한 결과, '과적합', '생성 텍스트 암기' 등 부정적 현상이 원본 대비 10% 수준으로 완화됨을 실증

ⓓ **고품질 데이터 선별(2025):** AI 학습데이터 규모를 약 250배 감소시켜도, 고품질 정보 선별 시 동등한 수준의 AI 모델 학습효과를 보일 수 있음을 증명*

* 구글은 자사 유해 광고 콘텐츠 판별 AI를 실제 운영하는 환경에 해당 데이터 선별 적용 시, 약 10,000배의 데이터 규모 절감이 가능할 것으로 시사

❍ **(데이터 투명성 및 표준화) 데이터의 전 생애주기**를 투명하게 **기록·문서화 하는 프레임워크를 개발**하고, 실무 적용을 통한 **효과성을 진단**

☞ **핵심 용어**

- **크루아상(Croissant):** 구글리서치가 제안한 AI 학습용 데이터의 메타데이터 형식 표준으로, Kaggle·Hugging Face·OpenML 등 글로벌 AI 플랫폼이 표준으로 채택
- **SDLC:** 소프트웨어의 생애주기를 요구사항 분석→설계→구현→테스트→유지보수의 단계로 설명한 소프트웨어 공학 이론

ⓐ **SDLC 기반 문서화(2021):** AI 데이터셋 구축을 SDLC에 대입*하여, 요구사항 분석~유지보수의 5단계 별 문서화 전략을 제시

* 데이터셋을 기술적 인프라로 간주함으로써 데이터와 소프트웨어의 생애주기를 동일한 선상에서 비교 분석한 것이 특징

ⓑ **데이터 카드(2022):** 대내외 전문가 약 40명과 글로벌 AI 기업 12곳이 참여해, 형식 중심이 아닌 사용자 중심의 유연한 데이터셋 문서화 프레임워크 개발

검토 결과 (계속)

ⓒ **데이터셋 실무자 문제 규명(2024):** 구글 내 데이터셋 실무자 인터뷰를 통해, 현업에서 AI 데이터 품질 관리가 어려운 5가지 시스템적 원인* 식별
* 품질 기준 모호, 직관 중심의 판단, 엑셀 의존성, 표준화 도구 부재, 확증 편향

ⓓ **크루아상(2024):** AI 데이터의 메타데이터 형식 표준으로서, Kaggle, Hugging Face, OpenML 등 글로벌 플랫폼이 표준으로 채택

<참 고> 크루아상(Croissant)의 메타데이터 계층 구조

계 층	세부 내용
① **데이터 세트 계층**	명칭·개요·버전·라이선스 등 데이터 세트의 기본정보
② **리소스 계층**	형식적 구조(FileObject: 파일 URL·인코딩, FileSet: 파일별 범주)
③ **구조 계층**	콘텐츠 정보와 표현 방식(정형·비정형, Field: 개별 정보)
④ **의미 계층**	AI 활용에 유용한 정보(데이터 세트 분할 추천·주요 유형 등)
⑤ **RAI 확장 계층**	신뢰성·윤리 보장을 위한 정보

❍ **(정보보호 및 합성데이터 방법론)** 데이터 고갈 현상에 대응해 **합성데이터를 활용**하고, 민감정보를 다루는 **최신(Emerging) 보안 기술의 적용을 실증**

☞ **핵심 용어**
- **합성데이터(Synthetic Data):** 실제 데이터의 통계적 특성과 패턴을 모방하여 AI 모델·알고리즘을 통해 인공적으로 생성한 가상 데이터
- **개인정보보호 강화기술(PETs):** 데이터 활용과 개인정보 보호의 균형을 위해 등장한 기술적 방법론·소프트웨어의 통칭(차등 개인정보보호, 동형 암호 등)

ⓐ **증강 데이터 효용성 평가(2021):** 증강 데이터의 품질 측정 방법을 제안(친화도(Affinity)·다양성(Diversity)하고, 적정 수준의 증강 기술 활용을 권고

ⓑ **LLM 합성데이터 생성(2023):** LLM만을 이용해 합성데이터 셋을 구축 후 AI를 학습시켜도 성능 향상 효과가 있음을 입증*
* 단, 연구자 직접 검증 결과 합성데이터의 11%에서 오류가 존재하는 것으로 드러나, 인간의 개입이 필요함을 시사

ⓒ **필드스왑(2024):** 정형 서식 문서의 핵심 구문을 자동 식별해 정합성 높은 합성데이터를 생성하는 기술을 제안*
* 실험 결과, 100건 이하의 원본만으로도 AI 성능(F1)이 최대 11점 상승

ⓓ **프라이버시 보호 기술(2025):** 개인정보보호와 데이터 활용 간 제로섬 관계를 해결할 수 있는 미래 기술군(PETs*)을 제시하며, 차세대 필수 기술로 소개
* 차등 개인정보보호, 생성적 적대 신경망(GAN), 동형 암호 등

검토 결과 (계속)

❍ **(데이터 윤리 및 다양성)** AI 데이터를 **'만드는 사람'의 관점**에서 생산 **품질을 강화**하고, 추상적 **윤리 개념을 '측정 가능한' 개념으로 전환**하는데 관심

☞ **핵심 용어**

- **데이터 어노테이션(Data Annotation):** 사전 정의된 지침에 따라 데이터에 의미를 부여(주석)하는 과업으로, 고품질 AI 학습용 데이터 구축의 핵심 공정
- **AI 윤리(AI Ethics):** AI의 설계·개발·배포·사용이 인간의 가치·권리·사회 규범에 부합하도록 보장하는데 중점을 둔 연구·실천 영역

ⓐ **어노테이션 종사자 실태(2022):** 데이터 종사자 47명을 대상으로 한 인터뷰를 통해, AI 시장 확대의 혜택이 근로자에게 충분히 돌아가지 않는 현실 진단

ⓑ **어노테이션 종사자 다양성(2023):** 종사자 다양성이 데이터 품질에 영향을 미치나 실제 인력 확보에 반영되지 않는 '인식과 실천의 괴리'를 규명*

* 본 연구는 다양성 경시 문화가 데이터 품질을 하락시키는 요인임을 시사

ⓒ **소수 언어권 데이터 확보(2024):** 소수 언어권 공동체(예: 마오이족)를 윤리적 방식으로 설득해 데이터를 확보하기 위한 5대 권고안*을 제시

* 인권·언어권, 공동체 중심 연구, 관계적 윤리, 주권·통제, 문화적 해석

ⓓ **AI 윤리 정량 평가(2025):** '11~'23년까지 발행된 학술 논문의 AI 윤리 평가 지표 791개를 11대 윤리 원칙별로 집대성*

* 윤리 측정 방법을 목록화한 최초 시도로 참고 가치가 높으나, 과거 연구 문헌을 기반으로 해 최신 AI 모델에 적용할 수 있는 지표는 소수

시사점

❍ **(총평)** 구글리서치의 연구는 AI 데이터 관리 체계 마련을 위해 고민해야 할 **다방면의 주제를 포괄**하고 있다는 점에서 **분석 의의가 있음**

❍ **(실천 가능성 측면)** 구글이 소개한 다양한 방법론의 **실제 적용을 위해**서는 **투입 자원**(인적자원, 소요 비용, 기대효과 등)에 대한 **면밀한 검토가 필요**

⇨ **AI 데이터 관리 체계 마련은 '신뢰성'과 '관리 부담' 간의 절충점을 탐색하는 과정**

한계 및 향후 연구

❍ **(한계) 본 검토는 구글리서치의 연구에 한정**되므로, 국제 규범(ISO/IEC, NIST AI RMF, EU AI Act)과의 비교 검증을 통해 **포괄 범위를 확장하는 것이 효과적**

❍ **(향후 연구)** 국내 데이터 생산 주체의 여건을 고려한 **실천 가능한 AI 데이터 관리 기준·범위 도출**을 위한 **실증 연구가 필요**

- AI 데이터 관리 수준별 'AI 성능 향상'과 '관리 비용'의 절충점(Sweet Spot)을 탐색하는 실증 연구 추진을 계획 중('27년)

<참 고> 데이터 중심 AI를 위한 실증 연구 추진 계획(안)

참고 문헌

○ Sambasivan et al.(2021), Everyone wants to do the model work, not the data work
○ Aroyo et al.(2022), Data Excellence for AI : Why Should You Care
○ Nystrom et al.(2022), Deduplicating Training Data Makes Language Models Better
○ Krause et al.(2025), Achieving 10,000x training data reduction with high-fidelity labels
○ Hutchinson et al.(2021), Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure
○ Zaldivar et al.(2022), Data Cards: Purposeful and Transparent Dataset Documentation
○ Akhtar et al.(2024), Croissant: A Metadata Format for ML-Ready Datasets
○ Qian et al.(2024), Understanding the Dataset Practitioners Behind LLMs
○ Cubuk et al.(2021), Tradeoffs in Data Augmentation
○ Gekhman et al.(2023), TrueTeacher: Generating Synthetic Data using LLMs
○ Xie et al.(2024), FieldSwap: Data Augmentation for Effective Form-Like Document Extraction
○ Amin et al.(2025), Google's Approach to Protecting Privacy in the Age of AI
○ Wang et al.(2022), Whose AI Dream? In search of the aspiration in data annotation
○ Kapania et al.(2023), Annotator Diversity in Data Practices
○ Wudiri et al.(2024), Socially Responsible Data for Large Multilingual Language Models
○ Rismani et al.(2025), Responsible AI measures dataset for ethics evaluation

※ 데이터 품질관리 연구 동향에 대해 국가데이터연구원에서 자체적으로 연구 · 검토한 내용임

DSRI
Data
Research
Brief

vol. 2026-2
ISSN 2765-1002(Online)

데이터 통합을 활용한 경제구조통계 개선

- 전처리와 자료 비교 가능성이 통계 활용 가능성에 미치는 영향 -

박성률
국가데이터연구원
데이터방법연구실

김민규
국가데이터연구원
데이터방법연구실

● **(연구 목적)** 행정자료와 조사자료를 결합할 때 전처리와 자료 간 비교 가능성이 결합 가능성 및 통계 활용성에 미치는 영향을 분석하였다.

● **(분석 자료)** 공정거래위원회 통신판매업 자료, 경제총조사 모집단 명부 및 경제총조사 시범예행조사 자료를 이용하였다.

● **(핵심 결과)** 사업자등록번호 정확 연계율은 3.86~10.41%였으며, 사업체명·대표자명·주소를 활용한 결정적 연계의 추가 성과는 거의 없었다.

● **(핵심 해석)** 연계 레코드에서도 핵심 분석 변수의 불일치율이 높아 기술적 결합이 곧 통계적 활용 가능성을 의미하지 않음을 확인하였다.

● **(함의)** 외부 자료의 국가통계 활용을 위해 결합 알고리즘 고도화뿐 아니라 자료수집, 표준화, 전처리 및 메타데이터 관리가 선행될 필요가 있다.

연구 배경 및 목적

❍ **(연구 배경)** 경제구조통계는 경제 현황과 산업 구조를 파악하고 정책 수립과 경제 분석을 지원하는 핵심 국가통계
- 응답률 저하, 응답 부담 증가 및 조사비용 상승으로 조사 중심 통계생산 방식의 지속가능성에 대한 문제가 제기
- 행정·민간 자료를 기존 조사자료와 결합하여 활용하려는 시도가 확대되고 있음
- 서로 다른 목적으로 생성된 자료는 모집단 범위, 변수 정의, 자료구조, 기준 시점 및 기록 방식이 상이할 수 있음
- 이 차이가 충분히 조정되지 않으면 자료의 양이 증가하더라도 실제로 결합할 수 있는 레코드와 통계분석에 활용할 수 있는 레코드의 범위는 제한될 수 있음

❍ **(연구 목적)** 행정자료와 경제총조사 자료의 결합 과정에서 전처리와 자료의 구조적 조건이 결합 가능성 및 통계 활용 가능성을 어떻게 제약하는지 실증적으로 확인

☞ **(핵심 질문)** 레코드가 기술적으로 결합되었다는 사실만으로 해당 결합자료를 국가통계 작성에 활용할 수 있는가를 검토함

데이터 및 방법

❍ **(활용 데이터)** 생성 목적과 모집단 범위가 서로 다른 세 가지 사업체 자료를 활용
- 공정거래위원회 통신판매업 자료, 경제총조사 모집단 명부, 경제총조사 시범예행조사 자료
- 통신판매업 자료에는 영업 중인 사업체와 폐업 사업체가 함께 포함될 수 있음
- 경제총조사 자료는 조사 목적에 부합하는 활동 사업체를 중심으로 구축됨

❍ **(전처리)** 자료 간 구조적 차이를 확인하고 연계 가능성을 높이기 위해 자료 진단, 스키마 정렬 및 개념 조화의 세 단계로 전처리 수행
- **(자료 진단)** 자료별 모집단 범위와 레코드 수를 비교하고 결측치, 중복 레코드, 조사 대상 외 사업체 및 식별·분석 변수의 이용 가능성 점검
- **(스키마 정렬)** 자료별 변수명, 자료형 및 표현 형식을 비교하고 사업자등록번호, 상호명, 대표자명 및 주소의 대응 관계를 설정
- **(개념 조화)** 변수 개념과 코드 체계를 통일하고 상호명·주소를 표준화하며 결측, 해당 없음 및 구조적 미수집 값을 구분함

❍ **(이력관리)** 전처리 과정의 변수 변환과 대응 관계를 변환표와 변경 이력으로 기록하여 재현성과 추적 가능성을 확보함

❍ **(레코드 연계)** 정확 연계와 규칙 기반 결정적 연계의 두 단계로 수행
- **(정확 연계)** 사업자등록번호가 정확히 일치하는 레코드를 동일 사업체로 판정
- **(결정적 연계)** 미연계 레코드에 대해 사업체명, 대표자명, 주소를 대상으로 Jaro-Winkler 유사도 0.90 이상인 경우에만 연계

❍ **(평가 지표)** 연계율, 추가 연계율, 일치율, 불일치율, 정밀도, 재현율 산출

❍ **(해석 유의)** 별도 수작업 검증자료가 없어 정밀도와 재현율은 사업자등록번호 정확 연계 결과를 참조 집합으로 한 일관성 지표로 해석함

주요 결과

❍ 결과는 결합 단계에서 갑자기 문제가 발생한 것이 아니라, 결합 이전의 모집단 차이와 변수 구조의 불완전성이 전처리 손실, 낮은 연계율, 결정적 연계의 제한 및 연계 후 변수 불일치로 이어지는 누적 구조를 보임

❍ **(전처리) 전체 자료의 약 29~44% 제외**
- 폐업 사업체, 조사 대상 외 사업체, 중복 레코드 및 주요 변수 누락 자료를 제외한 결과, 자료별로 전체 레코드의 28.73~43.87%가 분석 대상에서 제외

<표1> 자료별 규모 및 제외 현황

지 역	자 료	전체 레코드	제외 레코드	제외율	활용 레코드
부산 북구	통신판매업	5,925	2,534	42.77%	3,391
	모집단 명부	23,250	7,455	32.06%	15,795
	시범예행조사	9,837	3,423	34.80%	6,414
경북 경산시	통신판매업	10,769	3,094	28.73%	7,675
	모집단 명부	34,844	12,893	37.00%	21,948
	시범예행조사	16,418	7,202	43.87%	9,216

주요 결과 (계속)

- 주요 제외 사유
  · 통신판매업 자료는 폐업 사업체
  · 경제총조사 모집단 명부와 시범예행조사 자료는 조사 대상 범위에 포함되지 않는 사업체가 주된 제외 사유
- 이는 자료별 모집단 범위가 동일하지 않으며, 결합 이전 단계에서 이미 실제 결합 가능한 레코드의 범위가 크게 제한되었음을 의미함

❍ **결합 변수와 분석 변수의 완전성이 서로 다르게 분포**

- 결합에 사용되는 변수와 통계분석에 필요한 변수의 완전성이 동일한 자료에 함께 확보되지 않았음
  · **(결합 변수)** 통신판매업 자료의 사업자등록번호 결측률은 부산 북구 0.03%, 경북 경산시 0.29%로 낮았음
  · **(분석 변수)** 시범예행조사 자료의 온라인 거래 여부 결측률은 부산 북구 37.09%, 경북 경산시 54.25%로 높았음
- 결합 변수가 완전한 자료에는 분석 변수가 부족하고, 분석 변수를 포함한 조사 자료에는 결측이 많아 레코드 연계와 통계 활용을 동시에 충족하는 자료 범위가 추가로 축소됨

<표2> 자료 및 변수별 결측 현황

자 료	변 수	부산 북구	경북 경산시
통신판매업	사업자등록번호	0.03%	0.29%
	주소	0.12%	0.31%
모집단 명부	사업자등록번호	4.00%	1.96%
	대표자명	0.09%	0.05%
시범예행조사	사업자등록번호	3.95%	1.92%
	온라인 거래 여부	37.09%	54.25%

❍ **(정확 연계) 사업자등록번호를 이용한 연계율은 3.86~10.41%에 그침**

- 고유식별자인 사업자등록번호가 존재함에도 연계율이 낮게 나타난 것은 동일 사업체가 각 자료에 동시에 포함되는 비율이 낮았기 때문임
- 즉, 연계율 제한은 식별자 성능보다 자료 간 모집단 비중첩의 영향을 크게 받았음

<표3> 통신판매업 자료 기준 정확 연계 결과

자 료	결합 대상	연계 레코드	연계율
부산 북구	시범예행조사	163	4.81%
	모집단 명부	353	10.41%
경북 경산시	시범예행조사	296	3.86%
	모집단 명부	588	7.66%

주요 결과 (계속)

❍ **(결정적 연계) 규칙 기반 결정적 연계의 성과는 거의 나타나지 않음**

- 미연계 레코드를 대상으로 사업체명, 대표자명 및 주소를 활용, 결정적 연계 수행
- 추가 연계 건수는 경북 경산시의 모집단 명부 1건을 제외하고 모두 0건

<표4> 통신판매업 자료 기준 결정적 연계 결과

자료	결합 대상	미연계	추가 연계 레코드	추가 연계율
부산 북구	시범예행조사	3,228	0	0.00%
	모집단 명부	3,038	0	0.00%
경북 경산시	시범예행조사	7,379	0	0.00%
	모집단 명부	7,087	1	0.01%

- 이는 미연계가 단순한 표기 차이에서 발생한 것이 아니라 모집단 비중첩, 식별정보 누락 및 자료 표현의 구조적 차이에서 비롯되었음을 보여줌
- 따라서 연계 알고리즘을 정교화하는 것만으로는 결합 성과를 크게 개선하기 어려움

❍ **연계 레코드에서도 분석 변수의 일관성이 낮음**

- 시범예행조사 자료와 연계된 레코드의 온라인 거래 여부를 비교한 결과
  · **(변수 일치율)** 부산 북구 27.53%, 경북 경산시 16.72%
  · **(불일치율)** 부산 북구 72.47%, 경북 경산시 83.28%

<표5> 통신판매업 자료 기준 연계 품질 점검 결과

자 료	분석 변수 일치율	불일치율	정밀도	재현율
부산 북구	27.53%	72.47%	88.67%	23.93%
경북 경산시	16.72%	83.28%	90.91%	13.56%

- **(연계 성능)** 엄격한 기준을 적용함에 따라 정밀도는 비교적 높게 나타났으나, 재현율은 낮음
  · 오연계는 줄일 수 있었지만 실제로 연결되어야 할 레코드의 상당 부분을 식별하지 못했음을 의미
  · 동일한 사업체로 연계된 경우에도 온라인 거래 여부가 서로 다른 비율이 높음
- 레코드가 동일 사업체로 연계되었다는 사실만으로 결합자료의 통계적 활용 가능성을 판단할 수 없으며, 결합 이후에도 핵심 분석 변수의 일관성을 별도로 검증해야 함을 의미함

시사점

❍ **데이터 통합의 성과는 결합 방법보다 자료 조건에 크게 좌우됨**
- 자료 간 모집단 범위와 변수 구조가 일치하지 않으면 고유식별자가 존재하거나 결합 알고리즘을 고도화하더라도 결합률을 크게 개선하기 어려움
- 데이터 통합의 가능 범위는 결합 단계 이전의 자료 수집과 전처리 과정에서 상당 부분 결정

❍ **기술적 결합 가능성과 통계적 활용 가능성을 구분해야 함**
- 동일 사업체가 기술적으로 결합되더라도 핵심 분석 변수가 누락되거나 자료별 값이 일치하지 않으면 통계분석에 활용하기 어려움
- 품질을 평가할 때 결합률뿐만 아니라 모집단 포괄성, 분석 변수의 완전성 및 결합 후 변수 일관성을 함께 평가해야 함

❍ **외부 자료의 국가통계 활용을 위해 자료 표준화가 선행되어야 함**
- 외부 자료를 국가통계에 활용하기 위해서는 사업체명, 대표자명, 주소 및 사업자 상태와 같은 핵심 변수의 정의와 표현 방식을 표준화할 필요가 있음
- 자료 제공기관과 통계작성기관 간 메타데이터 기준을 마련하는 것도 중요함

❍ **전처리를 단순한 자료정제 작업이 아닌 품질 통제 단계로 관리해야 함**
- 전처리는 결합을 위한 기술적 준비 과정에 그치지 않고 실제로 결합이 가능한 자료와 분석이 가능한 자료의 범위를 결정하는 핵심 통제 단계
- 변수 대응 관계, 변환 규칙, 제외 기준 및 결측 처리 방법을 명시하고 변경 이력을 체계적으로 관리해야 함

❍ **데이터 통합의 품질관리는 전체 통계 생산 과정에서 수행되어야 함**
- 자료수집, 개념 정의, 표준화, 전처리, 결합 및 결과 검증은 상호 독립된 절차가 아님
- 초기 단계의 자료 불일치가 이후 결합과 분석 단계로 전이될 수 있으므로, 데이터 통합의 품질관리는 개별 결합 방법이 아니라 전체 통계 생산 과정과 데이터 관리 책임 체계(data stewardship)를 중심으로 설계할 필요가 있음

한계 및 향후 연구

❍ **본 연구 결과를 전국 또는 다른 산업 분야에 직접 일반화하는 데 한계가 있음**
- 별도의 수작업 검증자료가 확보되지 않아 사업자등록번호를 이용한 정확 연계 결과를 참조 자료로 사용
- 제시한 정밀도와 재현율은 실제 결합 정확도보다는 참조 자료와의 일관성을 나타내는 지표로 해석해야 함
- 통신판매업 자료의 수집 시점, 갱신 주기 및 운영상 오류가 결합 결과에 미치는 영향도 충분히 분석하지 못함

❍ **향후 연구에서는 다음 사항을 검토할 필요가 있음**
- 전국 단위 및 다양한 산업 분야로 분석 범위 확대
- 수작업 검증 표본을 활용한 연계 정확도 평가
- 연계 임곗값에 대한 민감도 분석
- 확률적 연계 및 기계학습 기반 연계 방법과의 비교
- 자료수집 시점과 갱신 주기를 고려한 시점 정합성 평가
- 메타데이터 표준화 방안 마련
- 데이터 이력 및 처리 과정을 관리하는 데이터 계보(data lineage) 체계 구축
- 결합 이후 결측 대체와 통계적 보정의 적용 가능성 검토

결론

❍ 본 연구는 낮은 연계율 자체보다 그 원인이 결합 이전의 모집단 차이, 변수 완전성 부족 및 자료 표현의 구조적 차이에 있다는 점을 실증적으로 확인
- 또한 연계된 레코드에서도 핵심 분석 변수의 불일치가 높아, 결합 성공과 통계적 사용 가능성을 동일하게 간주해서는 안 됨

❍ 외부 자료를 국가통계에 활용하기 위해서는 결합 알고리즘의 고도화만을 우선할 것이 아니라, 자료수집 단계의 모집단 정의, 변수 표준화, 전처리 규칙, 메타데이터 및 처리 이력, 결합 후 변수 일관성 검증을 포함하는 전 과정 품질관리 체계를 구축해야 함

❍ 데이터 통합의 핵심은 자료를 결합하는 데 있는 것이 아니라, 결합된 자료가 통계적으로 해석 가능하고 재현 가능한 방식으로 관리되었는지를 확인하는 데 있음

참고 문헌

○ Christen, P. (2012). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer

○ Eurostat. (2019). Quality assurance framework of the European Statistical System: Version 2.0. Publications Office of the European Union

○ Herzog, T. N., Scheuren, F. J., & Winkler, W. E. (2007). Data quality and record linkage techniques. Springer

○ Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), 334-350

○ Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13

○ United Nations Economic Commission for Europe. (2019). Generic Statistical Business Process Model: GSBPM version 5.1

○ United Nations Economic Commission for Europe. (2020). A guide to data integration for official statistics: Version 2.0

○ Winkler, W. E. (2006). Overview of record linkage and current research directions (Research Report Series No. RRS2006/02). U.S. Census Bureau

⇨ (**연구보고서 원본**) 국가데이터연구원홈페이지 > 연구간행물> 연구보고서 > **데이터과학 기술을 활용한 경제구조통계 자료수집 개선 연구**

DSRI
Data
Research
Brief

vol. 2026-2
ISSN 2765-1002(Online)

AI통계분류의 확장 가능성: 산업·직업분류를 넘어서

이규호
국가데이터연구원
데이터과학연구팀

- **(연구 목적)** 본 연구는 건설경기동향조사의 공종(19종), 발주자(27종) 분류코딩 작업의 자동화를 위해 **ELECTRA 기반 한국어 사전학습 모델(KcELECTRA)**을 활용한 인공지능 분류 모델을 **시험 구축**하고 실무 활용성을 사전 검증하였다.
- **(핵심 결과)** 경인지방데이터청의 10년간 수주자료 **약 9만 건**을 학습데이터로 활용하고, 2025년 실제 수주 데이터 3,254건으로 모델의 분류 정확도를 검증하였다. 그 결과, 일반적인 분류(Flat) 모델에서는 공종 88%, 발주자 43%의 정확도를 보였으나, **계층형 다중 작업 학습(Hierarchical Multi-Task Learning)** 방법론을 적용하여 **공종 91%, 발주자 94%**로 정확도를 대폭 향상시켰다.
- **(향후 활용)** 본 연구 결과는 지방청 단위의 분류코딩 업무 효율화에 활용될 수 있는 가능성을 보여주었으며, 나아가 **한국표준산업분류·직업분류 외 기타 분류체계를 가진 통계조사**에 대한 **AI통계분류 확산에 기여**할 수 있는 가능성을 보여주었다.

연구 배경 및 목적

❍ **(연구 배경)** 국가데이터처는 **2022년 인공지능 통계분류 자동화 시스템(AI통계분류시스템)을 구축**하여 인구총조사, 지역별고용조사, 전국사업체조사, 건설업조사 등 대규모 조사에 활용하고 있고, 산업·직업 분류 외 자체 분류체계를 사용하는 조사에도 **실무적용을 확대**하고 있음

[그림1] 인공지능기반 통계분류 자동화 시스템

연구 배경 및 목적 (계속)

❍ **(시스템 배경)** AI통계분류시스템은 ELECTRA, Roberta 등 **사전학습언어모델 기반 지도학습 방법론**을 통해 사업체명, 사업체가 하는일, 근무 부서, 직무, 내가 하는일 등 **조사표 텍스트로부터** 한국표준산업분류·직업분류 코드를 **자동 부여하는 체계**로, 나라통계시스템과 연계되어 연간 1천만 건 이상의 분류 코딩 업무를 지원함

❍ **(문제 인식)** 건설경기동향조사는 종합건설업체의 국내 건설공사 수주액 및 기성액을 공종별·발주자별로 파악하여 건설활동의 단기동향을 신속하게 제공하는 **지정통계 (승인번호 제101016호)**임. 매월 청별로 약 1,500건의 수주 데이터가 입수되며, 공사·발주자명 등 텍스트 자료를 담당자가 수동으로 검토하여 분류코드를 부여하는 작업이 반복적으로 수행되어 담당자 업무 부담과 분류 일관성 확보의 어려움이 상존함

❍ **(연구의 필요성)** 2021년 경인지방데이터청에서 **KoBERT 모델을 사용하여 자동 분류를 실무에 적용하기 위한 시도**를 하였으나, 실제 운용 정확도가 **공종 30.6%, 발주자 20.9%**에 그쳐 실무 적용에 한계가 있음

[그림2] 경인지방데이터청 건설 공종 발주자 부호 내검 프로그램

❍ **(연구 목표)** 본 연구는 ELECTRA 기반 사전학습 언어모델을 활용하여 건설경기동향조사의 공종(19종) 및 발주자(27종) 분류코딩을 자동화하는 AI 모델을 시험 구축하고, 실무 활용 가능성을 검증하는 것을 목적으로 진행되었음

데이터 및 방법

❍ **(활용 데이터)** 경인지방데이터청에서 입수한 2015년 1월부터 2025년 8월까지의 건설경기동향조사 수주자료 **총 90,994건**을 활용하였음. 주요 피처(Feature)로는 공사명, 기업체명, 공사지역, 발주자명 등의 텍스트 정보를 사용하였고, 정답 라벨(Label)은 **공종 부호(3자리, 19종)와 발주자 부호(4자리, 27종)**임

<표1> 데이터 구성

구 분	기 간	건 수	용 도
학습셋(Training)	2015.1. ~ 2024.12.	80,155건	모델 학습
검증셋(Validation)	2015.1. ~ 2024.12.	8,907건	학습 중 성능 확인
평가셋(Test)	2025.1. ~ 2025.8.	3,254건	최종 성능 평가
합계	-	**92,316건**	-

❍ **(AI 모델 선정)** 사전학습 언어모델인 **KcELECTRA-base**(한국어 ELECTRA)를 기반으로 지도학습 방식의 텍스트 분류 모델을 구축하였음. ELECTRA는 구글이 개발한 Transformer 기반 모델로, **기존 BERT 대비 약 1/4 수준의 컴퓨팅 리소스**로도 유사한 성능을 달성할 수 있다는 장점이 있음

❍ **(분석 방법)** 모델 평가에는 정분류율(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score(macro)를 활용하여 분류 단위별 성능을 종합적으로 분석하였음. 과적합 방지를 위해 **조기 종료(Early Stopping)와 Checkpoint 기법**을 적용하였음

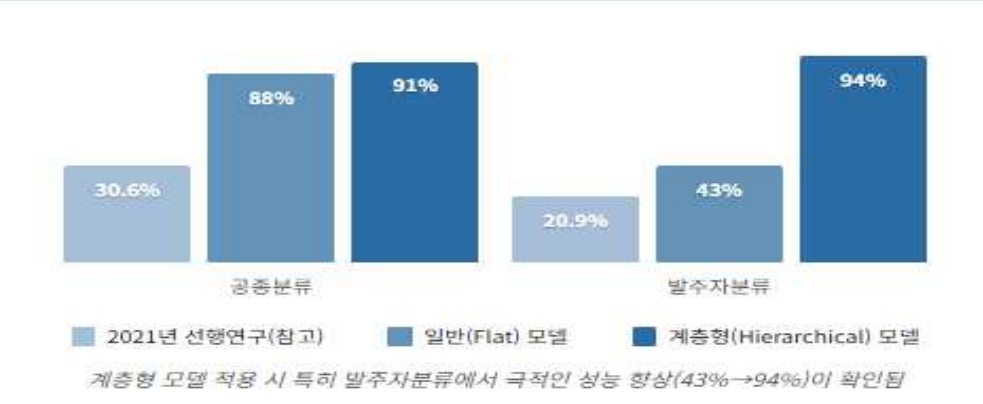
[그림3] 분석 프로세스 개요

주요 결과

❍ **(시행 착오)** 먼저 일반(Flat) 구조의 ELECTRA 모델로 시험 구축한 결과, 공종분류는 정분류율 88%(F1-macro 0.86)로 양호한 성능을 보였으나, 발주자 분류는 검증셋에서 93%를 기록했음에도 2025년 실제 평가 데이터에서는 **43% (F1-macro 0.25)로 급격히 저하**되었음. 이는 발주자 분류 체계 27종 중 **민자유치사업 11개 코드의 학습 데이터가 2% 미만**에 불과한 극심한 불균형에 기인했음. 이를 보완하기 위해 공사면 텍스트를 피처에 추가하거나(정확도 43%, Precision 감소), 3자리 분류와 키워드 조건 처리를 결합하는 방식(정확도 93%이나 키워드 의존성 과다)을 검토하였으나, 근본적 개선에는 이르지 못하였음

주요 결과 (계속)

[그림4] 모델별 정분류율(%) 비교 – 2025년 평가 데이터(3,254건) 기준



❍ **(최종 모델)** 이에 분류 코드의 일부 계층적 구조(예:2->21->215->2151)를 반영한 **계층형 다중 작업 학습(Hierarchical Multi-Task Learning)** 방법론을 적용하였음. 하나의 ELECTRA 모델에 분류 단위별 다중 분류기(Head)를 배치하여 계층별 코드를 동시에 학습하도록 설계한 결과, 발주자분류는 정분류율이 **43%에서 94%로** 대폭 향상되었고, 공종분류 또한 **88%에서 91%(F1-macro 0.91)로** 추가 향상되었음

<표2> 최종 모델(계층형) 성능 요약

구 분	학습데이터	평가데이터	정분류율	F1-macro	$Precision_{macro}$	$Recal_{macro}$
공종분류(19종)	80,155	3,254	**91%**	**0.91**	0.91	0.90
발주자분류(27종)	80,155	3,254	**94%**	**0.79**	0.80	0.80

시사점

❍ **(업무량 감소)** 분류코딩 업무의 실질적 자동화 기능을 확인하였음. 계층형 모델을 통해 **공종 91%, 발주자 94%**의 정분류율을 달성함으로써 지방청 담당자의 수동 코딩 작업 부담을 크게 줄일 수 있을 것으로 기대됨. 특히 **AI 예측 확률이 낮은 건(예:80% 미만)**에만 담당자가 집중 검토하는 방식으로 자료처리 업무의 효율성을 극대화하고 개선된 데이터를 재학습 해, 모델의 예측 정확도를 높일 수 있음

❍ **(계층형 방법론)** 계층형 분류 방법론의 유효성을 실증적으로 확인하였음. 발주자 분류와 같이 데이터 불균형이 심하고 계층적 구조를 가진 분류체계에 대해 계층형 다중 작업 학습이 매우 효과적임을 확인하였음. 이는 한국표준산업분류·직업분류뿐 아니라, **자체 분류체계를 사용하는 다양한 통계조사**에 대해서도 **AI 자동분류를 확대 적용할 수 있는 방법론적 기반**을 제공함

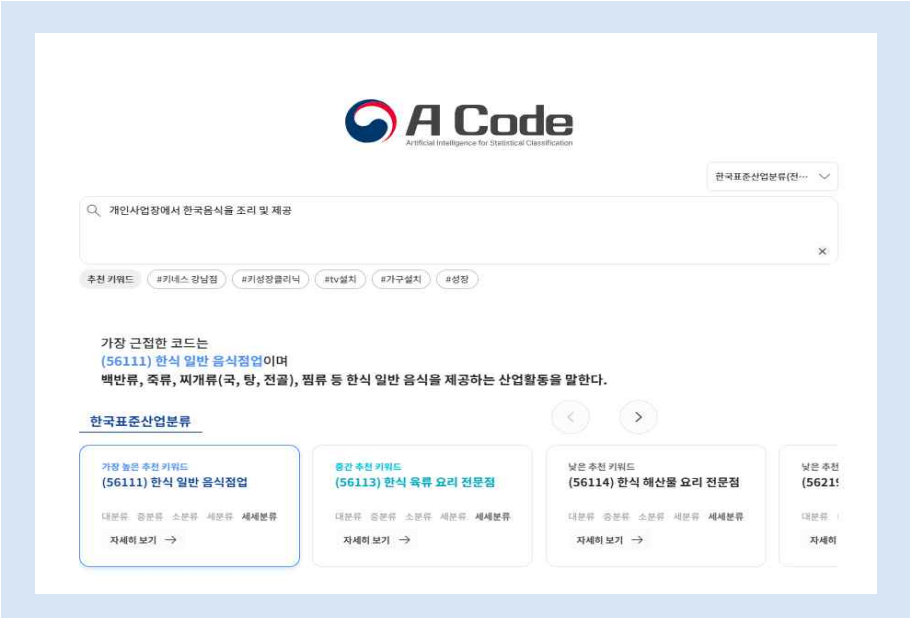
❍ **(실무 활용)** AI통계분류시스템의 적용 범위를 확장할 수 있음. 현재 5종의 대규모 조사에 한정된 시스템을 건설경기동향조사와 같은 월별 조사로 확대함으로써, 통계 생산 전반에 걸친 **AI 기반 자동화 체계(AX) 구축**에 기여할 수 있을 것으로 기대됨

한계 및 향후 연구

❍ **(학습데이터 한계)** 본 모델은 **경인지방데이터청의 수주 데이터**만으로 학습했다는 점에서 한계가 있음. 특히 발주자분류의 F1-macro(0.79)는 정분류율(94%)에 비해 상대적으로 낮은데, 이는 민자유치사업 등 소수 분류단위의 데이터 부족에 기인함. 향후 **전국 단위 데이터 확보**를 통해 소수 클래스의 학습 데이터를 보강할 필요가 있음

❍ **(향후 과제)** 전국 5개 지방청의 데이터를 모두 학습할 경우 지역별 공사·발주자 특성을 반영한 표준 모델로의 고도화가 가능할 것으로 기대됨. 실무 활용을 위해서 5개 지방청 전체 수주 데이터를 통합한 **전국 표준 모델 구축**을 통해 **나라통계시스템과의 실시간 연계**를 검토해 볼 수 있음

[그림5] AI통계분류시스템 메인 홈페이지



참고 문헌

○ 오교중 외 (2023), 인공지능 통계분류 자동화 확대 적용 연구, 국가데이터연구원

○ 임경민 (2023). AI 통계분류 결과분석 및 실무활용성 제고방안 연구, 국가데이터연구원

○ Kevin Clark et al. (2020). "ELECTRA: Pre-training text encoders as discriminators rather than generators."

○ Wehrmann et al. (2018). "Hierachical Multi-Label Classification Networks."

○ Yinhan Liu et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach."

⇨ (**연구보고서 원본**) 국가데이터연구원홈페이지 > 연구간행물> 연구보고서 > **AI기반 건설경기동향 자동분류 코딩 연구**

발 행 일 2026년 7월 31일
발 행 처 국가데이터처 국가데이터연구원
발 행 인 김 진
주 소 35220 대전광역시 서구 한밭대로 713 통계센터 6~8층
전화번호 042-366-7118
홈페이지 mods.go.kr/dsri