통계연구(2026), 제31권 제2호, 27-48/ DOI https://doi.org/10.22886/jkos.2026.31.2.27

공식 통계 생산을 위한 데이터 통합의 절차적 관리 프레임워크[1)]

김민규[2)] . 박성률[3)]

요약

공공 및 민간 자료의 활용이 확대됨에 따라, 데이터 통합은 공식 통계 생산의 핵심 인프라로 자리 잡고 있다. 기존 연구는 레코드 연계나 통계적 매칭 등 주로 개별 기법의 기술적 특성에 초점을 두어, 데이터 통합 전 과정에서 발생하는 오류와 불확실성을 어떻게 설계·관리할 것인지에 대한 논의는 상대적으로 제한적이었다. 이에 본 연구는 데이터 통합을 개별 기법의 집합이 아닌, 공식 통계 생산 과정에 내재한 절차적 관리 프레임워크로 정립한다. 본 연구는 통합 목적, 자료 구조, 통합 방법 차이를 설계 변수로 설정하고, 이를 바탕으로 데이터 통합을 「①전처리 및 정합성 점검 — ②결합 — ③결측값 대체 — ④보정 및 대표성 점검 — ⑤최종 품질 점검」의 5단계로 구조화하였다. 각 단계는 서로 다른 오류와 불확실성을 관리하고, 다음 단계로의 이행 여부를 판단하는 관리적 관문으로 기능한다. 또한 통합 목적, 자료 구조, 통합 방법에 따라 단계별 관리 강도와 판단 기준이 달라지는 조건부 구조를 가지며, 최종 품질 점검에서 확인된 문제는 이전 단계로 환류되어 통합 설계를 재조정하는 근거로 활용된다. 특히 본 연구는 과정 중심의 단계별 품질 점검과 결과 중심의 최종 평가를 결합한 이원적 품질관리 구조를 제시함으로써, 공식 통계 생산의 재현성과 투명성을 높이는 기틀을 마련하고자 한다.

주요용어 : 데이터 통합, 프레임워크, 품질관리, 공식 통계, 관리자

1. 서론

공공 및 민간[4)] 부문에서 다양한 자료(data source)[5)]가 생성·축적됨에 따라, 공식 통계

1) 본 연구는 김민규・박성률(2026)이 수행한 “데이터 통합 방법 체계화 연구”의 일부 내용을 수정・보완하여 작성하였습니다.

2) 제1저자, 교신저자, 국가데이터처 국가데이터연구원 데이터방법연구실, 통계연구사, E-mail: mgkim79@korea.kr

3) 제2저자, 국가데이터처 국가데이터연구원 데이터방법연구실, 통계연구관, E-mail: srpark08@korea.kr

4) 공공 자료(public sector data source)란 정부·공공기관이 생성·보유한 자료를 의미하며, 통계 목적으로 수집한 통계자료(statistical data)와 통계 목적 외의 행정자료(administrative data)를 포함한다. 민간 자료(private sector data source)란 기업 및 민간기관이 생성・보유한 자료로, 영리 목적 자료(commercial data)와 비영리 목적 자료(non-commercial data)를 포함한다.

5) 자료와 데이터는 용어 모두 데이터 통합의 핵심 요소이지만, 각각 담당하는 역할과 수준이 다르다. 정책·기술적 판단이 필요한 데이터 통합에서는 이를 엄밀히 구분하여야 한다. 자료(data source)는 가공되지 않은 1차 정보로, 설문 응답, 행정 기록, 센서 관측값 등이 이에 해당한다. 자료는 데이터의 출발점이자, 현실로부터 취득한 원천 정보이다. 데이터(data)는 자료를 일정한 기준에 따라 정의·분류·정제·구조화한 결과물이다. 자료가 “현상을 담은 원본”이라면 데이터는 “분석이 가능한 형태로 변환된 정보”라 할 수 있다. 요컨대, 자료는 데이터의 원천이며, 데이터는 통합의 기반이다.

생산 환경에서도 행정자료와 민간 자료의 활용이 확대되고 있으며, 서로 다른 출처의 자료를 결합하여 활용 가능성을 높이는 데이터 통합(data integration)이 통계 생산의 핵심 요소로 자리 잡고 있다. 이러한 변화는 고해상도 통계(high-resolution statistical data)[6] 및 맞춤형 통계 생산을 제고하는 한편, 자료 간 개념 불일치, 결합 오류, 대표성 저하와 같은 품질 위험을 동반한다. 그러나 기존 연구(*e.g.,* 이혜정 외, 2022; Cibella et al., 2009; Emmenegger et al., 2022; Liseo & Tancredi, 2009) 및 국제기구 지침(*e.g.,* UNECE, 2018; UN ESCAP, 2020)은 개별 통합 기법의 선택이나 결합 방법에 초점을 맞추어 왔으며, 데이터 통합 전(全) 과정에서 발생하는 오류와 불확실성을 체계적으로 관리하기 위한 절차적 틀을 충분히 제시하지는 못하였다.

구체적으로, 데이터 통합 관련 연구는 주로 레코드 연계(record linkage; *e.g.*, Christen, 2012; Winkler, 2006), 통계적 매칭(statistical matching; *e.g.*, D'Orazio, 2017; Rässler, 2002), 데이터 융합(data fusion; *e.g.*, Boonstra & del Pino, 2025; Han & Si, 2025)과 같은 개별 방법론의 특성에 중점을 두었다. 이러한 접근은 특정 단계에서 문제를 해결하는 데에는 유용하나, 통합 과정 전반에서 오류가 어떻게 누적·전이되며, 각 단계의 의사결정이 최종 산출물인 통합 데이터(integrated data set)의 품질에 어떠한 영향을 미치는지에 대한 설명에는 한계를 지닌다. 특히 공식 통계의 경우, 산출 결과에 대한 공적 책임성과 재현 가능성이 요구되므로, 데이터 통합은 단순한 기술적 선택이 아니라 관리 가능한 절차적 의사결정 과정으로 설계될 필요가 있다. 본 연구는 이러한 문제의식에 기반하여 데이터 통합을 개별 기법의 집합이 아닌 절차적 관리 프레임워크

6) 고해상도 통계(high-resolution statistical data)는 국제기구의 문헌을 바탕으로 본 연구에서 제시하는 개념이다. 국제기구들은 데이터의 세분성(granularity), 시의성(timeliness), 정밀성(precision) 향상을 새로운 과제로 제시하고 있다. UNECE(2024)는 통계가 더욱 자세하고 빈번하며 신속하게 생산되어야 함을 강조하였다("continued demand for more timely, frequent and granular official statistics." p. 1). 또한, 새로운 자료원과 기술이 전통적인 통계 생산 체계를 넘어서는 정밀한 데이터 구축의 가능성을 제시하였다("The new sources…, but they may allow for obtaining new data and achieving timeliness, frequency and granularity that were not possible before." p. 12). 이러한 논의는 통계의 공간 단위 세분화(spatial disaggregation)로 확장된다. UN-GGIM Europe(2022)는 통계와 지리정보의 통합을 단위 수준에서 구현해야 한다고 명시하였다("the Global Statistical Geospatial Framework (GSGF) principles encourage the link between statistical data and geography to be made at the unit record level by storing geometry against the unit record, typically on a point-based framework,…." p. 12). Tiecke 등(2017)은 "The resulting high-resolution population datasets have applications in infrastructure planning, vaccination campaign planning, disaster response efforts and risk analysis such as high accuracy flood risk analysis. (p. 1)"이라 언급하며, 고해상도 데이터가 인프라 계획, 보건 대응, 재난 관리 등 미시 단위 의사결정에 직접 활용될 수 있음을 실증적으로 보여주었다. 또한 Zheng 등(2025)은 "Accurate, up-to-date, and highly resolved measurements of economic well-being are essential for monitoring and achieving international goals of poverty alleviation. Granular estimates of household poverty and wealth are critical for understanding whether these goals are being met, as well as for targeting and evaluating anti-poverty interventions in regions where progress is lagging. (p. 2)"이라 밝히며, 지역 수준의 경제·복지 변동을 포착하기 위해 정밀한 시공간 측정(highly resolved measurement)의 필요성을 강조하였다. 종합하면, 고해상도 통계의 핵심은 단순한 데이터 정밀도 향상을 넘어, ①시간·공간적 세분화, ②미시 단위 기반, ③국지적 변동 탐지, ④정책 활용 가능성 확보에 있다.

(procedural management framework)로 정의하고자 한다. 이를 위해 본 연구는 다음의 연구 질문에 답하고자 한다.

첫째, 공식 통계 생산에서 데이터 통합의 품질을 전 주기에 걸쳐 관리하는 데 필요한 최소한의 절차적 단계는 무엇인가?

둘째, 통합 목적, 자료 구조, 통합 방법의 차이는 데이터 통합 절차의 구성과 품질 관리 강도를 어떻게 달리 설계하도록 요구하는가?

본 연구는 관련 연구를 바탕으로 데이터 통합의 개념을 다시금 정립하고, 통합 목적·자료 구조·통합 방법에 따른 유형을 체계적으로 분류한 후, 이를 공식 통계 생산 과정에 적용할 수 있는 절차적 프레임워크로 제시한다. 즉, 본 연구는 데이터 통합을 사후 점검의 대상이 아닌 단계별로 품질관리 기준과 환류 구조가 내재한 관리 과정으로 구조화하였다.

2. 데이터 통합 개념 정립

2.1 데이터 통합이란?

국제기구(UNECE, 2020; UN ESCAP, 2020; UNSD, n.d.)와 주요 국가 통계기관(Australian Bureau of Statistics, n.d.-a, n.d.-b; Frenette et al., 2025; Statistics New Zealand, n.d.-a, n.d.-b)에서 제시하는 데이터 통합의 정의는 대체로 서로 다른 출처의 자료를 결합하여 분석이 가능한 형태로 만드는 과정으로 요약될 수 있다. 이러한 정의들은 다출처 자료 활용의 중요성과 통합의 필요성을 강조하는 데에는 기여하였으나, 데이터 통합을 개별 단계의 병렬적 결합 또는 기술적 처리 과정으로 인식하는 경향이 강하다. 이러한 관점에서는 데이터 통합이 전처리, 결합, 조정, 평가로 이어지는 순차적 의사결정 과정으로 명확히 구조화하지 못한다는 한계를 지닌다. 그 결과, 데이터 통합 과정에서 발생하는 오류와 불확실성이 어느 단계에서 생성되어 이후 단계로 어떻게 전이되는지에 대한 체계적인 분석이 미흡하였다.

국제기구 지침(*e.g.*, UNECE, 2020; UN ESCAP, 2020) 역시 데이터 통합의 필요성과 전반적인 절차를 제시한다는 점에서 의의가 있으나, 통합 전 과정을 단계 간 상호의존성과 오류 전이 구조를 포함한 관리 체계로 재구성하는 데까지는 이르지 못하였다. 특히 각 단계의 품질 위험이 어떻게 관리되는지, 그리고 특정 단계의 판단 결과가 후속 단계의 설계 범위를 어떻게 제약하는지에 대해서는 제한적으로만 다루어지고 있다. 이에 본 연구는 데이터 통합을 개별 기법의 선택 문제가 아니라, 절차 전반에서 형성·관리되는 과정으로 재정의한다. 즉, 통합 데이터의 품질은 개별 단계의 산출물이 아니라, 단계 간 상호작용과 오류 전이를 통해 누적하여 형성되는 결과로 볼 수 있다. 이는 기존 국제기구 지침에서 제시된 데이터 통합의 원칙과 절차를 단순한 참고 틀을 넘어, 절차적 관리 단위로 재구성할 필요성을 시사한다.

본 연구는 해당 국제 지침을 대체하는 것이 아니라, 그 구조를 절차적 관리 체계로 확장하여 데이터 통합 과정에서 요구되는 판단 구조와 품질 통제 메커니즘을 구체화하고자 한다. 특히, 단계 간 상호 의존성과 오류 전이 구조를 명시적으로 반영함으로써, 데이터 통합 과정의 관리 가능성을 절차 수준에서 구현하고자 한다. 이러한 관점에서 본 연구는 데이터 통합을 ①통합 목적, ②자료 구조, ③통합 방법의 핵심 요소로 재구성한다. 세 요소는 기존의 데이터 통합(UNECE, 2019a, 2020; UN ESCAP, 2020) 및 공식 통계 생산(Eurostat, 2017, 2019; OECD, 2011; UNECE, 2011, 2019b) 관련 논의에서 강조된 개념적 구성 요소에 기반하며, 이를 절차 설계와 품질관리 강도를 결정하는 핵심 설계 변수(design parameters)로 재정의한다. 이들 요소는 데이터 통합 과정에서 요구되는 품질관리의 성격을 구조적으로 규정하며, 각각 다른 품질 차원으로 연결된다.

첫째, 통합 목적은 통합 데이터의 활용 수준과 요구되는 품질 수준을 규정하는 요소로서, 결측, 편의, 대표성 문제를 통제하는 결과 품질과 연결된다.

둘째, 자료 구조는 다출처 자료의 이질성이 데이터 통합의 제약 조건으로 작용하지 않도록 관리하는 요소로서, 형식적·개념적 일관성(structural and conceptual coherence)을 확보하는 문제와 연결된다.

셋째, 통합 방법은 결합 과정에서 자료 간 관계를 설정하는 방식과 관련되며, 이 과정에서 발생할 수 있는 구조적·논리적 오류를 통제하는 논리적 정합성(logical consistency)과 연결된다.

2.2 공식 통계 관점에서의 데이터 통합

공식 통계에서 데이터 통합은 단순한 기술적 결합 행위로 접근하기보다는, 관리 체계로 이해될 필요가 있다. 통합 데이터는 통계 산출의 기초 자료(*e.g.*, Asian Development Bank, 2018)로 활용되며, 그 결과는 정책 결정과 공공 의사결정에 영향을 미친다(*e.g.*, Australian Bureau of Statistics, 2020). 따라서 데이터 통합은 개별 기법의 적용이나 일회적 처리 과정이 아니라, 사전에 설계되고 통제 가능한 관리 대상으로 인식할 필요가 있다.

이와 같은 관점에서 데이터 통합은 형식적·개념적 일관성, 논리적 정합성, 결측·편의·대표성 관리라는 요건이 충족될 때 공식 통계 생산에 적합한 상태로 기능한다. 이는 데이터 통합의 성과를 특정 단계의 성공 여부로 판단할 수 없으며, 통합 과정 전반에서 이루어지는 의사결정이 통계 품질에 누적해서 영향을 미친다는 점을 전제로 한다.

본 연구는 공식 통계 생산을 위한 데이터 통합을 다음과 같이 정의한다.

첫째, 서로 다른 출처에서 수집한 자료를 개념적·구조적으로 표준화하는 과정이다.

둘째, 통합 목적에 부합하도록 개별 데이터를 유기적으로 결합하는 기술적 행위이다.

셋째, 이 모든 과정을 통해 검증된 품질의 통계 산출물을 도출하는 '전(全) 주기적 관리 체계'이다.

이는 데이터 통합을 개별 기법의 선택 문제가 아니라 통계 생산 체계 전반의 설계 문제로 위치 지우며, 이후 제시되는 통합 유형 분류와 절차적 관리 프레임워크 논의의 개념적 출발점을 제공한다. 또한, 통합 목적과 자료 특성에 따라 절차 설계의 차이를 설명하고, 통합 데이터의 품질을 사후 점검이 아닌 단계별 관리 대상으로 전환하여 이론적 기반을 제공한다.

3. 데이터 통합의 유형 분류

앞서 정의한 바와 같이, 공식 통계에서의 데이터 통합은 전(全) 주기에 걸쳐 관리되어야 하는 절차적 과정이다. 이러한 관점에서 데이터 통합은 단일한 방식으로 수행될 수 없으며, 통합의 목적, 활용 자료의 구조, 그리고 적용되는 통합 방법에 따라 관리 요건과 절차 설계가 달라진다. 따라서 데이터 통합을 유형화하는 작업은 단순한 분류를 넘어, 통합 절차의 설계와 단계별 관리 강도를 결정하기 위한 전제 조건으로 이해될 필요가 있다.

본 장에서는 데이터 통합을 ①통합 목적, ②자료 구조, ③통합 방법의 세 가지 기준에 따라 체계적으로 분류한다. 이 유형 분류는 데이터 통합의 개념적 차이를 설명하기 위한 것이 아니라, 이후 제시되는 절차적 관리 프레임워크에서 단계의 수행 여부, 반복 여부, 그리고 단계별 품질관리 강도를 조정하기 위한 설계 변수로 기능한다. 즉, 본 장에서 제시하는 유형은 데이터 통합을 설명하기 위한 범주가 아니라, 통합 절차를 구성하기 위한 설계 논리의 일부이다.

설계 변수는 데이터 통합의 구성 방식과 관리 수준을 규정하는 조건으로 작용한다. 예컨대, 자료 구조의 이질성이 높은 경우 전처리 및 정합성 점검 단계의 반복 수행과 관리 강도가 강화되며, 통합 방법에 따라 결합 이후 단계에서 요구되는 불확실성 관리 방식과 품질 점검 기준이 달라진다. 이러한 유형화에 기초하여, 데이터 통합은 통합 목적과 자료 특성에 따라 조건부로 구성·관리되는 구조로 이해할 수 있다.[7)]

3.1 통합 목적에 따른 데이터 통합 유형

데이터 통합의 절차적 설계는 통합 데이터의 활용 목적에 따라 근본적으로 달라진다. 공식 통계에서 데이터 통합은 통계 산출과 공표에 직접 활용되기도 하고, 통계 생산 기반을 구축하거나 자료의 품질을 진단·보완하기 위한 중간 단계로 수행되기도 한다. 이러한 차이는 통합 과정에서 요구되는 품질 수준과 관리 강도를 결정하므로, 통합

7) 조건부 구조는 데이터 통합 프레임워크의 운영 방식으로 구체화된다. 통합 목적은 단계별 품질관리의 우선순위를 결정하고, 자료 구조는 전처리 및 정합성 점검의 반복 가능성과 관리 강도를 규정하며, 통합 방법은 결합 이후 결측값 대체, 보정 및 대표성 점검, 최종 품질 점검에서 적용되는 판단 기준의 성격을 달리 설정하도록 한다. 따라서 본 장의 유형 분류는 데이터 통합의 개념적 차이를 설명하기 위한 분류 체계에 그치지 않고, 데이터 통합 프레임워크의 관리 강도, 반복 여부, 판단 기준, 환류 방향을 조정하는 절차 설계 논리로 기능한다.

목적에 따른 유형 구분은 데이터 통합 절차 설계의 출발점으로 이해할 수 있다.

본 연구에서는 통합 목적에 따라 데이터 통합을 ①활용 목적 통합과 ②기반 마련 목적 통합으로 구분[8]한다. 활용 목적 통합은 정책 수요와 통계 활용 요구에 대응하여 분석 가능성과 정책 연계성을 충족하는 것을 목적으로 하며, 통합 결과(*i.e.*, integrated data set)는 즉시 통계 산출, 분석, 공표에 활용된다. 이 경우 통합 결과의 정확성, 대표성, 안정성이 핵심 품질 목표로 설정된다. 반면, 기반 마련 목적 통합은 공식 통계 생산을 지속적으로 뒷받침하기 위한 제도적·기술적 기반 구축을 목적으로 하며, 그 결과는 표본 틀 보완, 자료 품질 진단, 향후 통합을 위한 인프라 구축 등 간접적 용도로 활용된다. 이 경우 통합 과정의 포괄성, 지속 가능성, 그리고 메타데이터의 일관성이 중요한 관리 대상이 된다.

목적에 따른 구분은 데이터 통합을 단일한 절차나 동일한 품질 기준으로 설계하는 접근에서 벗어나, 목적에 따른 절차적 구성과 품질 점검 수준을 차별화해야 함을 보여준다. 활용 목적 통합에서는 결합 이후의 결측 관리, 편의 및 대표성 조정, 최종 품질 점검이 강화되어야 하며, 기반 마련 목적 통합에서는 전처리 및 정합성 점검을 통해 자료 간 차이를 체계적으로 파악하고 메타데이터를 구축하는 과정이 상대적으로 중요하다. 따라서 통합 목적에 따른 유형 구분은 통합 목적에 부합하는 절차적 구성과 품질관리 전략을 설계하기 위한 기준으로 기능한다.

3.2 자료 구조에 따른 데이터 통합 유형

데이터 통합에서 통합 대상은 자료 구조에 따라 ①정형(structured), ②반정형(semi-structured), ③비정형(unstructured)으로 구분할 수 있으며, 이러한 구조 차이는 기술적 처리 방식(*e.g.*, 전처리, 결합 등)을 넘어 통합 과정에서 관리해야 할 품질 위험의 성격과 전이 경로에 영향을 미친다. 따라서 자료 구조에 따른 유형 구분은 단순한 설명적 분류가 아니라, 품질 위험 요인을 사전에 식별하고 절차 설계를 조정하기 위한 기준으로 이해될 필요가 있다.

정형 자료는 행·열 구조의 테이블 형태로 변수 정의와 분류 체계가 비교적 명확한 자료로, 형식적·개념적 일관성이 확보되면 결합 이후 단계의 관리 부담이 상대적으로 제한적이다. 반정형 자료는 일정한 구조를 갖추고 있으나 변수 정의와 형식이 표준화되어 있지 않아 정형화를 위한 전처리 절차가 요구되며, 전처리 설계의 적절성은 이후 결합 가능성에 영향을 미친다. 비정형 자료는 텍스트, 이미지, 영상, 음성 등 구조화되지 않은 자료로, 데이터 통합에 활용하기 위해서는 추출·분류·정제 과정을 통해 해석 가능한 변수를 생성한 후, 다른 자료와 결합해야 한다. 이 과정에서는 정보 손실과 해석의 불확실성, 생성 변수의 재현성 문제가 주요 관리 쟁점으로 작용한다.

한편, 결합 대상 자료 간 구조가 이질적일수록 전처리의 복잡성은 증가하며, 통합 과정에서 발생하는 불확실성이 결과에 미치는 영향도 커질 수 있다(Doan et al., 2012; UNECE, 2020). 결합 자료별 구조에 따라 통합은 ①구조화된 자료 간 통합과 ②이질적

8) 기존 문헌에서 확립된 분류라기보다, 데이터 통합을 절차적 관리 관점에서 체계화하기 위해 본 연구에서 도입한 구분이다.

자료 간 통합으로 구분할 수 있다.

구조화된 자료 간 통합은 주로 정형 자료 간 결합을 중심으로 이루어지며, 변수 정의와 코드 체계가 비교적 명확하게 정리되어 있어 정합성(consistency) 확보가 용이하다. 다만, 시점 차이, 단위 불일치, 누락 또는 중복 문제 등은 여전히 주요 관리 쟁점으로, 이는 통합 결과에 영향을 미칠 수 있다(UNECE, 2019a, 2020; UN ESCAP, 2020). 구조화된 자료 간 통합에서는 통합 이후 통합 데이터가 기존 통계 생산 체계와 일관되게 관리되고 있는지가 핵심 검토 대상이 된다.

반면, 이질적 자료 간 통합은 정형 자료와 반정형 또는 비정형 자료 간 결합으로, 자료 생성 목적과 수집 방식의 차이, 불명확한 변수 정의로 인해 개념 정합성(consistency of definition) 확보가 중요한 과제가 된다. 이 경우 전처리의 중요성이 더욱 커지며, 통합 과정에서 발생하는 불확실성이 이후 분석 결과에 미치는 영향도 상대적으로 크다.

3.3 통합 방법에 따른 데이터 통합 유형

데이터 통합 방법은 통합 대상의 특성과 가용 정보 수준에 따라 달라지며, 이는 통합 절차에서 수행되어야 할 기능과 관리 초점을 규정한다. 본 연구는 데이터 통합에서 활용되는 방법을 ①레코드 연계(record linkage), ②통계적 매칭(statistical matching), ③데이터 융합(data fusion)으로 구분하고, 이를 단일 방법 선택의 문제가 아니라 절차 내에서 수행되는 기능적 구성 요소로 해석한다. 즉, 통합 방법에 따른 유형 구분은 “어떤 방법을 선택할 것인가?”가 아니라, “통합 과정에서 어떤 기능을 어떤 단계에서 수행할 것인가?”를 결정하기 위한 기준으로 접근한다.

레코드 연계는 서로 다른 자료에서 동일한 통계 단위[9]를 식별·연결(identifying and linking)하는 방법(Christen, 2012; Harron, 2016; Winkler, 2006)이다. 레코드 연계는 연결 변수(linkage variables; 고유식별자 및 공통 변수)의 가용성이나 자료 품질 제약으로 모든 레코드를 완전하게 결합하지 못하는 경우, 결측과 불균형이 발생할 수 있으며 이는 결합 이후 단계로 전이된다. 따라서 연계 정확도가 가장 중요한 품질 기준이며, 과대 또는 과소 연계는 통합 결과의 신뢰성에 직접적인 영향을 미친다.

통계적 매칭은 연계 공백을 완화하기 위한 보완적 수단으로서 공통 보조 변수(auxiliary variables)로 미관측 변수(unobserved variables)를 추정하여, 정보 범위를 확장하는 방법이다(D’Orazio et al., 2006; D’Orazio, 2017; Rässler, 2002). 이에 통계적 매칭은 정확한 결합보다는 분포 보전과 관계 구조 유지에 초점을 둔다. 따라서 가정 설정의 타당성, 모형 선택의 민감도, 결과 분포의 안정성 등이 주요 품질관리 대상이 되며, 이는 이후 단계에서 재검토되어야 할 요소로 작용한다.

9) 레코드 연계에서는 일반적으로 동일 대상(same entity)을 식별·연결하는 것을 전제로 한다. 그러나 공식 통계 생산 환경에서는 분석과 산출이 통계적으로 정의된 관측 단위(statistical unit)를 기준으로 이루어지므로, 동일 개체와 통계 단위가 일치하는 경우가 대부분이다. 이에 본 연구에서는 통계 생산 맥락을 전제로 하여 ‘동일 개체’ 대신 ‘통계 단위’를 중심 개념으로 사용한다.

데이터 융합은 결합 여부와 무관하게 상충·중복된 정보를 조정하고, 결측을 보완하여 일관된 데이터 세트를 구축하는 절차적 틀이다(Boonstra & del Pino, 2025; Han & Si, 2025). 이는 레코드 연계나 통계적 매칭을 통해 결합한 정보의 불완전성을 보완하고 통합 데이터의 일관성과 활용 가능성을 확보하기 위한 조정 원칙[10]으로 기능한다. 이러한 의미에서 데이터 융합은 대체(imputation)와 보정(calibration)을 포함하여 통합 과정 전반에서 발생하는 불확실성을 조정하는 기능을 수행한다.

통합 방법에 따른 유형 구분은 데이터 통합을 단일 기법의 선택 문제로 환원하지 않고, 결합 이후 단계로 전이되는 불확실성과 품질 위험을 관리하기 위한 절차적 기준을 제공한다. 이러한 기능적 구분은 다음 장에서 제시할 절차적 관리 프레임워크에서 각 단계의 역할과 상호 연계를 이해하는 핵심 전제가 된다.

①통합 목적, ②자료 구조, ③통합 방법에 따른 유형 분류는 데이터 통합 절차의 단계 구성과 품질관리 강도를 결정하는 설계 변수로 작용한다. 즉, 데이터 통합 절차는 모든 자료 환경에 동일하게 적용되는 고정적 순서가 아니라, 설계 변수에 따라 단계별 수행 강도와 판단 기준이 달라지는 조건부 구조이다. 예컨대 활용 목적 통합에서는 결측값 대체, 보정 및 대표성 점검, 최종 품질 점검의 관리 강도가 강화되는 반면, 기반 마련 목적 통합에서는 전처리 및 정합성 점검과 메타데이터 기록이 상대적으로 중요해진다. 또한 이질적 자료 간 통합에서는 전처리 및 정합성 점검 단계의 반복 가능성이 커지며, 통계적 매칭 중심 통합에서는 결합 이후 불확실성 관리와 대표성 점검 기준이 강화된다. 이러한 점에서 통합 목적, 자료 구조, 통합 방법은 데이터 통합 프레임워크의 단계별 관리 강도와 판단 기준을 조정하는 절차적 설계 변수로 기능한다.[11]

4. 데이터 통합 프레임워크

본 장에서는 데이터 통합을 공식 통계 생산 과정에 적용할 수 있는 절차적 관리 프레임워크로 재구성하여 제시한다. 데이터 통합은 목적, 자료 구조, 방법론에 따라 절차의 구성과 관리 강도가 달라지는 조건부 과정이며, 이러한 설계 판단은 이후 절차의 단계 구성과 관리 방식의 기준으로 작용한다.

---

10) 이는 특정 기법을 지칭하기보다는, 데이터 통합 과정에서 발생한 결합 누락, 추정 불확실성, 편의 가능성을 조정하는 절차적 판단의 집합에 가깝다.

11) 절차적 설계 변수는 단계별 품질관리 강도의 차이로 구체화된다. 활용 목적 통합에서는 통합 결과가 통계 산출, 분석, 공표에 직접 활용되므로 결측값 대체, 보정 및 대표성 점검, 최종 품질 점검의 관리 강도가 상대적으로 강화된다. 반면 기반 마련 목적 통합에서는 향후 통합을 위한 자료 기반 구축과 메타데이터 축적이 핵심이므로 전처리 및 정합성 점검의 관리 강도가 강화된다. 자료 구조 측면에서는 구조화된 자료 간 통합의 경우 변수 정의, 코드 체계, 시점, 단위, 분류의 정합성 확인이 핵심 판단 기준이 되지만, 이질적 자료 간 통합에서는 자료 구조화, 개념 조화, 정보 손실 여부, 생성 변수의 재현성 검토가 추가적으로 요구된다. 통합 방법 측면에서는 레코드 연계 중심 통합의 경우 오연계와 미연계, 연결 변수의 품질, 연계 정확도 관리가 강화되며, 통계적 매칭 중심 통합에서는 공동 지지영역 확보, 조건부 분포 보존, 모형 민감도 분석이 중요한 판단 기준으로 작동한다. 데이터 융합 중심 통합에서는 상충 정보 조정, 결측 보완, 분포 정렬, 최종 일관성 평가가 상대적으로 중요해진다.

본 연구가 제안하는 프레임워크는 데이터 통합을 단계 간 상호 연계된 절차적 관리 과정으로 이해한다. 통합 과정에 발생하는 오류와 불확실성은 단계별로 상이한 성격을 가지며, 적절히 관리되지 않으면 이후 단계로 전이되어 최종 산출물의 품질에 영향을 미친다. 따라서 데이터 통합은 각 단계가 독립적으로 수행되는 과정이 아니라, 이전 단계의 판단과 결과를 전제로 다음 단계가 구성되는 절차적 관리 과정으로 이해된다.

앞서 논의한 바와 같이 기존 국제기구의 지침은 표준적인 절차를 제공한다는 점에서 유용한 준거가 되지만, 데이터 통합 과정을 하나의 관리 체계로 구조화하고 단계 간 오류 전이 구조를 체계적으로 반영하는 데에는 한계가 있다. 특히 각 단계에서 발생하는 품질 위험의 성격과 그 판단 결과가 후속 단계에 미치는 영향은 제한적으로 다루어진다. 이에 본 연구는 데이터 통합을 개별 기법의 선택 문제가 아니라, 단계별 판단과 환류 구조가 내재한 절차적 관리 프레임워크로 재구성한다. 또한 품질관리를 ①과정 중심의 단계별 점검과 ②결과 중심의 최종 평가를 결합한 이원적 구조로 제시하며, 각 단계는 다음 단계로의 이행 여부를 판단하는 관리적 관문으로 기능한다.

본 연구에서 제안하는 데이터 통합의 절차적 관리 프레임워크는 기존 데이터 결합 실무나 국제기구의 지침에서 제시한 일반적 절차를 단순히 반복한 것이 아니다. 기존 데이터 결합 실무가 주로 자료 준비, 결합 수행, 결합 결과 확인과 같은 기술적 처리 순서를 중심으로 구성된다면, 본 연구는 각 단계를 품질 위험의 발생과 전이를 통제하는 관리적 관문으로 설정한다. 즉 본 연구는 새로운 결합 기법을 제안하는 데 있는 것이 아니라, 기존 실무와 국제기구 지침에서 분산적으로 다루어진 품질관리 요소를 공식 통계 생산에 적합한 단계별 판단 구조와 환류 구조로 재구성한 데 있다.[12)]

4.1 단계 구분의 논리와 프레임워크 개요

데이터 통합 과정에서 발생하는 오류와 불확실성은 단일한 성격을 가지지 않으며, 각 단계에서 상이한 형태로 생성·누적된다. 이러한 이질성은 데이터 통합을 단순한 선형 처리 과정이 아니라 단계별로 구분된 관리 구조로 설계해야 할 이론적 근거를 제공한다. 예를 들어, 자료 간 개념 불일치나 구조적 차이로 인한 오류는 이후 단계에서의 통계적 보정(correction)만으로는 완화하기 어렵지만, 결합 과정에서 발생하는 미연계나 불완전한 정보는 대체(imputation)나 보정(calibration)을 통해 일정 부분 통제가 가능하다. 이러한 차이는 단계별로 요구되는 관리 방식과 품질 통제 전략이 다름을 의미하며, 데이터 통합 절차를 단계별로 구분하여 설계할 필요성을 뒷받침한다.

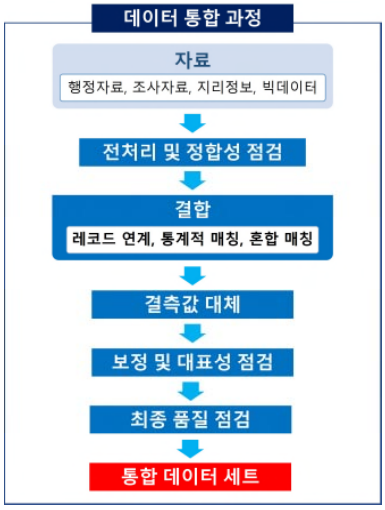
본 연구는 데이터 통합을 「전처리 및 정합성 점검 – 결합 – 결측값 대체 – 보정 및 대표성 점검 – 최종 품질 점검」의 단계로 구조화한다. 이러한 구분은 각 단계에서 요구되는 설계 판단의 성격과 관리 대상이 되는 오류 유형을 기준으로 구성된 논리적 구조이다. 즉, 데이터 통합의 다섯 단계는 단순한 작업 순서라기보다, 각 단계에서 확인된 품질 위험이 후속 단계의 수행 조건과 관리 강도를 규정하는 판단 구조로 이해되어야 한다.

12) 본 연구의 데이터 통합 5단계 절차는 기존 데이터 결합 실무나 국제기구 지침에서 제시한 일반 절차를 부정하거나 대체하려는 것이 아니다. 기존 접근이 자료 준비, 결합, 품질 확인 등 데이터 통합에 필요한 주요 절차와 원칙을 제시한다면, 본 연구는 이를 공식 통계 생산의 관점에서 재배열하여 각 단계의 품질 위험, 후속 단계로의 전이 가능성, 선행 단계로의 환류 필요성을 판단하는 절차적 관리 구조로 재구성하는 데 초점을 둔다.

각 단계는 독립적인 절차가 아니라, 이전 단계의 산출물과 판단 결과를 전제로 다음 단계가 구성되는 상호 의존적 구조를 갖는다. 전처리 및 정합성 점검 단계는 데이터 통합 가능성의 전제 조건을 설정하고, 결합 단계는 데이터 간 관계 구조를 확정하는 분기점으로 기능한다. 이후 결측값 대체와 보정 단계는 결합 결과로 파생된 불확실성과 대표성 왜곡을 통계적으로 조정하는 수렴 과정이며, 최종 품질 점검 단계는 절차 전반의 타당성과 안정성을 종합적으로 평가하는 상위 판단 단계이다.

이와 같은 구조에서 각 단계는 다음 단계로의 이행 여부를 판단하는 관리적 관문(gatekeeper)으로 기능한다. 관리적 관문은 해당 단계에서 설정된 품질 기준의 충족 여부를 평가하여 후속 단계의 진행 가능성을 결정하는 판정 장치이며, 기준이 충족되지 않으면 다음 단계로의 진입은 유보되고 선행 단계로의 환류가 이루어진다. 이러한 단계적 관리 구조는 오류의 전이를 사전에 통제하는 장치이다. 단계별 산출물과 품질 진단 결과는 다음 단계의 설계 조건으로 작용하며, 특정 단계에서 확인된 품질 한계는 이후 단계의 처리 방식과 관리 강도를 제약한다. 동시에 최종 단계에서 발견된 구조적 결함은 필요에 따라 이전 단계로 환류되어 통합 절차의 일부를 재설계하도록 요구한다. 이러한 환류 구조를 통해 데이터 통합은 일회성 결합 행위를 넘어, 단계별 판단과 학습을 통해 지속적으로 품질을 개선하는 절차적 관리 과정으로 이해된다.

<그림 4.1> 데이터 통합 프레임워크



<그림 4.1>은 데이터 통합 프레임워크의 구조로 각 단계는 선형적으로 배열되어 있으나, 이는 단순한 시간 순서를 의미하지 않는다. 각 단계는 관리적 관문을 통해 다음 단계로의 이행 여부를 판단하며, 필요시 선행 단계로 환류되는 구조를 갖는다.

이러한 절차적 구조에 기반하여 데이터 통합 프레임워크에서 각 단계는 관리적 관문을 중심으로 구성되며, 단계별 판단 기준에 따른 절차의 이행과 환류가 결정되는 구조를 갖는다. 관리적 관문은 개념적 수준에 그치지 않고, 각 단계에서 충족되어야 할 판단 기준으로 구체화할 필요가 있다. 이는 단계별 품질 판단을 실제 통계 생산 과정에서 점검이 가능한 형태로 전환하는 것을 의미한다. 더불어, 관리적 관문은 획일적인 임곗값

(threshold)의 적용이 아니라, 사전에 정의된 품질 요건의 충족 여부를 판단하는 구조로 작동한다. 따라서 정량 지표[13]는 단계별 판단을 보조하는 근거로 기능하고, 정성 판단은 해당 지표가 생산 맥락에 비추어 적절하게 해석되었는지를 확인하는 역할을 한다.

<표 4.1> 데이터 통합 단계별 관리적 관문과 주요 판단 기준

단계	관리적 관문	주요 판단 기준 예시
전처리 및 정합성 점검	결합이 가능한 상태인가?	• 변수 대응 관계가 정의되어 있는가? • 개념·시점·단위·분류·범위 불일치가 허용 가능 수준으로 조정되었는가? • 자료 구조에 따라 구조화, 개념 조화, 정보 손실 여부가 점검되었는가? • 통계적 정합성이 허용 범위 내에 있는가? • 정합성 점검 결과가 메타데이터로 기록되었는가?
결합	데이터 구조가 설계 의도에 부합하게 형성되었는가?	• 결합 방법에 따라 오연계·미연계, 공동 지지영역, 조건부 분포 보존 여부가 점검되었는가? • 레코드 연계의 정확도·재현율·오연계율·누락률 등이 수용 가능 수준인가? • 통계적 매칭의 공동 지지영역 확보·조건부 주변 확률분포 보존·민감도 분석 결과가 타당한가?
결측값 대체	결측으로 인한 정보 공백 및 분포 왜곡이 적절히 보완되었는가?	• 결측 메커니즘을 고려한 방법이 적용되었는가? • 결측 발생 구조가 결합 실패, 변수 불일치, 특정 자료의 변수 부재 등과 구분되어 검토되었는가? • 대체 후, 분포 왜곡 및 불확실성 과소평가 문제가 발생하지 않은가? • 대체 방법의 안정성과 타당성을 확보하였는가?
보정 및 대표성 점검	모집단 수준의 분포와 대표성이 확보되었는가?	• 외부 기준자료와의 정합성이 확보되었는가? • 통합 목적에 따라 대표성, 가중값 안정성 여부, 외부 기준자료와의 정합성이 적절히 점검되었는가? • 특정 집단의 과대·과소 대표가 완화되었는가? • 극단 가중값 및 가중값 분포의 변동성이 허용 가능 수준인가?
최종 품질 점검	통합 데이터는 실제 활용 가능한 수준인가?	• 특정 집단 잔존 편의가 허용 범위 내에 있는가? • 설계 변화에 따른 결과 민감도가 과도하지 않은가? • 최종 점검 결과가 어떠한 선행 단계의 재검토를 요구하는지 확인되었는가? • 통합 데이터가 통계 산출 및 공표 가능한 수준의 품질을 충족하는가?

13) ①전처리 및 정합성 점검 단계에서는 개념·시점·단위·분류·범위의 정합 비율과 기술통계·분포 동일성 검정 등의 통계 지표를 활용한다(Harron et al., 2017; Herzog et al., 2007). ②결합 단계에서 레코드 연계는 정확도, 정밀도, 재현율, 오연계율, 누락률 등을 통해(Christen, 2012; Winkler, 2006), 통계적 매칭은 공동 지지영역, 조건부 주변 확률분포 보존, 민감도 분석 등을 활용한다(D'Orazio et al., 2006; Rässler, 2002). ③결측값 대체 단계에서는 대체 이후 분포 왜곡 및 불확실성 과소평가 여부를 평가하는 동시에, 관측값 모의 대체(Abayomi et al., 2008), 교차 검증(Emmanuel et al., 2021; Stekhoven &Bühlmann, 2012) 등을 통해 대체 방법의 타당성과 안정성을 점검한다. ④보정 및 대표성 점검 단계에서는 외부 기준자료와의 정합성, 특정 집단의 과대·과소 대표 완화 정도, 가중값 변동성 등을 검토한다(Deville & Särndal, 1992). ⑤최종 품질 점검 단계에서는 정확성·완전성·대표성·정합성·시의성·접근성을 중심으로 통합 데이터의 품질 수준을 평가하고(Eurostat, 2021; UNECE, 2024; UNSD, 2022), 외부 기준자료와의 차이를 통한 잔존 편의와 설계 선택 변화에 따른 결과 민감도 등을 확인한다.

<표 4.1>의 판단 기준은 모든 데이터 통합 사례에 동일하게 적용되는 고정 기준이 아니라, 통합 목적, 자료 구조, 통합 방법에 따라 관리 강도와 해석 방식이 달라지는 조건부 기준이다. 따라서 관리적 관문은 단순한 통과·실패 판정 장치가 아니라, 다음 단계 이행, 동일 단계 반복, 선행 단계 환류 여부를 판단하는 절차적 장치로 기능한다. 이러한 점에서 관리적 관문은 통합 유형과 품질 위험의 성격에 따라 조정되는 조건부 판단 구조로 이해할 수 있다.

4.2 전처리 및 정합성 점검: 통합 가능성 확보

전처리 및 정합성 점검(preprocessing and consistency assessment)은 데이터 통합의 출발점으로, 다출처 자료 간의 이질성이 이후 단계에서 결합 가능성의 제약으로 작용하지 않도록 관리하는 역할을 한다. 이 단계의 목적은 자료 간 차이를 제거하는 데 있는 것이 아니라, 서로 다른 출처에서 생성된 자료가 동일한 논리적 맥락에서 결합·해석할 수 있도록 최소한의 공통 기반을 형성하는 데 있다. 이러한 관점에서 전처리 및 정합성 점검은 데이터 통합을 가능하게 하는 전제 조건을 관리하는 단계로 이해될 필요가 있다.

전처리는 스키마 정렬(schema alignment)과 개념 조화(harmonization)[14]를 통해 자료 간 형식적·개념적 일관성을 확보하는 과정이며, 정합성 점검[15]은 이러한 조정 결과가 논리적 모순 없이 구현되었는지를 확인하는 절차이다(Rahm & Do, 2000; Rahm & Bernstein, 2001; UNECE, 2019a; UN ESCAP, 2020). 중요한 점은 전처리 및 정합성 점검이 단순한 자료 정제나 표준화 작업으로 환원되어서는 안 된다는 것이다. 이 단계에서 이루어지는 판단은 이후 결합 단계에서 어떤 변수들이 실제로 연계 대상이 될 수 있는지, 어떤 정보가 비교 가능한 상태로 간주될 수 있는지를 결정한다. 따라서, 전처리 및 정합성 점검 단계에서는 그 결과를 명시적으로 기록(*i.e.*, 메타데이터)하고, 이후 단계에서 이를 참조할 수 있도록 관리하는 기능을 포함해야 한다. 이러한 기록은 결합 실패나 품질 저하가 발생했을 때 그 원인을 추적하고 통합 절차를 재설계하는 데 중요한 근거로 활용된다. 이는 데이터 처리 이력(data lineage)을 구성하며, 데이터 통합 과정

14) 스키마 정렬은 서로 다른 자료에서 동일 개념일 가능성이 큰 변수를 식별하고, 대응 관계(mapping relation)를 설정하는 과정으로, 개념 조화 단계에서 변환과 표준화를 수행할 수 있도록 변수명, 자료 유형, 코드 체계, 단위 등의 형식적 일관성(coherence)을 확보하는 데 있다(Rahm & Do, 2000; Rahm & Bernstein, 2001). 개념 조화는 자료 간 변수의 의미와 정의를 공통 기준에 맞추어 변환·표준화함으로써 개념적 일관성을 확보하는 과정이다(UN ESCAP, 2020). 이는 단순한 외형을 일치시키는 것을 넘어, "동일 현상을 동일하게 관측·해석·비교할 수 있도록 만드는 과정"이다(Eurostat, 2014; Rahm & Bernstein, 2001).

15) 정합성은 자료의 논리적 일관성을 전제로 결합 가능성을 판단하는 기준이다(UNECE, 2019a; UN ESCAP, 2020). 스키마 정렬과 개념 조화를 통해 형식적·개념적 일관성을 확보했더라도, 자료가 기준에 부합하지 않는 경우가 있다. 가령, 변수 단위를 모두 연 단위로 변환했음에도 일부가 월 단위로 남아 있을 수 있고, 한국표준산업분류(KSIC)로 변환하는 과정에서 일부 코드가 잘못 연결될 수도 있다. 이러한 오류, 왜곡, 정보 손실을 최소화하여 통합 데이터의 품질을 사전에 확보하기 위해서는 정합성 점검이 필수적이다. 정합성 점검은 전처리 단계와 연계해 반복 수행되며, 그 결과는 메타데이터에 체계적으로 기록한다(UN ESCAP, 2020).

에서의 의사결정 경로를 추적할 수 있도록 함으로써 통합 절차의 투명성과 재현 가능성을 확보하는 기반이 된다.

전처리 및 정합성 점검은 이후 모든 단계의 품질 한계를 규정하는 초기 관리 단계로 기능한다. 이 단계에서는 자료 간 개념 정의, 코드 체계, 시점·단위·분류·분포 일관성 등을 검토하고, 결측·오류·중복 등의 품질 문제를 식별함으로써 자료 간 정합성과 비교 가능성을 판단한다. 이는 단순한 사전 진단에 그치지 않고, 결합 단계에서 허용 가능한 범위와 정확성 수준을 결정하는 기준으로 작용하며, 이후 대체 및 보정 단계에서 요구되는 통계적 가정의 강도를 구조적으로 규정한다.

전처리 및 정합성 점검에서 형식적·개념적·논리적 일관성이 충분히 관리되지 않으면 결합 이후 발생하는 결측이나 편의, 대표성 문제는 단순한 추정의 문제가 아니라, 통합 설계 자체의 한계로 귀결될 수 있다. 따라서, 이 단계에서 도출된 자료별 품질 점검 결과는 과정 중심 품질관리의 1차 산출물로 축적되며, 최종 품질 점검 단계에서 통합 설계의 전제가 실제 통합 데이터에 유지되었는지를 판단하는 기초 자료로 기능한다.

더불어, 전처리 및 정합성 점검 단계의 관리 강도는 자료 구조에 따라 달라진다. 구조화된 자료 간 통합에서는 변수 정의, 코드 체계, 시점, 단위, 분류의 정합성 확인이 핵심 판단 기준이 된다. 반면 반정형 또는 비정형 자료가 포함된 이질적 자료 간 통합에서는 변수 추출, 자료 구조화, 개념 조화, 정보 손실 여부, 생성 변수의 재현성 검토가 추가로 요구된다. 이 경우 정합성 기준이 충족되지 않으면 결합 단계로 이행하기보다 전처리 및 정합성 점검 단계로 환류하여 통합 가능성을 재검토해야 한다.

4.3 결합: 관계 구조의 형성과 통합 설계의 분기점

결합(matching)은 서로 다른 데이터의 통계 단위 또는 정보 구조의 대응 관계를 설정하는 과정으로, 데이터 통합에서 가장 가시적인 단계이자 이후 단계의 불확실성 규모를 결정하는 핵심적인 분기점이다. 전처리 및 정합성 검증이 통합 가능성의 전제 조건을 형성하는 단계라면, 결합 단계는 그 조건에서 어떠한 관계를 허용할 것인가를 선택하는 단계에 해당한다. 이 단계에서의 선택은 이후 대체 및 보정 단계에서 관리해야 할 결측 구조, 불확실성의 형태, 대표성 왜곡 가능성을 구조적으로 규정한다.

본 연구에서는 결합 방법을 레코드 연계(record linkage), 통계적 매칭(statistical matching), 혼합 매칭(hybrid/synthetic matching)으로 구조화한다. 이는 상호 배타적인 결합 기법을 병렬적으로 나열하기 위한 것이 아니라, 데이터의 식별 가능성·불확실성 수준에 따라 결합 전략이 절차적으로 분기한다는 점을 체계화하기 위한 것이다. 한편, 레코드 연계, 통계적 매칭, 혼합 매칭은 결합 단계에서 관계를 설정하는 구체적 적용 방식으로서, 데이터 통합 전(全) 과정에 적용되는 방법론적 유형(*i.e.,* 레코드 연계, 통계적 매칭, 데이터 융합)과 구별된다.[16]

16) 본 연구에서 레코드 연계(record linkage), 통계적 매칭(statistical matching), 데이터 융합(data fusion)은 데이터 통합 방법론의 유형을 구분한 개념으로, 특정 단계가 아닌 데이터 통합 전 과정에 걸쳐 적용되는 절차적 구성 요소를 의미한다. 반면, 결합 단계에서의 레코드 연계, 통계적 매칭, 혼합 매칭은 이러한 방법론이 결합 과정에서 구체적으로 구현되는 방식이다.

레코드 연계는 동일한 통계 단위를 직접 연결함으로써 명확한 ‘1:1’ 또는 ‘1:N’ 관계 구조를 형성할 수 있는 강력한 결합 방법이다(Christen, 2012; Harron, 2016; Winkler, 2006). 연결 변수(*i.e.,* 고유식별자, 공통 변수)의 품질이 충분히 확보된 경우, 결합 결과의 해석 가능성과 정확성은 크게 향상된다. 그러나 연결 변수의 오류·누락·불일치가 존재하면 특정 집단에서 체계적인 미연계가 발생할 수 있으며, 이는 이후 단계에서 결측과 대표성 왜곡의 형태로 전이된다. 따라서 레코드 연계는 정확성을 극대화하는 전략인 동시에 결합 실패가 구조적 결측으로 고착될 위험을 내포한다.

통계적 매칭은 연결 변수가 불완전하거나 부재한 상황에서 공통 보조 변수를 활용하여 정보 범위를 확장하는 방법이다(D’Orazio et al., 2006; Rässler, 2002). 이는 레코드 연계 이후 발생한 정보 공백을 완충하는 기능을 수행함으로써, 대체 단계에서의 불확실성 규모를 조정한다. 즉, 통계적 매칭은 단순한 보완 기법이 아니라, 결합 범위를 확장하는 동시에 불확실성의 성격을 재구성하는 설계 선택에 해당한다.

혼합 매칭은 레코드 연계와 통계적 매칭을 보완적으로 결합하여, 정확성(accuracy)과 정보 범위의 균형을 도모하는 방법이다. 현실의 데이터 환경에서는 고유식별자(unique identifiers)나 공통 변수(common variables)가 충분하지 않은 경우가 많기에 레코드 연계를 통해 식별이 가능한 범위까지 정확성을 확보하고, 미연계 레코드(non-linked records)는 통계적 매칭을 통해 보완함으로써 결합 단계에서 형성된 불확실성 구조를 완화한다. 즉, 혼합 매칭은 정확성 극대화와 정보 확장의 상충 관계를 조정하는 설계 전략으로 이해할 수 있다.

한편, 결합 단계의 주요 품질 위험은 ①과대 결합(false linkage) 또는 과소 결합(missed linkage)에 따른 결합 오류, ②미연계로 인한 정보 공백의 확산으로 구체화할 수 있다. 과대 결합은 서로 다른 개체를 동일한 대상으로 잘못 식별하는 오류이며, 과소 결합은 동일 개체가 존재함에도 불구하고 연계되지 않아 정보가 단절되는 문제를 의미한다. 또한 미연계로 인한 정보 공백은 특정 집단이 체계적으로 결합에서 제외되면 분포 왜곡과 대표성 문제로 이어질 수 있다. 이러한 결합 오류는 이후 결측값 대체 단계에서 추가적인 불확실성을 유발하고, 보정 단계에서는 가중값 조정이나 분포 보정의 강도를 증가시키는 요인으로 작용한다.

결합 단계에서는 결합 규칙 설계, 정확도 검증, 그리고 결합 결과에 대한 품질 평가가 필수적으로 수행되며, 해당 결과는 후속 단계의 처리 방식과 품질관리 강도를 결정하는 기준으로 활용된다. 따라서, 결합 전략은 “데이터 조건에 따라 정확성과 정보 범위 간의 균형을 어떻게 설계할 것인가?”의 문제로 이해할 수 있으며, 이러한 설계 판단은 통합 절차 전반의 불확실성 구조와 품질관리 부담을 구조적으로 규정한다.

결합 단계의 관리적 관문은 통합 방법에 따라 서로 다른 기준으로 작동한다. 레코드 연계 중심 통합에서는 오연계율, 미연계율, 연결 변수의 품질, 연계 정확도가 핵심 판단 기준이 되며, 기준을 충족하지 못하면 결합 규칙 재설계 또는 전처리 단계의 연결 변수 정비로 환류할 수 있다. 통계적 매칭 중심 통합에서는 공동 지지영역 확보, 조건부 분포 보존, 모형 민감도 분석이 핵심 판단 기준이 되며, 기준 미충족 시 공통 보조 변수 또는 모형의 재설정이 요구된다. 따라서 결합 단계의 판단 결과는 단순히 결합 성공 여부를 확인하는 데 그치지 않고, 이후 결측값 대체와 보정 및 대표성 점검 단계의 관리 강도를 결정하는 조건부 판단 근거로 기능한다.

4.4 결측값 대체: 식별 가능성과 추정 구조의 복원

결측값 대체(imputation)는 변수 간의 구조와 관계를 활용하여 관측되지 않은 항목을 추정이 가능한 값(plausible value)으로 치환하는 과정이다(Little & Rubin, 2019). 이러한 접근은 단일 데이터 내 결측뿐 아니라, 데이터 통합 과정에도 적용된다. 결합 단계에서 레코드 연계나 통계적 매칭이 완전하게 이루어지지 못한 경우, 일부 변수는 관측되지 않은 상태로 남게 되며, 이러한 결측은 이후 통계 산출 과정에서 편의(bias)와 분산의 증가를 유발할 수 있다. 따라서 결측값 대체는 단순한 자료 보완 절차가 아니라, 통합 데이터의 완결성과 내적 일관성을 확보하는 결정적 과정으로 이해할 필요가 있다.

데이터 통합에서 발생하는 결측은 전통적인 조사자료의 무응답과는 다른 구조적 특성을 가진다. 결측은 주로 ①결합 실패로 인한 미연계, ②데이터 간 변수 정의 불일치, ③특정 데이터에만 존재하는 변수의 편재성 등으로 발생하며, 이는 통합 대상의 특성과 데이터 구조에 의해 체계적으로 형성된다. 따라서, 결측값 대체 단계의 핵심 과제는 단순히 값을 보완하는 것이 아닌 결측으로 인해 약화된 식별 가능성(identifiability)을 어떻게 복원할 것인가에 있다. 이에 결측값 대체는 결측 메커니즘을 고려하지 않은 일괄적 처리보다는, 결측 발생의 원인과 구조를 반영한 방식으로 설계되어야 한다.

결측값 대체에서의 핵심 품질 위험은 분포 왜곡(distortion of the distribution)과 불확실성의 과소평가(underestimation of uncertainty)이다. 대체 과정이 통합 대상의 이질성을 충분히 반영하지 못할 경우, 특정 특성이 과대 또는 과소 추정될 수 있으며, 이는 통계 결과의 신뢰성에 직접적인 영향을 미친다. 따라서 대체 결과는 이후 단계에서 반드시 점검되어야 하며, 결측값 대체는 데이터 통합의 종결 단계가 아니라 보정 및 최종 품질 점검으로의 이행을 매개하는 전이 단계(transitional stage)로 이해할 필요가 있다. 특히 통계적 매칭 중심의 통합에서는 결측값 대체가 단순한 보완 절차에 그치지 않고, 결합 과정에서 형성된 추정 불확실성과 정보 불일치를 반영하는 관리 단계로 기능한다. 따라서 대체 결과의 분포 변화와 불확실성 반영 여부는 보정 및 최종 품질 점검 단계에서 재검토되어야 한다.

4.5 보정 및 대표성 점검: 분포 및 대표성 구조 재정렬

결측 대체 이후에도 통합 데이터가 모집단을 충분히 대표하지 못하는 경우가 존재한다. 이는 단순한 결합 실패의 문제가 아니라, 자료가 생성·수집되는 방식 자체에서 비롯되는 구조적 대표성 손실(loss of representativeness)에 기인한다. 대표성 손실은 자료의 포괄 메커니즘에 따라 다른 형태로 나타난다. 첫째, 확률 표본에 기반한 조사자료는 추출 확률에 근거한 설계 가중값(design weight)을 적용하더라도 무응답 편향이나 표본 추출 편향이 발생할 수 있다. 둘째, 등록 기반의 행정자료는 표면적으로는 모집단 전체를 포괄하는 것처럼 보이지만, 실제로는 누락(under-coverage)이나 중복(over-coverage)을 내포할 수 있다. 셋째, 플랫폼·로그 데이터와 같은 비확률적·자발적 참여에 기반한 민간 자료는 특정 집단의 과대·과소 대표를 내재할 수 있으며, 이는 통합 데이터에 체계적으로 반영될 위험이 있다. 이러한 대표성 손실은 통합 데이터의 분포 구조를 왜곡시키며, 단순한

결합이나 결측값 대체만으로는 교정되기 어렵다.

이러한 구조적 한계를 보완하기 위해 보정(calibration)이 요구된다. 보정은 외부 기준자료(reference data)를 활용하여 가중값을 조정하거나 분포를 재정렬함으로써 통합 데이터가 모집단 구조와 보다 일관되도록 만드는 단계이다. 이는 단순한 사후 점검이 아니라, 결합 및 대체 단계에서 형성된 불균형을 모집단 수준에서 재조정하는 관리 절차에 해당한다.

그러나 보정은 모든 대표성 문제를 해결하는 만능 장치는 아니다. 외부 기준자료 자체가 시점 차이, 분류 불일치, 등록 누락 등의 제약이 있다면, 보정 결과 역시 구조적 한계를 내포할 수 있다. 또한, 과도한 세분화나 다수의 보정 변수를 동시에 적용할 경우 일부 집단에서 극단적인 가중값이 발생하여 추정 결과의 변동성이 확대될 수 있다. 이는 대표성 확보와 추정 안정성 간의 균형을 요구한다.

따라서 보정 및 대표성 점검 단계는 통합 데이터의 분포 구조가 모집단을 얼마나 충실히 반영하고 있는지를 평가하고, 필요시 이를 재조정하는 절차로 이해하여야 한다. 이 단계에서의 판단은 데이터 통합 설계가 구조적 대표성 한계를 어느 수준까지 통제하였는지를 검증하는 기준으로 기능하며, 최종 품질 점검 단계에서 통합 데이터의 활용 가능성과 통계적 신뢰성을 판정하는 근거로 연결된다.

보정 및 대표성 점검 단계의 관리 강도는 통합 목적에 따라 달라진다. 활용 목적 통합에서는 통계 산출과 공표 가능성이 중요하므로 대표성, 가중값 안정성, 외부 기준자료와의 정합성을 엄격히 점검해야 하며, 기반 마련 목적 통합에서는 향후 통합을 위한 포괄성, 자료 관리 가능성, 메타데이터 축적 가능성이 상대적으로 중요하다.

4.6 최종 품질 점검: 통합 설계의 타당성 및 안정성 평가

최종 품질 점검은 데이터 통합 절차의 마지막 단계로서, 「전처리 및 정합성 점검 – 결합 – 결측값 대체 – 보정 및 대표성 점검」을 거쳐 구축된 통합 데이터가 실제 활용 가능한 수준의 품질을 확보하였는지를 종합적으로 평가하는 절차이다. 본 연구에서 품질관리는 단계별로 수행되는 ①과정 중심 품질 점검과 통합 데이터 자체의 품질을 판단하는 ②결과 중심 품질 점검의 이원적 구조로 설계되었다. 이에 최종 품질 점검은 각 단계에서 수행된 과정 중심 점검을 반복하는 것이 아니라, 그 결과가 누적된 최종 통합 결과물의 신뢰 가능성과 활용 적합성을 판단하는 상위 단계로 기능한다.

최종 품질 점검에서 핵심적으로 고려해야 할 품질 위험은 데이터 통합 절차의 복잡성으로 인해 단계별 오류와 불확실성이 누적되거나 상쇄되어 가시화되지 않을 수 있다는 점이다. 개별 단계에서는 허용 수준으로 판단되었던 오류라 하더라도, 결합·대체·보정의 상호작용을 거치며 통합 데이터의 분포 구조와 추정 안정성을 저해할 수 있다. 따라서 최종 품질 점검은 개별 단계 지표의 재확인이 아니라, 통합 결과를 종합적으로 조망하는 관점에서 수행되어야 한다.

본 연구는 최종 품질 점검을 다음 세 가지 축으로 구성한다.

첫째, 결합 성과와 오류 수준을 결과 차원에서 평가한다. 이는 결합 품질 지표를

활용하되, 단계별 점검의 반복이 아니라 최종 통합 데이터에서 관찰되는 결합 구조가 해석 가능하고 안정적인지, 그리고 미연계로 인한 정보 공백이 결과 분석에 미치는 영향을 중심으로 검토한다.

둘째, 대표성 관점에서 통합 데이터의 분포 구조를 점검한다. 보정을 통해 의도한 모집단 정렬이 통합 결과에 충분히 반영되었는지, 특정 집단의 과대·과소 대표가 남아 있는지 등을 종합적으로 검토함으로써 통합 결과의 모집단 타당성을 평가한다.

셋째, 통합 데이터의 안정성(stability)을 평가한다. 이는 결합 전략, 대체 방법, 보정 설정 등 설계 선택의 변동이 최종 산출 결과에 얼마나 민감하게 반영되는지를 점검하는 것으로, 통합 데이터의 재현 가능성과 신뢰 가능성을 판단하는 핵심 기준이 된다.

최종 품질 점검에서의 환류 방향은 확인된 품질 위험의 성격에 따라 달라진다. 변수 정의, 시점, 단위, 분류 체계의 불일치가 확인되면 전처리 및 정합성 점검 단계로 환류해야 하며, 오연계나 미연계 문제가 확인되면 결합 단계의 규칙과 연결 변수를 재검토해야 한다. 결측 보완 이후 분포 왜곡이나 불확실성 과소평가가 확인되면 결측값 대체 단계로 환류하고, 모집단 분포와의 괴리나 특정 집단의 잔존 편의가 확인되면 보정 및 대표성 점검 단계를 재수행해야 한다. 이러한 환류 구조는 최종 품질 점검을 단순한 사후 평가가 아니라, 품질 위험의 발생 위치를 식별하고 해당 단계의 판단 기준을 재조정하는 상위 관리적 관문으로 기능하게 한다.

5. 결론

본 연구는 공식 통계 생산 환경에서의 데이터 통합을 개별 기법의 적용 문제를 넘어, 단계별 설계 판단과 품질 통제가 구조화된 절차적 관리 프레임워크로 재정립하였다. 기존 연구가 레코드 연계나 통계적 매칭과 같은 특정 결합 방법의 정확성에 초점을 두었다면, 본 연구는 데이터 통합 전(全) 과정에서 오류와 불확실성이 생성·전이되는 구조를 관리 관점에서 체계화하였다는 점에서 차별성을 가진다.

첫째, 데이터 통합의 전 주기에 적용할 수 있는 절차적 단계 구조를 제시하였다. 「전처리 및 정합성 점검 – 결합 – 결측값 대체 – 보정 및 대표성 점검 – 최종 품질 점검」의 다섯 단계는 단순한 시간적 순서가 아니라, 오류 유형과 통제 가능성에 따른 판단 구조로 구성된다. 이를 통해 데이터 통합은 선형적 처리 절차가 아니라, 단계별 의사결정이 이후 단계의 품질 한계를 구조적으로 규정하는 관리 체계로 이해될 수 있다.

둘째, 통합 목적, 자료 구조, 통합 방법을 절차 설계의 핵심 변수로 체계화하였다. 통합 목적은 단계별 품질관리의 우선순위를, 자료 구조는 전처리 및 정합성 점검의 관리 강도와 반복 가능성을, 통합 방법은 결합 이후 결측 구조와 대표성 조정의 필요성을 규정한다. 이는 데이터 통합 유형 분류가 단순한 설명 범주가 아니라,

프레임워크의 관리 강도, 판단 기준, 환류 방향을 조정하는 설계 논리로 기능함을 의미한다.

셋째, 데이터 통합 품질관리를 과정 중심 품질 점검과 결과 중심 품질 점검의 이원 구조로 제시하였다. 과정 중심 점검은 단계별 오류 발생과 전이를 통제하는 역할을 하며, 결과 중심 점검은 통합 설계 전체의 타당성과 안정성을 종합적으로 평가하는 상위 판단 체계로 기능한다. 이를 통해 품질은 사후 검증의 대상이 아니라, 설계와 평가가 결합한 판단 체계로 재정립된다.

본 연구는 데이터 통합을 '기법의 집합'이 아니라 '단계적으로 구조화된 판단 체계'로 개념화함으로써, 공식 통계 생산에서 요구되는 관리 기준을 절차 수준에서 제시하였다. 특히, 본 연구에서 제안한 데이터 통합 프레임워크는 공식 통계 생산에서 통합 과정의 블랙박스화(opacity induced by black-boxing)를 완화하고, 단계별 판단 기준을 통해 오류 전이를 사전에 통제할 수 있는 관리 기반을 제공한다는 점에서 실무적으로도 의의가 있다.

다만 본 연구는 데이터 통합 절차의 개념적·구조적 정립에 초점을 두었으며, 제안한 프레임워크의 실증 적용과 효과 검증은 수행하지 못하였다. 향후 연구에서는 실제 통계 생산 사례를 통해 단계별 설계 선택이 추정 결과의 편의, 분산, 대표성 및 안정성에 미치는 영향을 분석할 필요가 있다. 또한, 단계별 품질 지표의 정교화와 환류 구조의 제도적 구현 방식에 관한 연구가 병행된다면, 데이터 통합 절차는 공식 통계 생산에서 더욱 실질적인 관리 기준으로 발전할 수 있을 것이다.

(2026년 2월 26일 접수, 2026년 4월 8일 수정, 2026년 5월 18일 채택)

참고문헌

이혜정, 이기호, 안수인, 임종호, 이상혁, 조용찬 (2022). 보건복지 분야 데이터 경제 활성화를 위한 다출처 데이터 연계, 통합, 활용 방안 연구. 한국보건사회연구원.

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 57, 273-291.

Asian Development Bank. (2018). User Guide for ADB Statistical Business Register.

Australian Bureau of Statistics. (2020). Labour Account Australia: Methodology.

Australian Bureau of Statistics. (n.d.-a). Multi-Agency Data Integration Project (MADIP).

Australian Bureau of Statistics. (n.d.-b). Data integration.

Boonstra, P.S. and del Pino, P.O. (2025). A comparison of some existing and novel methods for integration historical models to improve estimation of coefficients in logistic regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 188, 46-67.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science and Business Media.

Cibella, N., Fortini, M., Scannapieco, M., Tosco, L., and Tuoto, T. (2009). Theory and practice in developing a record linkage software. In Insights on Data Integration Methodologies. Eurostat.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association, 87*, 376-382.

Doan, A., Halevy, A., and Ives, Z. (2012). *Principles of Data Integration*. Morgan Kaufmann.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Wiley.

D'Orazio, M. (2017). Statistical matching and imputation of survey data. ESSnet Project on Data Integration.

Emmanuel, T., Maupong, T., Mpoeleng, D. et al. (2021). A survey on missing data on machine learning. *Journal of Big Data*, 8, 140.

Emmenegger, J., Münnich, R., and Schaller, J. (2022). Evaluating data fusion methods to improve income modelling. *Research Papers in Economics,* No. 3/22, 1-29.

Eurostat. (2017). European statistics code of practice. Retrieved from https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-18-142

Eurostat. (2019). Quality Assurance Framework of the European Statistical System (Version 2.0).

Eurostat. (2021). Quality assurance framework of the European Statistical System.

Frenette, M., Chan, W., and Handler, T. (2025). Leveraging Statistics Canada data integration opportunities for program evaluation. *Economic and Social Reports*, 5, 1-5.

Han, P. and Si, Y. (2025). Frontiers in data integration. *Journal of the Royal Statistical Society Series A*, 188, 24-26.

Harron, K.L. (2016). An introduction to data linkage. Administrative Data Research Network.

Harron, K.L., Doidge, J.C., Knight, H.E. et al. (2017). A guide to evaluation linkage quality for the analysis of linked data. *International Journal of Epidemiology*, 46, 1699-1710.

Herzog, T.N., Scheuren, F.J., and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer.

Liseo, B. and Tancredi, A. (2009). Model based record linkage: A Bayesian perspective. In Insights on Data Integration Methodologies. Eurostat.

Little, R.J.A. and Rubin, D.B. (2019). *Statistical analysis with missing data (3rd ed.)*. Wiley.

OECD. (2012). Quality Framework and Guidelines for OECD Statistical Activities (Version 2011/1).

Rahm, E. and Do, H.H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 24, 3-13.

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

Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Springer.

Statistics New Zealand. (n.d.-a). Integrated Data Infrastructure (IDI). Retrieved from https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/

Statistics New Zealand. (n.d.-b). Data in the IDI. Retrieved from https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/data-in-the-idi/

Stekhoven, D.J. and Bühlmann, P. (2012). MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28, 112-118.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25, 1-21.

Tiecke, T.G., Liu, X., Zhang, A. et al. (2017). Mapping the world population one building at a time. World Bank.

UNECE. (2011). Using Administrative and Secondary Sources for Official Statistics: A Handbook of principles and Practices. United Nations.
UNECE. (2017). Quality indicators for the Generic Statistical Business Process Model (GSBPM): For statistics derived from surveys and administrative data sources (Version 2.0).
UNECE. (2018). Guidelines on the use of statistical business registers for business demography and entrepreneurship statistics (ECE/CES/STAT/2018/5).
UNECE. (2019a). Guidelines on data integration for measuring SDGs.
UNECE. (2019b). Generic Statistical Business Process Model (Version 5.1).
UNECE. (2020). A guide to data integration for official statistics (Version 2.0).
UNECE. (2024). In-depth review of timeliness, frequency and granularity of official statistics (ECE/CES/2024/6).
UN ESCAP. (2020). Asia-Pacific guidelines to data integration for official statistics.
UN-GGIM Europe. (2022). Core Spatial Data Theme Statistical Units: Recommendation for Content (Version 1.1).
UNSD. (2022). Guidance and Toolkit for Quality Assessment of Administrative Data for Official Statistics.
UNSD. (n.d.). Glossary of statistical terms: Data integration.
Winkler, W.E. (2006). Overview of record linkage and current research directions. U.S. Census Bureau.
Zheng, Z. Wu, T., Lee, R. et al. (2025). Dynamic, high-resolution wealth measurement in data-scarce environments (Policy Research Working Paper No. 11058). World Bank.

A Process-Oriented Management Framework for Data Integration in Official Statistics[17)]

Mingyu Kim[18)] · Seong Ryul Park[19)]

Abstract

The increasing use of administrative, private-sector, and other multi-source data has made data integration an essential component of official statistics production. Previous studies have mainly focused on specific techniques such as record linkage and statistical matching, while providing limited discussion of how errors and uncertainties generated across the integration process should be managed. The present study conceptualizes data integration as a procedural quality management framework embedded in the production system of official statistics. It classifies data integration by integration purpose, data structure, and integration method, and positions these classifications as design variables that shape the operation of the integration process. Based on this perspective, this study proposes a five-stage framework: preprocessing and consistency verification, matching, imputation, statistical adjustment, and final quality assessment. Rather than operating as a fixed linear sequence, these stages function as conditional gatekeepers whose management intensity, decision criteria, and feedback direction vary according to the type of integration and the nature of quality risks. The framework also incorporates feedback mechanisms through which problems identified in later stages can lead to the repetition or redesign of preceding stages. The contribution of this study lies not in proposing a new linkage technique, but in reorganizing dispersed quality management elements from existing practices and international guidelines into a coherent procedural framework for official statistics production. By doing so, it provides a conceptual basis for managing data integration as a structured statistical production process rather than a one-time technical procedure.

Key words : data integration, framework, quality management, official statistics, gatekeeper

17) This study was developed by revising and expanding selected parts of the research conducted by Mingyu Kim and Seong Ryul Park (2025), titled “Establishing a Methodological Framework for Data Integration.”

18) First author and Corresponding author, Researcher, Data and Statistics Methodology Division, Data and Statistics Research Institute, E-mail: mgkim79@korea.kr

19) Second author, Senior Researcher, Data and Statistics Methodology Division, Data and Statistics Research Institute, E-mail: srpark08@korea.kr