사회과학 연구데이터 큐레이션의 문제점과 유형화*

- 리포지토리의 데이터 품질 평가를 중심으로 -

Challenges and Categorizations of Research Data Curation in Social Science: Focusing on the Repository's Data Quality Review

박 석 훈 (Sukhoon Park)** 김 혜 진 (Hyejin Kim)***
신 지 민 (Jimin Shin)**** 허 혜 옥 (Hye-ok Heo)*****
김 석 호 (Seokho Kim)******

목 차

초 록

이 연구는 리포지토리의 데이터 큐레이션 과정 중 데이터 품질 평가 및 선별 단계에 집중하여 기탁된 연구데이터에서 발견되는 문제점들을 분석하고 유형화하였다. 연구데이터의 가치 제고와 재이용 활성화를 위한 데이터 큐레이션의 중요성이 강조되고 있으나, 리포지토리의 큐레이션 작업에 대한 실증적 연구는 부족하다. 이 연구에서는 해외 리포지토리 가이드라인을 검토하여 품질 평가 유형 및 문항을 제시하고, 이를 바탕으로 한국사회과학자료원의 장기 미구축 데이터세트 166건을 분석하였다. 분석 결과, 데이터세트 완결성, 데이터 무결성, 파일 형식, 데이터 문서화, 법적·윤리적 문제 등 다섯 가지 유형 중 데이터세트 완결성과 법적·윤리적 문제가 가장 빈번하게 발생하면서도 리포지토리가 단독으로 해결하기 어려운 문제로 나타났다. 이 연구는 기탁된 데이터를 평가 및 선별할 때 식별되는 문제 유형을 구체적인 기준을 통해 분석하여 리포지토리의 데이터 큐레이션 과정을 이해하는 데 기여한다는 점에서 의의가 있다.

ABSTRACT

This study systematically analyzes and categorizes the challenges found in deposited research data by focusing on the appraisal and selection phases of the data curation process. While the importance of data curation for enhancing the value of research data and promoting data reuse has been emphasized, there has been a lack of empirical research on repositories' data curation practices. This study reviews international repositories' guidelines to identify quality assessment types and criteria and applies them to analyze 166 long-term non-archived datasets from Korea Social Science Data Archive (KOSSDA). The analysis reveals five types of challenges: dataset completeness, data integrity, file format, data documentation, and legal/ethical issues. Dataset completeness and legal/ethical issues are the most frequent and difficult challenges to resolve independently. This study contributes to a better understanding of the repositories' data curation process by analyzing the challenges identified during data appraisal and selection phases through concrete criteria.

키워드: 데이터 큐레이션, 연구데이터 관리, 데이터 품질, 연구데이터 재이용, 데이터 리포지토리
Data Curation, Research Data Management, Data Quality, Research Data Reuse, Data Repositories

---

* 이 논문은 2024년 한국사회학회 후기사회학대회에서 발표한 내용을 전면 수정·발전시킨 것임.
이 연구는 서울대학교 기반연구사업의 지원을 받아 수행되었음(과제번호: 0448B-20240005).
** 서울대학교 사회학과 석사졸업(parksukhoon96@gmail.com / ISNI 0000 0005 2396 8812) (제1저자)
*** 서울대학교 한국사회과학자료원 연구원(kiyo3@snu.ac.kr / ISNI 0000 0005 2398 0336) (공동저자)
**** 서울대학교 한국사회과학자료원 연구원(slgm96@snu.ac.kr / ISNI 0000 0005 2398 7538) (공동저자)
***** 서울대학교 한국사회과학자료원 연구원(hyeokh@snu.ac.kr / ISNI 0000 0005 1762 7704) (공동저자)
****** 서울대학교 한국사회과학자료원 원장: 서울대학교 사회학과 교수
(seokhokim@snu.ac.kr / ISNI 0000 0004 6111 3527) (교신저자)
논문접수일자: 2025년 1월 19일 최초심사일자: 2024년 1월 31일 게재확정일자: 2024년 2월 19일
한국문헌정보학회지, 59(1): 333-354, 2025. http://dx.doi.org/10.4275/KSLIS.2025.59.1.333

1. 서 론

국내에서 오픈사이언스(Open Science) 논의를 비롯하여 연구데이터의 공개 및 재이용과 관련한 연구가 주목받고 있다. 학계와 정부부처를 중심으로 연구데이터의 공개와 이를 위한 관리에 대한 관심이 높은데, 연구데이터를 체계적으로 관리하고 공유하면 연구의 투명성과 신뢰성이 향상되는 동시에, 학문적 협력이 활성화되고 제한된 연구자원을 효율적으로 활용할 수 있기 때문이다(Tenopir et al., 2020). 이러한 연구데이터의 효율적인 활용을 위해서는 적절한 제도와 시스템을 운영해야 한다는 인식이 일반화되고 있다(한상우, 2023). 이러한 인식은 과학기술 분야에서 보다 활발하다고 할 수 있는데, 대표적으로 한국과학기술정보연구원(KISTI)에서는 연구데이터 관리 가이드라인, 데이터 관리 계획(DMP) 가이드라인 등을 수립하고, 국가연구데이터플랫폼인 DataON을 운영하여 연구자들에게 연구데이터의 검색, 분석, 활용을 위한 통합플랫폼을 제공하고 있다(한나은 외, 2024). 인문사회 분야에서도 과학기술 분야와 마찬가지로 오픈사이언스 정책의 필요성과 중요성에 대해 공감하고 있지만, 구체적인 실행 전략까지는 논의되고 있지 않다(정영철 외, 2020).

오픈사이언스에 대한 공감대가 실천으로 이어지기 위해서는 어떻게 해야 하는가? 연구데이터를 비롯한 디지털 정보의 확산에 따라 데이터를 보존하는 것뿐만 아니라, 데이터를 큐레이션(Curation)해야 한다는 주장이 등장하고 있다(Constantopoulos & Dallas, 2007). 과거와는 달리 연구자들이 다루는 데이터의 양과 축적 속도가 비교할 수 없을 정도로 증가하고 있어, 전문 인력에 의한 체계적인 데이터의 관리는 더욱 중요성을 얻고 있다(Thomer et al., 2022). 연구자들 사이에서도 데이터 큐레이션의 가치는 더욱 공감을 얻고 있는데, 한 연구에 따르면 연구자의 97%가 데이터 큐레이션이 데이터 공유에 긍정적인 영향을 미친다고 응답하였다(Marsolek et al., 2023).

연구데이터 관리 및 데이터 큐레이션에 대한 학술적 논의는 문헌정보학을 중심으로 활발히 이루어져 왔으며, 특히 최근 5년간 연구데이터 관련 학술 문헌의 출판이 더욱 증가하고 있다(한상우, 2023). 국외 데이터 큐레이션 연구 동향(최재은, 2024), 국내 과학기술특성화 대학의 연구데이터 관리 서비스(김주섭, 김선태, 2023), 국외 데이터 라이프 사이클 분석(김주섭 외, 2019), 과학기술분야 정부출연연구기관의 연구데이터 관리 방안(한나은 외, 2024) 등 다양한 연구가 수행되었다. 그러나 실제 자료를 기탁받는 리포지토리에서 큐레이션이 어떻게 이루어지고 있는지에 대한 연구는 찾아보기 어렵다.

이 연구에서는 데이터 큐레이션의 여러 단계 중 평가 및 선별(Appraise and Select) 과정에 주목한다. 데이터 큐레이션은 단일의 활동이 아니라 전문적인 인력에 의해 이루어지는 일련의 과정이며, 그중 평가 및 선별은 리포지토리가 기탁받은 데이터를 마주하는 첫 번째 단계이다(Higgins, 2008). 리포지토리는 연구데이터 생산에 직접적으로 관여하기보다 생산된 데이터를 기탁받는 주체이다. 리포지토리는 기탁된 데이터를 평가 및 선별함으로써 큐레이션 작업에 투자해야 할 시간과 노력의 규모를 결정한다. 따라서 평가 및 선별 단계는 이후 수행될 큐레

이션 작업의 범위와 방향을 결정하는 중요한 과정이며, 이를 고찰함으로써 데이터 큐레이션 과정 전반에서 발생할 수 있는 문제점을 파악할 수 있다.

이 연구는 데이터 큐레이션 관련 문헌과 해외 리포지토리의 가이드라인을 검토하여 리포지토리가 기탁된 연구데이터를 어떤 기준과 과정을 통해 평가 및 선별하는지 제시하고자 한다. 특히 "리포지토리에 기탁된 데이터세트에서는 어떠한 문제 유형이 식별되는가?"라는 질문에 답함으로써, 큐레이션과 관련한 학술적 논의가 실증적으로 어떻게 적용되는지, 사회과학 연구데이터의 특성을 고려한 데이터 큐레이션은 어떻게 이루어지는지 구체적으로 보여주고자 한다.

논문의 구성은 다음과 같다. 먼저 연구데이터 재이용을 위한 데이터 큐레이션의 중요성과 그 과정에 관한 학술적 논의를 설명한다. 다음으로 이 연구에서 대상으로 하는 서울대학교 한국사회과학자료원(KOSSDA)의 장기 미구축 데이터에 대한 설명과 분석방법을 제시한다. 이어지는 장에서는 해외 주요 리포지토리의 가이드라인에 따라 도출한 기탁 데이터 품질 평가 유형과 기준을 소개하고, 분석한 결과를 조사자료와 질적자료, 기탁자 유형별로 구분하여 살펴본다. 결론에서는 논문의 주요 내용을 요약하고, 후속 연구를 위한 시사점을 제시한다.

2. 이론적 논의

2.1 데이터 큐레이션: 개념과 필요성

디지털 데이터의 재이용성을 높이는 방법 중 대표적으로 논의되는 개념으로는 디지털 큐레이션(Digital Curation)이 있다. 디지털 큐레이션이란 "현재와 미래 이용을 위해 신뢰할 수 있는 디지털 정보를 유지하고 그 가치를 더하는 활동으로, 달리 말해 디지털 정보의 전체 수명 주기 동안 적극적으로 관리하고 평가(Appraisal)하는 활동(Pennock, 2007, 1)"을 의미한다. 한편, 데이터 큐레이션(Data Curation)은 디지털 큐레이션의 하위 범주로서(김판준, 2015; 신영란, 정연경, 2012), "데이터를 이용하기에 적합하고 보관이 용이하며, 장기적으로 접근할 수 있도록 만드는 과정(Thomer et al., 2022, 2)"으로 정의되며, 현재뿐만 아니라 미래에도 데이터의 활용 가능성을 높이는 일련의 과정을 뜻한다.

데이터 큐레이션은 왜 필요한가? 먼저 데이터 큐레이션은 연구데이터의 품질을 높임으로써 데이터 재이용에 직접적으로 기여한다. 연구데이터 공유에서의 암묵적인 전제 중 하나로 연구데이터를 많이 공유하고 보급할수록 데이터 재이용이 늘어날 것이라는 전제가 있지만, 이들 간의 관계가 단선적인 것만은 아니다(Faniel & Zimmerman, 2011). 연구자들은 연구를 위해 데이터를 선택할 때 데이터 생산자의 평판, 문서화뿐만 아니라 큐레이션 과정의 투명성과 데이터의 품질을 중요시하는데(Yoon, 2017), 큐레이션의 수준은 데이터세트의 재이용을 가장 유의미하게 예측하는 요인이기도 하다(Fear, 2013). 따라서 기탁된 연구데이터의 공개 가능성을 판단하고 필요한 개선사항을 식별하여 해결하는 데이터 큐레이션 작업은 연구데이터의 재이용을 위해 필수적이라고 할 수 있다.

또한 큐레이션된 연구데이터는 연구자의 노력을 줄여주며 연구성과에 기여한다. 연구자의 업무 중 대부분은 데이터를 실제 분석 가능한 형태로 만드는 작업에 할애되는데(Thomer et al., 2022), 이는 수집 단계의 데이터를 곧바로 분석에 적용하기 어렵기 때문이다. 따라서 큐레이션을 통해 데이터의 품질을 높이는 과정은 연구자들의 분석에 이르기까지의 수고를 덜어줄 뿐만 아니라, 신뢰할 만한 연구결과 도출에 기여한다는 점에서 연구데이터 관리에 필수적인 작업이라고 할 수 있다.

2.2 데이터 큐레이션과 FAIR 원칙

좋은 연구데이터 관리는 새로운 지식의 발견과 혁신은 물론 데이터의 출판 이후 지식의 통합과 재이용으로 이어지는 데 반드시 필요한 핵심적인 기제이다(Wilkinson et al., 2016). Wilkinson과 동료 연구자들은 연구자들이 디지털 출판에서 얻을 수 있는 부가가치를 극대화하는 데 도움을 주는 원칙인 FAIR 원칙을 제시하였다. FAIR 원칙이란 검색 가능성(Findable), 접근 가능성(Accessible), 상호운용 가능성(Interoperable), 재이용 가능성(Reusable)의 줄임말로, 국내에서도 연구데이터 관리의 표준으로 자리잡고 있다(국가과학기술연구회, 2019).

하지만 FAIR 원칙을 준수하는 것이 완전한 연구데이터 관리를 의미하는 것은 아니다. 최근 연구에서는 FAIR 원칙의 한계가 지적된 바 있는데, FAIR 원칙은 인간 연구 대상에서 생산된 데이터를 이용할 때 발생하는 인식론적, 법적・윤리적 문제를 직접적으로 다루지 않는다. 또한, 질적자료의 경우 1990년대부터 데이터 큐레이션과 관련한 관행들이 확립되어 왔지만, 소셜 빅데이터 큐레이션에서는 연구 수행, 투명성 제고, 연구참여자 보호 등의 합의가 이루어지지 않아 큐레이션과 관련한 전략 수립이 필요한 상황이다(Mannheimer, 2024). FAIR 원칙의 한계를 보완하고 장기적으로 접근 가능한 디지털 데이터를 제공하기 위해서는 FAIR 원칙을 달성해야 하는 최종 목표보다는 작업상의 기본적인 목표로 삼아야할 것이다. 따라서 리포지토리는 다루는 데이터의 특성, 기술 수준, 법적・윤리적 고려사항 등을 고려하여 FAIR 원칙 그 이상의 데이터 큐레이션을 수행해야 한다.

2.3 데이터 큐레이션에서 평가 및 선별의 중요성

데이터 큐레이션은 어떠한 과정을 통해서 이루어지는가? 데이터 큐레이션의 세부 활동들은 동시에 진행되는 것이 아니라, 일정한 순서와 피드백 과정을 거치면서 완성된다. 대표적인 모델로는 디지털 큐레이션 센터(Digital Curation Centre, DCC)의 DCC 큐레이션 생애주기 모형(DCC Curation Lifecycle Model)이 있다(Higgins, 2008). DCC 큐레이션 생애주기 모형은 크게 8가지 과정 – 개념화(Conceptualise), 생성 및 접수(Create and Receive), 평가 및 선별(Appraise and Select), 투입(Ingest), 보존(Preservation Action), 보관(Store), 접근・이용・재이용(Access, Use, and Reuse), 변환(Transform) –으로 이루어진다. 물론 이 과정은 완전한 것이 아니며, 학문 분야별로 수정 및 개선해서 적용할 필요가 있다(Higgins, 2008).

이 연구에서는 큐레이션의 여러 단계 중, '평가 및 선별' 단계의 중요성에 주목한다. Higgins(2008)의 DCC 모형 이후에 제안된 여러 큐레이션 모형에서도 평가 혹은 선별 단계를 포함하고 있는데(Constantopoulos et al., 2009; Ball, 2012), 이 단계에서 리포지토리는 기탁된 데이터가 이용자들에게 서비스하기에 적합한지 검토하고, 데이터의 특성을 고려하여 구축 단위와 범위, 큐레이션 활동 내용을 결정한다. 이러한 평가 및 선별 단계는 데이터 큐레이션의 전체 과정에서 핵심적인 위치를 차지한다. 이 단계에서의 판단은 이후의 투입, 보존, 보관 등의 단계에 직접적인 영향을 미치며, 각 단계에서 수행될 작업의 범위와 깊이를 결정한다. 예를 들어, 평가 단계에서 식별된 데이터의 문제점은 투입 단계에서의 처리 방식을 결정하고, 데이터의 특성에 따른 보존 전략 수립에도 영향을 미친다. 따라서 이 단계에 대한 체계적인 이해는 전체 큐레이션 프로세스를 설계하고 운영하는 데 필수적이다.

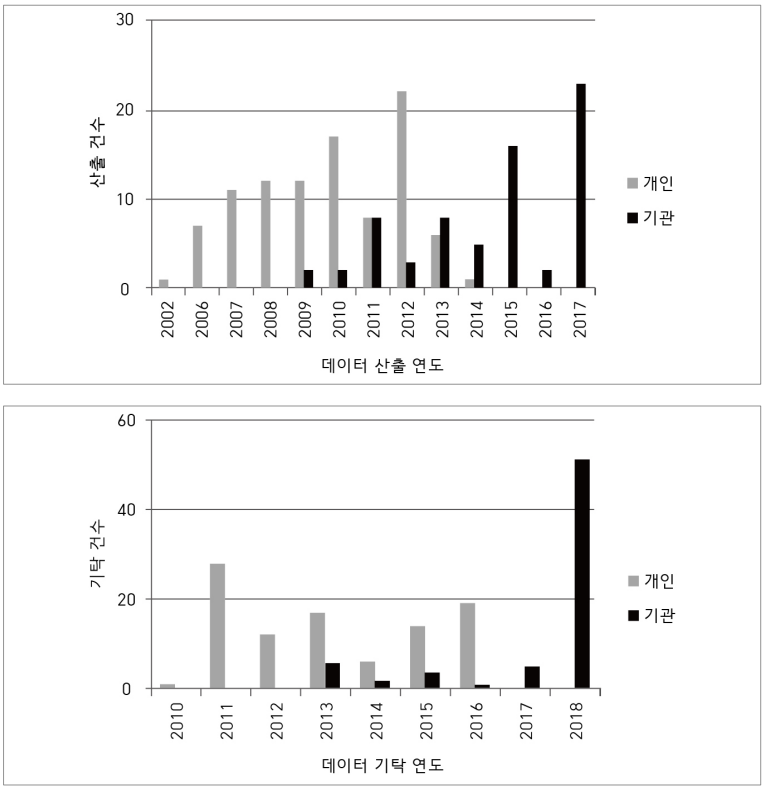
그럼에도 불구하고 기탁된 데이터를 큐레이션하는 리포지토리의 활동은 물론, 데이터 큐레이션의 각 단계에 대한 연구를 찾아보기 어렵다. Plantin(2019)에 따르면, 이는 데이터 큐레이션이 지닌 이중적 특성에 기인한다. 즉, 데이터 큐레이션은 기술적 작업으로서 관리자나 데이터 큐레이터에게는 보이지만, 실제 데이터를 이용하는 연구자 및 이용자들에게는 그 작업이 드러나지 않는 특성을 지니기 때문이다. 이러한 연구 공백을 메우기 위해 이 연구는 실제 리포지토리의 데이터 큐레이션 과정, 특히 평가 및 선별 단계에서 발생하는 문제점들을 체계적으로 분석하고 유형화하고자 한다. 구체적으로 해외 대표적인 사회과학 리포지토리인 미국 대학 간 정치사회연구 컨소시엄(Inter-university Consortium for Political and Social Research, ICPSR), 영국 데이터 서비스(UK Data Service, UKDS), 오스트리아 사회과학 데이터 아카이브(Austrian Social Science Data Archive, AUSSDA)의 기탁 데이터 평가 기준을 검토하여 분석틀을 구성하고, 이를 바탕으로 실제 기탁된 데이터세트의 문제점을 실증적으로 분석한다.

3. 연구자료와 연구방법

이 연구는 한국사회과학자료원에서 수행한 「KOSSDA 장기 미구축 데이터 해소 연구 Ⅰ-Ⅱ」의 구축대상인 데이터세트 166건을 연구대상으로 한다. 한국사회과학자료원은 국내 대표적인 사회과학 데이터 아카이브로서, 개인연구자를 비롯하여 대학 부설 연구소, 정부출연연구기관, 민간연구기관 및 리서치회사로부터 기탁받은 자료를 데이터베이스로 구축하여 국내외 연구자들에게 제공하고 있다. 「KOSSDA 장기 미구축 데이터 해소 연구 Ⅰ-Ⅱ」는 기탁 시점으로부터 2년 이상 경과하였으나 미구축 상태로 남아있는 연구데이터(이하 '장기 미구축 데이터')의 구축 가능성을 평가·선별하여 구축하기 위한 목적으로 2022년부터 2023년까지 수행되었다. 한국사회과학자료원은 기탁된 연구데이터를 가능한 당해연도에 구축하는 것을 원칙으로 하며, 이에 따라 장기 미구축 데이터의 선별 기준을 기탁 시점으로부터 2년 이상 경과한 것으로 정의하였다.

연구대상 데이터세트는 자료유형별로 조사자료 138건(83.1%), 질적자료 28건(16.9%)으로 구성된다. 조사자료는 사회과학 분야의 횡단조사, 패널조사, 종단 및 국제비교 자료 등을 포함하고 있으며, 질적자료는 인터뷰, 기록문서, 관찰기록 등이다. 기탁자 유형으로는 개인기탁 97건(58.4%), 기관기탁 69건(41.6%)으로 구분되며, 질적자료 28건은 모두 개인기탁인 반면, 조사자료는 개인기탁과 기관기탁이 각각 69건(50.0%)으로 동일한 비율을 보인다. 기탁 연도는 2010년부터 2018년까지 분포되어 있으며, 2018년(51건, 30.7%), 2011년(28건, 16.9%), 2013년(23건, 13.9%) 순으로 기탁이 많았다(〈그림 1〉 참조).

이 데이터세트들은 사회과학 분야 연구데이터로서의 특성과 장기 미구축 데이터로서의 특성을 동시에 가진다. 사회과학 연구데이터는 인간을 연구 대상으로 하므로 높은 수준의 연구윤리 준수가 요구된다. 즉, 연구대상자의 연구 참여와 이를 통해 산출된 데이터의 활용과 공유에서의 동의, 연구대상자뿐만 아니라 데이터에 포함된 제3자의 개인정보 보호가 필요하다. 또한, 사회과학 연구데이터는 자료수집 당시의 맥락정보가 데이터의 해석과 활용에 핵심적인 요소가 되므로 자료수집방법과 과정을 상세히 기재한 문서 및 메타데이터가 중요하다.



〈그림 1〉 연도별 데이터 산출 및 데이터 기탁 건수

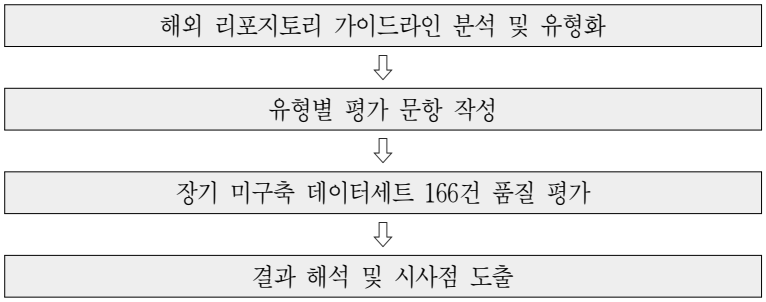
장기 미구축 데이터로서의 특징은 다음과 같다. 먼저 기탁 시점과 구축 시점 간 차이가 크다는 점이다. 데이터 기탁은 데이터가 산출되고 연구자에 의해 1차로 이용·분석된 이후에 이루어지는데, 장기 미구축 데이터의 경우 데이터세트의 기탁 시점과 구축 시점 간의 평균 차이가 7.7년이며, 산출부터 기탁까지의 평균 기간은 3년으로 이 기간까지 고려하면 실제 간격은 더욱 크다. 또한 사회과학 분야 연구윤리심의위원회(Institutional Review Board, IRB) 승인 제도가 도입된 2013년 이전의 데이터가 다수로 연구참여자 보호와 정보에 입각한 동의 등 제도화된 절차에 의해 산출되지 않은 경우가 많다. 연구대상 데이터는 2002년부터 2017년 사이에 산출되었는데, 이 중 2013년까지의 생산 데이터가 119건으로 71.7%를 차지하며, 개인 기탁 데이터의 경우에는 2014년 산출된 1건을 제외한 96건 모두가 2013년까지 생산된 데이터이다(〈그림 1〉 참조).

연구는 네 단계로 진행하였다. 첫 단계에서는 해외 주요 리포지토리의 가이드라인을 분석하여 데이터 품질 평가 유형을 도출하였다. 두 번째 단계에서는 평가 유형별로 조사자료와 질적자료의 특성을 반영한 세부 평가 문항을 마련하여 분석틀을 완성하였다. 세 번째 단계에서는 도출된 평가 문항에 따라 166건의 장기 미구축 데이터세트를 분석하였다. 분석의 신뢰도를 확보하기 위해 두 명의 큐레이터가 독립적으로 평가를 수행한 후 결과를 교차 검토하였다. 평가 기준 충족 여부가 모호한 경우는 큐레이터 간 논의를 통해 최종적으로 판단하였다. 또한, 각 문제 유형의 구체적 사례를 체계적으로 기록하여 향후 유사 사례 평가에 활용할 수 있도록 하였다. 마지막 단계에서는 평가 결과를 자료유형 및 기탁자 유형에 따라 범주화하고 특징을 분석하였다. 특히 발견된 문제들을 리포지토리가 단독으로 해결 가능한 사항과 기탁자와의 협력이 필요한 사항으로 구분하여 분석함으로써, 데이터 큐레이션 전략에 대한 시사점을 도출하고자 하였다. 이상의 연구 수행 절차는 〈그림 2〉와 같다.

4. 분석결과

4.1 데이터 품질 평가 유형

데이터의 품질을 보증하고 가치를 높이기 위



〈그림 2〉 연구 수행 절차

해서는 연구자의 데이터 공유를 고려한 데이터 관리와 리포지토리의 큐레이션이 필요하다. 리포지토리를 비롯한 데이터 관리 및 예산지원 기관은 연구자를 돕기 위한 연구데이터 관리 가이드와 리포지토리의 데이터 평가 기준 및 체크리스트를 마련하여 제공하고 있다. 리포지토리는 이러한 데이터 평가와 검토(Review)를 통해서 일차적으로는 수집 데이터를 선별하고 나아가 데이터 큐레이션 계획을 수립하는 기능을 한다.

연구자를 위한 데이터 관리 가이드와 리포지토리의 연구데이터 평가는 데이터 공유와 재이용을 전제로 하기 때문에 목적 대상은 다르지만 추구하는 데이터 품질상의 목표는 동일하다. 이러한 이유로 연구데이터 관리 가이드는 연구자와 리포지토리 두 주체의 요구를 충족하는 내용을 포함하고 있다.

연구자를 위한 연구데이터 관리 가이드는 연구데이터 생애주기에 따라 FAIR 원칙을 준수하기 위한 고려사항들을 서술하는데 파일 형식, 데이터 무결성, 문서화, 저장 및 백업, 보안, 법적・윤리적 문제 등의 내용을 포함한다. 리포지토리 기관은 데이터 평가 기준으로 데이터의 활용 가능성과 가용성(Availability)을 일차적으로 강조한다. ICPSR과 UKDS는 모두 연구 및 교육에의 활용 가능성과 가치, 현재 혹은 앞으로의 연구와 연구자의 수요 여부, 과학적・역사적 가치와 고유성, 그리고 데이터에 대한 지속적인 접근이 위협받는 데이터를 우선하여 수집한다. ICPSR은 그 외 평가 기준으로 보안・개인정보・기밀성의 고려와 저작권 및 다른 법적 문제, 데이터 품질, 파일 형식 등을 제시한다.

이 연구에서는 사회과학 분야 대표 리포지토리인 ICPSR과 UKDS, 그리고 AUSSDA의 자료를 참고하여 기탁 데이터의 품질 검토 내용을 살펴보고, 이를 정의하여 유형화하고자 하였다. ICPSR과 UKDS의 자료를 참고 대상으로 선정한 이유는 이들 기관이 리포지토리에서 활용할 수 있는 데이터의 평가와 선별 기준을 마련하고 있을 뿐만 아니라 연구자를 대상으로 한 데이터 관리 및 기탁 준비 자료도 풍부하게 제공하여 실용적인 관점에서 참고할 만한 검토 기준을 찾을 수 있었기 때문이다. AUSSDA의 경우, 연구자용 기탁 가이드가 주요한 사항을 간단명료하게 제시하면서도 데이터 클리닝 기술과 체크리스트 등 연구에 참고할 수 있는 도구를 포괄적으로 제공하고 있어 선정하였다. 구체적으로 이 연구에서 참고한 세 기관의 자료는 리포지토리의 데이터 평가 및 선별 관련 문서, 연구자를 위한 연구데이터 관리 가이드, 데이터 기탁 가이드라인 또는 기탁 준비 체크리스트 등이다(Butzlaff, 2022; ICPSR, n.d.a; ICPSR, n.d.b; UKDS, 2022; UKDS, n.d.a; UKDS, n.d.b). 이들 기관의 평가 문항을 분석하여 도출한 데이터 품질 평가의 유형은 〈표 1〉과 같다.

리포지토리는 입수된 연구데이터와 관련 문서의 검토를 통해 데이터세트의 구성 가능 여부와 그 단위를 결정할 수 있다. 이와 관련해서 ICPSR과 UKDS는 데이터와 모든 혹은 필요한 관련 문서가 기탁되었는지를 공통적으로 검토하고 있다. 이 연구에서는 이들 문항을 데이터세트 완결성으로 유형화하였다.

데이터세트의 품질은 데이터 재이용과 직결되는 문제로, 입수한 데이터세트의 품질을 평

〈표 1〉 해외 리포지토리 데이터 평가 문항의 유형별 재분류

기관명	평가 문항	평가 유형
ICPSR	데이터와 모든 관련 문서가 기탁되었는가?	데이터세트 완결성 (Dataset Completeness)
UKDS	필요한 모든 자료(Materials)를 받았는가?	
ICPSR	데이터와 문서 정보가 일치하는가?	데이터 무결성 (Data Integrity)
	데이터에 오류가 없는가? 데이터 수정(Revision)에 대한 권장사항이 있는가?	
UKDS	설문지와 파생 변수 모두에서 변수와 값 레이블이 완전하고 일관성이 있는가?	
	의미 있고 쉽게 설명이 가능한 변수 이름, 코드, 약어가 사용되었는가?	
AUSSDA	데이터와 문서자료를 비교하였을 때, 차이가 없는가?	
	데이터가 일관성이 있고, 품질의 문제가 없는가? (레이블 확인 및 철자 오류)	
ICPSR	데이터를 설명하는 데 필요한 세부 정보가 포함되었는가?	데이터 문서화 (Data Documentation)
UKDS	연구자가 데이터를 재이용하기에 문서화가 충분한가?	
AUSSDA	데이터 재이용을 높이기 위해 문서가 포괄적으로 제출되었는가?	
ICPSR	다양한 컴퓨팅 및 기술 환경에서 이용자가 접근하여 쉽게 사용할 수 있는 형식인가? 연구 가치를 손상시키지 않으면서도 쉽게 접근하고 사용할 수 있는 데이터 형식인가?	파일 형식 (File Format)
	다양한 통계 또는 분석 소프트웨어에서 사용할 수 있는 형식으로 변환 가능한가?	
	부가가치 소프트웨어의 파일 형식인가?	
UKDS	데이터 재이용에 적합한 파일 형식이 생성되었는가?	
	최소한의 비용 혹은 오픈소스로 변환이 가능한가?	
	기탁받은 데이터 형식은 장기 보존이 가능한가?	
AUSSDA	데이터 유형별 형식이 어떠한가?	
ICPSR	저작권자가 확인되었는가?	법적·윤리적 문제 (Legal and Ethical Issues)
	저작권 소유자가 기관의 배포 정책에 동의하였는가?	
	데이터가 개인정보보호 및 비밀유지를 위한 표준을 준수하였는가?	
	데이터가 퍼블릭 도메인 데이터인가?	
UKDS	법적·윤리적으로 데이터 재이용이 허용되었는가? (예: 데이터 공유 및 재이용에 대한 동의, IPR, 저작권, DPA 등)	
AUSSDA	USF 라이선스 계약이 되어 있는가?	
	데이터의 가명 및 익명화 절차가 수행되었는가?	

가하고 큐레이션 가능 여부를 검토하는 것은 필수적인 작업이다. 이와 관련하여 ICPSR과 AUSSDA는 데이터와 문서 정보의 일치 여부를 공통적으로 확인하고 있다. 이외 ICPSR은 이를 데이터 품질(Data Quality)의 개념으로 접근하여 데이터에 오류가 없는지를 점검하고, AUSSDA는 데이터의 일관성과 레이블 확인 및 철자 오류 등의 품질 확인 문항을 포함한다. 그리고 UKDS는 데이터 재이용성(Reusability)라는 개념으로 접근하여 설문지와 파생 변수에서 변수와 값 레이블이 완전하고 일관성이 있는지, 변수 이름과 약어 등이 의미 있고 다른 연구자들에게도 쉽게 설명 가능한지 등을 확인한다. 유럽 사회과학 데이터 아카이브 컨소시엄(Consortium of European Social Science Data Archives, CESSDA)의 데이터 관리 전문가 가이드에 따르면 데이터에 포함된 정보의 정확성, 일관성, 완전성을 보장하는 것을 데이터

무결성(Data Integrity)으로 정의하고 있는데(CESSDA, 2020), 이 연구에서도 해당 정의를 따랐다.

데이터 재이용을 위해서는 데이터 무결성과 함께 데이터에 대한 정보를 충분히 수집하고 이용자들에게 제공하는 작업이 중요하다. ICPSR, UKDS, AUSSDA 세 기관 모두 데이터를 설명하는 정보 제공의 문서화를 점검하며, UKDS와 AUSSDA는 문서화의 목적으로 데이터의 재이용을 강조한다. CESSDA(2020)는 데이터 세트의 재이용을 위한 맥락적 정보를 데이터 문서화(Data Documentation)로 정의하는데, 이 연구에서도 해당 정의를 따랐다.

이용자들이 제한 없이 데이터에 접근하고, 다른 데이터와 상호 호환하는 데 문제가 없으며, 데이터가 장기 보존되기 위해서는 적절한 데이터 형식을 갖추어야 한다. 세 기관에서 제시하는 주요 평가 기준으로는 사용하기에 편리한 데이터 형식인지, 연구 가치를 손상하지 않으면서도 쉽게 접근하고 사용할 수 있는지, 다른 소프트웨어와의 호환 및 변환이 가능한지, 오픈소스 소프트웨어인지 등이 있다. 이와 같은 내용을 확인하는 것을 파일 형식으로 유형화하였다.

데이터 권리자의 확인과 동의는 데이터의 공유와 재이용의 선결 조건이다. 데이터 권리자가 공유와 재이용, 리포지토리의 배포와 저장 및 가공에 대해 동의하지 않는다면 리포지토리는 데이터 큐레이션을 할 수 없다. 따라서 연구자가 데이터 권리자를 명확히 하고 라이선스 계약을 통해 리포지토리의 데이터 가공 및 배포 권한을 승인하며, 그 이용조건을 확인하는 것이 일차적으로 해결되어야 하는 중요한 사안이다. 또한 리포지토리는 데이터에 개인 및 민감 정보가 포함되어 있는지를 검토하고, 해당되는 경우 접근통제, 노출위험평가, 익명화, 보안 등 관련 규정과 장치를 적용할 수 있다. 세 기관은 저작권자의 확인 여부, 데이터 재이용의 허용, 데이터의 가명 및 익명화 절차 수행 여부 등을 평가 문항을 통해 점검하며, 이 연구에서는 이러한 데이터 저작권과 개인정보 보호 문제를 법적・윤리적 문제로 유형화하였다.

4.2 평가 문항 및 결과

앞 절에서 도출한 다섯 가지 평가 유형을 바탕으로, 이를 실제 데이터 평가에 적용하기 위한 구체적인 평가 문항을 마련하였다(〈표 2〉 참조).

한국사회과학자료원의 선별 및 평가 기준은 대부분 해외 리포지토리의 문항을 따랐지만, 국내 사례에 맞게 문항을 변형하거나 추가 문항을 생성하였다. 예를 들어, 데이터세트 완결성에서 해외 리포지토리가 제시하는 '모든 관련 문서', '필요한 모든 자료' 등은 실제 리포지토리의 입장에서 주관의 여지가 개입할 수 있어, 이를 '필수 구성요소'로 구체적으로 정의하여 평가하였다. 그리고 해외 리포지토리는 연구자로부터 기탁된 정보가 충분한지를 확인하는 문항에 한정되지만, 국내는 기탁된 관련 문서만으로는 데이터에 대한 문서화가 충분하지 않아 큐레이터가 메타데이터 및 관련 자료를 작성하는 것이 일반적이어서 이와 관련한 문항을 신규 생성하였다.

질적자료의 경우 한국사회과학자료원의 경험적 사례를 토대로 기탁 데이터의 특성을 고

〈표 2〉 해외 리포지토리와 한국사회과학자료원의 기탁 데이터 평가 문항 비교

평가 유형	평가 문항		
	해외 리포지토리	한국사회과학자료원	비고
데이터세트 완결성	• 데이터와 모든 관련 문서가 기탁되었는가?	• 데이터세트가 필수 구성요소를 갖추었는가?	
	-	• 데이터의 규모와 양을 고려하여 데이터세트로 구성 가능한가?	추가
데이터 무결성	• 데이터와 문서 정보가 일치하는가?	• 데이터와 관련 문헌의 정보가 일치하는가?	
	• 데이터에 오류가 없는가? 데이터 수정에 대한 권장 사항이 있는가?	• 내용이나 파일형태가 손상되었는가? • 분석에 필요한 사회인구학적 배경변수가 누락되었는가? • 데이터 변수와 설문지 문항이 일치하는가?	
	• 설문지와 변수에서 변수와 값 레이블이 완전하고 일관성이 있는가? • 쉽게 설명이 가능한 변수 이름, 코드, 약어가 사용되었는가?	• 변수명, 변수 레이블, 변수값 레이블 정보가 충분한가?	
	-	• 데이터에 동일 템플릿이 적용되었는가?	추가
		• 데이터가 일관된 기준으로 익명 처리되었는가?	추가
파일 형식	• 사용하기에 편리한 데이터 형식이 제공되었는가? • 데이터 재이용에 적합한 파일 형식이 생성되었는가? • 최소한의 비용 혹은 오픈소스로 변환이 가능한가? • 기탁받은 데이터 형식은 장기 보존이 가능한가?	• 데이터가 허용 가능한 파일 확장자인가?	
데이터 문서화	• 데이터를 설명하는 데 필요한 세부 정보가 포함되었는가? • 연구자가 데이터를 재이용하기에 문서화가 충분한가?	• 기탁자가 데이터 설명문서를 충분히 제공하였는가?	
	-	• 큐레이터가 메타데이터 작성에 충분한 설명문서를 얻을 수 있는가?	추가
법적·윤리적 문제	• 저작권자가 확인되었는가? • 저작권 소유자가 기관의 배포 정책에 동의하였는가? • 법적·윤리적으로 데이터 재이용이 허용되었는가?	• 데이터 권리자가 데이터의 공유와 재이용에 동의하였는가?	
	• 데이터가 개인정보보호 및 비밀유지를 위한 표준을 준수하였는가? • 데이터의 가명 및 익명화 절차가 수행되었는가?	• 데이터에 개인 또는 민감 정보의 식별 및 노출 위험이 있는가?	

려하여 평가 문항을 추가하였다. 질적자료는 연구과제 산출 데이터 중 일부만 기탁되거나 방대한 경우 한국사회과학자료원 메타데이터 기준에 맞춰 데이터세트를 구성할 수 있는지 확인이 필요하여 데이터세트 완결성에서 해당 문항을 추가하였다. 또한 기탁 데이터가 익명화를 포함하여 일관된 기준으로 정리되지 않거나 적용 템플릿도 상이한 경우가 많아 데이터 무결성에서 관련 문항을 생성하였다.

이렇게 마련된 데이터 평가 문항을 토대로 한국사회과학자료원의 장기 미구축 데이터세트 166건을 분석하였다. 분석은 자료유형과 기탁자 특성이 드러날 수 있도록 평가 기준을 세분화하였다. 평가 문항은 모든 자료유형에 적용되는 문항(공통), 조사자료에만 적용되는 문항(조사), 질적자료에만 적용되는 문항(질적)으

로 구분하였다. 또한 기탁자의 특성에 따른 차이를 분석하기 위해 기관기탁(기관)과 개인기탁(개인)으로 구분하였다. 연구기관의 규정과 조직 문화 내에서 생성되는 기관자료와 개인연구자가 독립적으로 수행한 연구자료는 차이가 있을 것으로 판단하였기 때문이다. 다만, 질적자료의 경우 모든 자료가 개인기탁 자료이므로 기탁자 유형에 따른 구분은 하지 않았다.

분석 결과는 〈표 3〉과 같다. 평가문항은 '예(Yes, Y)' 혹은 '아니오(No, N)'의 응답이 도출되는데, 문제가 되는 응답값을 집계 기준으로 삼았다. 예를 들어 데이터세트 필수 구성요소 문항은 이를 갖추지 못한 '아니오(집계 기준 N)'의 데이터세트의 빈도를 집계하였고, 반면에 개인 및 민감 정보 식별 위험의 문항은 위험이 있다고 판단되는 '예(집계 기준 Y)'의 응답값의 빈도를 기재하였다. 모든 퍼센트(%)는 집계 가능한 자료를 대상으로 한 유효 퍼센트를 기준으로 산출하였다.

첫 번째 유형인 데이터세트 완결성에서는 데

〈표 3〉 품질 평가 유형별 세부 문항 및 집계 결과

(건수, %)

평가 유형	평가 문항		집계 기준	조사자료			질적자료
				기관 (69건)	개인 (69건)	합계 (138건)	개인 (28건)
데이터세트 완결성	공통	데이터세트가 필수 구성요소를 갖추었는가?	N	25(36.2)	19(27.5)	44(31.9)	5(17.9)
	질적	데이터의 규모와 양을 고려하여 데이터세트로 구성 가능한가?	N	-	-	-	0
데이터 무결성	공통	데이터와 관련 문헌의 정보가 일치하는가?	N	0	15(21.7)	15(10.9)	0
	공통	내용이나 파일형태가 손상되었는가?	Y	0	0	0	9(32.1)
	조사	데이터 변수와 설문지 문항이 일치하는가?	N	6(8.7)	20(29.0)	26(18.8)	-
	조사	분석에 필요한 사회인구학적 배경변수가 누락되었는가?	Y	0	6(8.7)	6(4.3)	-
	조사	변수명, 변수 레이블, 변수값 레이블 정보가 충분한가?	N	0	26(37.7)	26(18.8)	-
	질적	데이터에 동일 템플릿이 적용되었는가?	N	-	-	-	9(39.1)[1]
	질적	데이터가 일관된 기준으로 익명 처리되었는가?	N	-	-	-	27(96.4)
파일 형식	공통	데이터가 허용 가능한 파일 확장자인가?	N	0	0	0	0
데이터 문서화	공통	기탁자가 데이터 설명문서를 충분히 제공하였는가?	N	1(1.4)	18(26.1)	19(13.8)	5(17.9)
	공통	큐레이터가 메타데이터를 작성하기에 충분한 데이터 설명문서를 얻을 수 있는가?	N	0	2(2.9)	2(1.4)	5(17.9)
법적·윤리적 문제	공통	데이터 권리자가 데이터의 공유와 재이용에 동의하였는가?	N	0	26(37.7)	26(18.8)	18(64.2)
	공통	데이터에 개인 또는 민감 정보의 식별 및 노출 위험이 있는가?	Y	6(8.7)	0	6(4.3)	26(92.9)

1) 28건 중 데이터세트 필수 구성요소인 데이터리스트가 없는 5건은 집계에서 제외 후, 유효 퍼센트(%)를 계산하였다.

이터세트의 필수 구성요소 충족 여부를 검토하였다. 필수 구성요소 충족 여부 검토는 기탁자가 제공한 데이터와 관련 파일이 큐레이션에 필요한 구성요소를 제대로 갖추었는지 확인하는 작업이다. 조사자료의 경우 데이터와 설문지를 필수 구성요소로 정의했으며, Excel 형식의 데이터 파일이 기탁된 경우에는 코드북과 코딩가이드 등 변수에 대한 설명문서도 필수 구성요소에 포함하였다. 질적자료에서는 데이터 파일과 함께 데이터리스트가 제출되었는지를 점검하였다.

검토 결과, 전체 138건의 조사자료 중 44건(31.9%), 28건의 질적자료 중 5건(17.9%)이 데이터세트의 필수 구성요소를 갖추지 않은 것으로 나타났다. 조사자료 44건 중 42건은 Excel 프로그램의 확장자(.xlsx, .xls)를 가진 데이터 파일을 기탁받았으나 코드북과 코딩가이드 등 변수에 대한 설명문서가 누락되었다. Excel 파일은 SPSS 프로그램의 데이터 파일(.sav)과 달리 변수명과 변수값만 기재할 수 있어, 변수 레이블, 변수값 레이블 등의 정보를 포함할 수 없다. 이는 SPSS 파일에서 큐레이터나 일반 이용자가 변수 설명문서가 없더라도 데이터 내 변수를 쉽게 식별할 수 있는 것과 대조된다. 따라서 큐레이터가 변수를 식별하려면 코드북이나 코딩가이드가 반드시 필요하다. 나머지 2건은 기탁된 데이터의 파일명이 내용과 일치하지 않았는데, 파일명은 설문지를 의미하지만 실제 파일의 내용은 설문지가 아닌 코드북이나 빈도표 등인 경우가 해당된다. 질적자료의 경우 메타데이터 단위의 데이터세트 구성은 28건 모두 가능했으나, 5건은 데이터 정보 확인에 필요한 데이터리스트가 누락되어 큐레이션이 불가능한 경우로 판단하였다.

두 번째 유형인 데이터 무결성에서는 기탁된 데이터의 상태를 확인하였다. 공통 기준으로는 데이터와 관련 문헌 간 정보 일치 여부, 데이터 파일의 내용이나 형태의 손상 여부를 기준으로 검토하였다. 데이터와 관련 문헌 간 정보가 일치하지 않는 경우는 조사자료 138건 중 15건(10.9%)이었으며, 이는 연구결과보고서의 표본수와 실제 데이터의 사례 수가 불일치한 경우였다. 이러한 불일치는 기탁된 데이터 원본이 최종 버전의 데이터가 아닐 수 있음을 시사한다. 내용이나 파일형태 손상은 질적자료 9건(32.1%)에서 발견되었는데, 모두 관찰기록 이미지의 낮은 해상도의 문제로 나타났다.

자료유형별 평가 기준을 살펴보면, 먼저 조사자료에서는 세 가지 기준을 적용하였다. 첫째, 데이터 변수와 설문지 문항 간 일치 여부로, 138건 중 26건(18.8%)에서 문제가 발견되었다. 대부분 설문지 문항에 있는 측정개념의 변수가 데이터에 없거나, 반대로 데이터의 변수가 설문지에 없는 경우였다. 또한 변수 정보 부족으로 설문지 문항과 데이터 변수 간의 대응 관계를 확인할 수 없는 기탁자료도 존재하였다. 예를 들어 연구결과보고서의 빈도표나 별도의 빈도표 파일이 있다면 큐레이터가 실제 데이터와 대조하며 변수와 설문 문항을 연결할 수 있지만, 이러한 파일이 없을 경우 데이터 구축이 불가능하다. 둘째, 사회과학 연구데이터에서 빈번히 사용되는 사회인구학적 배경변수의 포함 여부를 기준으로 평가한 결과, 전체 조사자료 138건 중 6건(4.3%)에서 성별, 연령 등 사회인구학적 배경변수가 누락된 것으로 확인되었다. 마지막으로 변수, 변수 레이블, 변수값 레이블이 충분한

지를 확인하였으며, 138건 중 26건(18.8%)에서 불충분한 것으로 나타났다. 대표적으로 SPSS 파일의 변수 정보란이 비어있는 채 기탁되었거나, Excel 파일을 SPSS 파일로 변환 후 기탁된 설문지와 대조하면서 큐레이터가 직접 변수 정보를 입력해야 하는 경우가 해당된다.

질적자료는 데이터 템플릿 적용과 익명 처리의 일관성을 추가로 검토하였다. 질적자료는 대개 텍스트 형태로, 동일한 템플릿으로 정리되었을 때 이용자가 주요 정보를 쉽게 파악할 수 있다. 예를 들어 인터뷰 데이터는 면접 정보와 면접대상자 정보를, 기록문서와 관찰기록은 파일명, 생산자, 산출시기 등의 정보가 동일한 형식으로 정리되어야 한다. 동일한 형식으로 정리되었을 때, 큐레이터의 데이터 가공 시간이 단축될 수 있으며, 이용자가 보다 용이하게 재이용할 수 있다. 검토 결과 질적자료 23건 중 9건(39.1%)이 일관된 템플릿을 적용하지 않았다. 리포지토리는 개인정보와 민감정보가 포함된 데이터에 대해 보안 강화, 익명화, 접근통제 등의 큐레이션 작업을 일관된 기준으로 수행해야 한다. 익명 처리의 일관성을 검토한 결과, 전체 질적자료 28건 중 27건(96.4%)이 직접 식별자를 익명화하지 않은 상태로 기탁되었다.

세 번째 유형인 파일 형식에서는 기탁된 데이터의 파일 확장자가 권장 또는 허용 가능한 형식인지 검토하였다. 리포지토리는 데이터의 접근성과 재이용을 위해서 기탁된 연구데이터를 지속 가능한 디지털 파일 형식으로 이용자들에게 제공해야 한다. 사회과학 분야 주요 리포지토리들은 표준 및 개방형 파일 형식을 원칙으로 두고, 권장 및 허용 가능 파일 형식에 대한 기준을 마련하여 제공하고 있다(Butzlaff, 2022; CESSDA, 2020; Corti et al., 2020; ICPSR 2020; Pienta & Reneau, 2023; UKDS, n.d.c). 대표적으로 이들 리포지토리는 마이크로소프트 오피스(Microsoft Office), SPSS 등 소프트웨어의 표준 파일 형식과 개방형 파일 형식의 PDF/A, CSV, TIFF 등을 권장한다. 〈표 4〉는 해외 주요 리포지토리의 가이드라인을 참고하고, 국내 사회과학 분야 기탁 데이터의 유형과 디지털 파일 형식을 고려하여 한국사회과학자료원에서 마련한 기준이다(한국사회과학자료원 아카이빙사업부, 2024).

해외 기관과 달리 한국사회과학자료원이 원자료 텍스트와 관련 문서 파일 형식에서 hwp, hwpx, doc, docx 등을 권장하는 것은 데이터의 품질 향상을 위한 큐레이터의 가공 작업이 대개의 경우 수반되기 때문이다. 한국사회과학자료원은 원자료 텍스트 유형의 대부분을

〈표 4〉 데이터 유형별 권장 및 허용 가능 파일 확장자

구분	데이터 유형	권장하는 파일 확장자	허용 가능한 파일 확장자
원자료	설문조사 데이터	sav, dta, csv	xlsx
	텍스트	hwp, hwpx, doc, docx	pdf, txt
	이미지	tif	jpeg, jpg, png, pdf
	오디오	flac	mp3, wma, wav
	비디오	mp4	wmv, mpge4, avi
관련 문서	문서, 스크립트 등	hwp, hwpx, doc, docx, xls, xlsx	pdf, txt

차지하는 전사자료의 경우 익명화를 포함한 데이터 가공을, 설문지・질문지・데이터리스트 등 관련 문서는 표준 포맷 적용, 데이터 인용 주요 정보 제공, 오탈자 수정 등의 작업을 수행하고 있어 편집가능한 파일 형식을 권장하고 있다.

이를 기준으로 기탁 데이터의 파일 형식을 검토한 결과, 조사자료와 질적자료 모두 권장 또는 허용 가능한 파일 확장자를 사용하는 것으로 나타났다. 다만 질적자료의 이미지 파일(18건)은 허용 가능한 파일 확장자(.jpg)로 현재 이용에는 문제가 없으나, 권장 파일 포맷(.tif)을 따르지 않아 향후 데이터의 장기 보존 측면에서 한계가 있다. 특히 이미지 및 오디오와 비디오 데이터는 최초에 높은 사양의 파일 형식으로 산출되지 않은 경우 데이터 큐레이터에 의한 업그레이드가 불가능하고 편집과 저장을 반복하면 손실의 위험이 높아 장기적인 접근과 이용에 제한이 있기 때문에, 파일 확장자 준수가 필요하다.

네 번째 유형인 데이터 문서화에서는 설명문서의 충분성과 메타데이터 작성 가능성을 평가하였다. 설명문서의 충분성에서는 기탁자가 데이터를 기탁할 때 연구계획서, 연구결과보고서 등 데이터를 설명하는 문서를 함께 제공했는지 검토하였다. 검토 결과 조사자료 138건 중 19건(13.8%)과 질적자료 28건 중 5건(17.9%)이 데이터 설명문서 없이 기탁된 것으로 나타났다. 특히 19건의 조사자료 중 18건이 개인기탁 자료로 나타났는데, 이러한 차이는 기관과 개인의 연구 수행 방식에서 기인하는 것으로 보인다. 연구기관은 조사와 연구 수행 시 표준화된 형식의 연구결과보고서를 발간해야 하므로, 자료를 기탁할 때 보고서를 포함한 조사에 대한 정보를 상세하게 제공할 수 있다. 반면 개인연구자는 조사설계 단계에서 데이터 공개를 고려하지 않은 경우 관련 정보의 필요성을 인식하지 못하여 데이터 설명문서 없이 자료를 산출하는 경향 때문인 것으로 판단된다.

메타데이터 작성 가능성 평가는 메타데이터를 국제표준인 DDI(Data Documentation Initiative)를 준수하여 작성할 수 있는지 여부로 판단하였다. DDI는 사회 및 행동 과학 분야 연구데이터를 설명하기 위해 설계된 기술 표준으로 ICPSR, UKDS, AUSSDA 등 주요 사회과학 리포지토리에서 적용하고 있다. 166건의 기탁자료 모두 기탁자에 의한 메타데이터는 작성되지 않았으며, 큐레이터가 연구 관련 문서나 외부 정보를 활용하여 메타데이터를 작성하였다. 그러나 조사자료 138건 중 2건(1.4%)과 질적자료 28건 중 5건(17.9%)은 이러한 추가 작업에도 불구하고 연구과제 및 데이터 관련 정보를 확인할 수 없어 메타데이터 작성이 불가능한 자료로 판단하였다. 이들 자료의 경우 다른 기탁자료와 마찬가지로 메타데이터가 작성되지 않았지만, 데이터 제공을 위한 기본 정보조차 확보할 수 없는 경우라는 점에서 문제 사례로 분류하였다.

마지막 유형인 법적・윤리적 문제에서는 저작권과 개인정보 보호를 검토하였다. 먼저 저작권 기준에서는 데이터 권리자가 데이터 공유와 재이용에 동의했는지를 기준으로 검토하였다. 검토 결과 전체 조사자료 138건 중 26건(18.8%)과 질적자료 28건 중 18건(64.2%)에서 저작권 기준을 충족하지 못하는 것으로 나타났다. 조사자료의 경우 21건은 동의서는 있으나 데이터 저

작권자의 서명이 없었으며, 5건의 경우 동의서가 없이 기탁되었다. 질적자료 18건은 연구책임자로부터 데이터 공유와 재이용에 대한 동의서는 확보되었으나, 관찰기록 9건과 기록문서 9건에 연구참여자가 직접 산출한 데이터가 포함되어 있어 저작권 침해의 소지가 있었다.

개인정보 보호 기준에서는 기탁된 데이터에 개인 또는 민감 정보의 식별 및 노출 위험이 있는지 검토하였다. 검토 결과 조사자료 6건(4.3%)에서 개방형 문항의 응답에 개인정보가 포함되어 있었고, 질적자료는 28건 중 26건(92.9%)에서 개인정보 식별 및 노출 위험이 발견되었다. 특히 질적자료의 경우 1건을 제외한 모든 자료가 익명 처리가 되지 않은 상태로 기탁되었다. 익명처리한 1건의 데이터도 직접 식별자는 제거되어 있었으나, 데이터 내 간접 식별자들을 결합하거나 외부 출처 정보와 연계하면 개인 식별이 가능한 것으로 확인되어, 체계적인 익명화 작업이 필요한 것으로 판단하였다.

4.3 요약 및 시사점

분석결과를 요약하면 다음과 같다. 조사자료는 데이터세트 필수 구성요소 미충족(31.9%), 데이터 변수와 설문지 문항 불일치(18.8%), 변수 정보 불충분(18.8%), 데이터 권리자의 공유・재이용 미동의(18.8%) 등이 주요 문제로 확인되었다. 반면, 질적자료는 익명 처리 미흡(96.4%), 개인정보 식별 및 노출 위험(92.9%), 데이터 권리자의 공유・재이용 미동의(64.2%) 등이 주된 문제로 식별되었다. 이러한 분석결과는 자료유형에 따라 리포지토리가 서로 다른 큐레이션 전략을 수립하고 작업 내용과 범위를 확정해야 함을 보여준다.

식별된 문제들은 리포지토리의 대응 방식에 따라 크게 두 유형으로 구분할 수 있다. 첫째는 리포지토리가 단독으로 해결하기 어려운 문제들로, 데이터세트 완결성과 법적・윤리적 문제가 여기에 해당한다. 리포지토리는 데이터를 직접 생산하지 않고 기탁받은 데이터에 대해서 큐레이션을 수행하기 때문에, 필수 구성요소 누락이나 저작권, 개인정보 보호 문제는 기탁자와의 긴밀한 협력을 통해서만 해결할 수 있다.

둘째는 리포지토리의 전문성을 통해 해결 가능한 문제들로, 데이터 무결성, 문서화, 파일 형식이 해당한다. 예를 들어 변수 정보가 불충분한 경우 큐레이터가 설문지 문항과 대조하여 보완할 수 있고, 익명 처리가 미흡한 경우 익명화 작업에 도움을 줄 수 있다. 또한 데이터 설명문서가 부족하면 외부 출처를 활용해 메타데이터를 작성할 수도 있다. 파일 형식 역시 .xlsx, .csv 등 변수 레이블 정보를 포함하지 않은 채 기탁된 경우에는 큐레이터의 작업을 통해 해결할 수 있다. 다만 이러한 작업에는 상당한 시간과 노력이 요구될 뿐만 아니라 큐레이터에 의한 오류가 발생할 여지가 있다. 따라서 연구자가 연구데이터 생산 및 분석 단계에서 데이터 설명문서 작성, 파일 형식 등을 고려한다면 보다 효율적인 관리가 가능할 것이다.

한편 조사자료의 데이터 품질은 기탁자 유형에 따라 상당한 차이를 보였다. 개인기탁 자료는 기관기탁 자료에 비해 데이터 무결성, 데이터 문서화, 법적・윤리적 문제의 세부 평가 기준을 충족하지 못하는 비율이 높은 것으로 확인되었다. 구체적으로 데이터 무결성 기준을 충족하지 못한 데이터세트는 개인기탁이 50건

(72.5%), 기관기탁이 6건(8.7%)이었다. 데이터 문서화에서는 개인기탁이 18건(26.1%), 기관기탁 1건(1.4%)이 평가 기준을 충족하지 못했으며, 법적·윤리적 문제는 개인기탁 26건(37.7%), 기관기탁 6건(8.7%)이 기준을 충족하지 못했다. 이러한 차이는 개인연구자의 경우 기관에 비해 연구 초기 단계에서 연구데이터 공개를 고려하지 않을 뿐만 아니라 법적·윤리적 문제, 데이터 관리에 대한 지식 부족 등의 이유로 데이터 관리의 필요성을 인식하지 못하는 경우가 많고, 연구데이터 관리 자체가 개인이 수행하기에 현실적으로 어려움이 따르는 작업이기 때문이다(김지현, 2012; 박성준 외, 2023; 한나은 외, 2024).

최종적인 평가 및 선별 결과, 166건 중 세부 평가 기준을 충족했거나 리포지토리가 큐레이션 작업을 통해 구축 가능한 데이터세트는 총 107건(64.4%)이었다. 반면, 해결 가능 여부를 추가적으로 검토해야 하는 구축 보류 상태의 데이터세트는 36건(21.7%), 리포지토리가 단독으로 해결할 수 없는 구축 불가 상태의 데이터세트는 23건(13.9%)으로 나타났다. 이처럼 전체 데이터세트의 35.6%가 큐레이터의 전문적인 작업에도 불구하고 이용자들에게 즉시 제공할 수 없다는 점은 기탁 단계 이전부터 데이터 큐레이션이 중요함을 실증적으로 보여주는 결과이다.

5. 결 론

이 연구는 사회과학 연구데이터의 큐레이션 과정에서 리포지토리가 마주하는 문제점들을 체계적으로 분석하고 유형화하였다. ICPSR, UKDS, AUSSDA의 가이드라인을 검토하여 도출한 다섯 가지 유형－데이터세트 완결성, 데이터 무결성, 파일 형식, 데이터 문서화, 법적·윤리적 문제－을 기준으로 장기 미구축 데이터세트 166건을 분석한 결과, 자료유형에 따라 서로 다른 문제점이 두드러졌다. 조사자료에서는 데이터세트의 필수 구성요소 미충족(31.9%)과 함께 데이터 변수와 설문지 문항 불일치, 변수 정보 불충분, 데이터 권리자의 공유·재이용 미동의(각 18.8%) 등이 주요 문제로 나타났다. 반면, 질적자료에서는 익명 처리 미흡(96.4%)과 개인정보 식별 및 노출 위험(92.9%)이 핵심 과제로 확인되었다. 또한 조사자료의 경우, 개인기탁과 기관기탁이 각각 절반을 차지하는 가운데 기탁자 유형에 따른 차이가 확인되었다. 개인기탁 자료는 기관기탁 자료에 비해 데이터 무결성(72.5%), 데이터 문서화(26.1%), 법적·윤리적 문제(37.7%) 등에서 평가 기준을 충족하지 못하는 비율이 현저히 높았다.

이 연구는 다음과 같은 점에서 의의를 갖는다. 첫째, 해외 리포지토리의 가이드라인을 바탕으로 구체적인 평가 문항을 제시함으로써 국내 사회과학 데이터에 적용할 수 있는 실천적인 평가 기준을 마련하였다. 둘째, 자료유형과 기탁자 특성에 따른 연구데이터 품질 문제를 실증적으로 분석함으로써, 데이터 특성에 적합한 차별화된 큐레이션 전략의 필요성을 뒷받침하였다. 셋째, 평가 및 선별 단계에서의 판단이 이후 큐레이션 작업의 범위와 방향을 결정한다는 점에서 리포지토리의 데이터 큐레이션 과정에 대한 이해를 제고하였다.

그럼에도 불구하고 이 연구는 다음과 같은 한계를 갖는다. 먼저, 연구자료의 기탁 시점과 구축 시점 간의 차이로 인해 장기 미구축 데이터의 특성이 현재의 일반적인 기탁 데이터와 다를 수 있다는 점이다. 최근에는 데이터 관리에 대한 인식이 높아지고 관련 정책이 강화되면서 기탁 데이터의 품질이 전반적으로 향상되었을 가능성이 있다. 또한, 단일 데이터 리포지토리 사례만을 분석했다는 점에서 연구결과를 일반화하는 데는 어려움이 있을 수 있다.

이러한 한계를 고려하여 후속 연구에서는 다음과 같은 과제들이 수행될 필요가 있다. 첫째, 데이터 무결성과 문서화 부족 등의 문제가 실제 큐레이션 과정에서 어떻게 해결되는지, 그 과정에서 리포지토리가 어떤 역할을 수행하는지에 대한 실증적 연구가 필요하다. 둘째, 다양한 리포지토리의 연구데이터 큐레이션 사례를 비교 분석함으로써 연구결과를 일반화하고 기관별 특성과 모범 사례를 도출할 필요가 있다. 마지막으로 본 연구에서 다룬 품질 평가 및 선별 단계를 넘어, 데이터 처리, 보급과 보존 그리고 재이용에 이르는 전체 큐레이션 과정에서 발생하는 문제점과 해결 방안에 대한 실증적 연구로 확장될 수 있다.

데이터 큐레이션은 상당한 비용과 노력이 요구되는 활동이지만, 이는 연구데이터의 가치를 제고하고 연구결과의 재현을 용이하게 하는 필수적인 과정이다(Peer, 2011). 특히 이 연구에서 확인된 것처럼 자료유형과 기탁자 특성에 따라 서로 다른 문제점이 존재하므로, 이러한 특성을 고려한 차별화된 큐레이션 전략이 필요하다. 이 연구의 분석 결과는 데이터 큐레이션이 단순한 기술적 처리를 넘어, 연구자와 리포지토리 간의 긴밀한 협력을 통해 이루어져야 함을 보여준다.

참 고 문 헌

국가과학기술연구회 (2019). 연구데이터 관리 가이드라인 (2019-07).

김주섭, 김선태 (2023). 대학도서관 연구데이터 관리서비스 현황과 제언: 과학기술특성화대학을 중심으로. 한국문헌정보학회지, 57(3), 279-301. https://doi.org/10.4275/KSLIS.2023.57.3.279

김주섭, 김선태, 전예린 (2019). 연구 데이터 관리를 위한 데이터 라이프 사이클 제안. 한국문헌정보학회지, 53(4), 309-340. https://doi.org/10.4275/KSLIS.2019.53.4.309

김지현 (2012). 대학 내 연구자들의 연구데이터 관리에 관한 연구. 한국도서관·정보학회지, 43(3), 433-455. https://doi.org/10.16981/kliss.43.3.201209.433

김판준 (2015). 디지털 큐레이션 연구동향 분석과 과제: 문헌정보학 분야를 중심으로. 정보관리학회지, 32(1), 265-295. https://doi.org/10.3743/KOSIM.2015.32.1.265

박성준, 신정우, 조용찬 (2023). 보건복지 조사데이터의 체계적 관리 및 활용 방안 연구: 한국보건사회

연구원 조사자료 관리 사례를 중심으로(워킹페이퍼 2023-01). 한국보건사회연구원.
신영란, 정연경 (2012). 국내 인문사회 연구데이터 아카이브의 개선방안에 관한 연구. 한국기록관리학회지, 12(3), 93-115.
정영철, 정소희, 이기호, 김은주, 진재현, 안수인 (2020). 전사적 연구데이터 관리 체계 마련을 위한 연구(연구보고서 2020-30). 한국보건사회연구원.
최재은 (2024). 텍스트 마이닝을 활용한 국외 데이터 큐레이션 연구 동향 분석. 정보관리학회지, 41(3), 85-107. https://doi.org/10.3743/KOSIM.2024.41.3.085
한국사회과학자료원 아카이빙사업부 (2024). 자료기탁 가이드 v.1.5.
한나은, 엄정호, 임형준 (2024). 과학기술분야 정부출연연구기관의 연구데이터 관리 방안 연구. 한국문헌정보학회지, 58(2), 151-175. https://doi.org/10.4275/KSLIS.2024.58.2.151
한상우 (2023). 키워드 네트워크 분석을 활용한 국내 연구데이터 연구동향 분석. 한국도서관·정보학회지, 54(4), 393-414. https://doi.org/10.16981/KLISS.54.4.202312.393
Ball, A. (2012). Review of Data Management Lifecycle Models. University of Bath.
Butzlaff, I. (2022). Data deposit guideline (Public version) v2.0. The Austrian Social Science Data Archive. Available: https://aussda.at/fileadmin/user_upload/p_aussda/Documents/Data-Deposit-Guideline_SUF_v2_0.pdf
CESSDA (2020). CESSDA Data Management Expert Guide. Bergen, Norway: CESSDA ERIC. Available: https://doi.org/10.5281/zenodo.3820473
Constantopoulos, P. & Dallas, C. (2007). Aspects of a digital curation agenda for cultural heritage. Available: http://www.dcu.gr/wp-content/uploads/2016/10/Aspects-of-a-digital-curation-agenda-for-cultural-heritage.pdf
Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y., & Papatheodorou, C. (2009). DCC&U: An extended digital curation lifecycle model. International Journal of Digital Curation, 4(1), 34-45. https://doi.org/10.2218/ijdc.v4i1.76
Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2020). Managing and Sharing Research Data: A Guide to Good Practice (2nd ed.). Los Angeles: SAGE.
Faniel, I. M. & Zimmerman, A. (2011). Beyond the data deluge: a research agenda for large-scale data sharing and reuse. International Journal of Digital Curation, 6(1), 58-69. https://doi.org/10.2218/ijdc.v6i1.172
Fear, K. (2013). Measuring and anticipating the impact of data reuse. Doctoral Dissertation, University of Michigan, Ann Arbor, United States.
Higgins, S. (2008). The DCC curation lifecycle model. International Journal of Digital Curation,

3(1), 134-140. https://doi.org/10.2218/ijdc.v3i1.48

ICPSR. (2020). Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (6th ed.). Available: https://doi.org/10.7302/3705

ICPSR. (n.d.a). Life of a Dataset. Available: https://www.icpsr.umich.edu/web/pages/datamanagement/life-of-dataset.html

ICPSR. (n.d.b). Selection and Appraisal. Available: https://www.icpsr.umich.edu/web/pages/datamanagement/lifecycle/selection.html

Mannheimer. S. (2024). Scaling Up: How Data Curation can Help Address Key Issues in Qualitative Data Reuse and Big Social Research. Cham: Springer International Publishing.

Marsolek, W., Wright, S. J., Luong, H., Braxton, S. M., Carlson, J., & Lafferty-Hess, S. (2023). Understanding the value of curation: A survey of researcher perspectives of data curation services from six US institutions. PLOS ONE, 18(11), e0293534. https://doi.org/10.1371/journal.pone.0293534

Peer, L. (2011). Building an open data repository: Lessons and challenges. http://dx.doi.org/10.2139/ssrn.1931048

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. Library and Archives Journal, Issue 1. Available: https://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

Pienta, A. & Reneau, K. (2023). Guide for Sharing Qualitative Data at ICPSR. Available: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/191150/Guide%20for%20Sharing%20Qualitative%20Data%20at%20ICPSR.pdf

Plantin, J.-C. (2019). Data cleaners for pristine datasets: Visibility and invisibility of data processors in social science. Science, Technology, & Human Values, 44(1), 52-73. https://doi.org/10.1177/0162243918781268

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Grant, B., Olendorf, R., & Sandusky, R. J. (2020). Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. PLOS ONE, 15(3), e0229003. https://doi.org/10.1371/journal.pone.0229003

Thomer, A. K., Akmon, D., York, J. J., Tyler, A. R. B., Polasek, F., Lafia, S., Hemphill, L., & Yakel, E. (2022). The craft and coordination of data curation: Complicating workflow views of data science. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW2), 1-29. https://doi.org/10.1145/3555139

UK Data Service. (2022). Collections development selection and appraisal criteria v7.0. Available:

https://ukdataservice.ac.uk/app/uploads/cd234-collections-appraisal.pdf

UK Data Service. (n.d.a). Curated data repository: In-house checks. Available: https://ukdataservice.ac.uk/help/deposit-data/deposit-in-the-curated-data-repository/curated-data-repository-in-house-checks/

UK Data Service. (n.d.b). Prepare your data for deposit. Available: https://ukdataservice.ac.uk/help/deposit-data/prepare-your-data-for-deposit/

UK Data Service. (n.d.c). Recommended Formats. Available: https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yoon, A. (2017). Data reusers' trust development. Journal of the Association for Information Science and Technology, 68(4), 946-956. https://doi.org/10.1002/asi.23730

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Choi, Jaeeun (2024). Analysis of research trends in data curation using text mining techniques. Journal of the Korean Society for Information Management, 41(3), 85-107. https://doi.org/10.3743/KOSIM.2024.41.3.085

Han, Na-Eun, Um, Jung Ho, & Yim, Hyung Jun (2024). A study on research data management methods for government-funded research institutes in the field of science and technology. Journal of the Korean Society for Library and Information Science, 58(2), 151-175. https://doi.org/10.4275/KSLIS.2024.58.2.151

Han, Sangwoo (2023). An analysis of domestic research trend on research data using keyword

network analysis. Journal of Korean Library and Information Science Society, 54(4), 393-414. https://doi.org/10.16981/KLISS.54.4.202312.393

Jung, YoungChul, Jung, Sohee, Lee, Ki-ho, Kim, Eun-ju, Jin, Jaehyun, & An, Su-in (2020). A Study on the Establishment of an Enterprise-wide Research Data Management System (Research Report 2020-30). Korea Institute for Health and Social Affairs.

Kim, Jihyun (2012). A study on university researchers' data management practices. Journal of Korean Library and Information Science Society, 43(3), 433-455. https://doi.org/10.16981/kliss.43.3.201209.433

Kim, JuSeop & Kim, Suntae (2023). Current status and proposal of university library research data management service: focused on science and technology specialized universities. Journal of the Korean Society for Library and Information Science, 57(3), 279-301. https://doi.org/10.4275/KSLIS.2023.57.3.279

Kim, JuSeop, Kim, Suntae, & Jeon, Yerin (2019). Data life cycle proposal for research data management. Journal of the Korean Society for Library and Information Science, 53(4), 309-340. https://doi.org/10.4275/KSLIS.2019.53.4.309

Kim, Pan Jun (2015). An analytical study on research trends of digital curation: focused on library and information science. Journal of the Korean Society for Information Management, 32(1), 265-295. https://doi.org/10.3743/KOSIM.2015.32.1.265

Korea Social Science Data Archive. Data Archiving Department (2024). Guide for Data Deposit V.1.5.

National Research Council of Science and Technology (2019). Research Data Management Guidelines (2019-07).

Park, Seongjun, Shin, Jeongwoo, & Cho, Yongchan (2023). A Study on the Management and Utilization of Health and Welfare Survey Data: Focusing on Survey Data Management Cases in KIHASA (Working Papers 2023-01). Korea Institute for Health and Social Affairs.

Shin, Young-Ran & Chung, Yeon-Kyoung (2012). A study on the improvement plans of the humanities and social sciences research data archives in Korea. Journal of Korean Society of Archives and Records Management, 12(3), 93-115.