사회과학 분야 연구자의 데이터요구와 데이터 재이용 행위에 관한 연구*

An Investigation on Data Needs and Data Reuse Behavior in the Field of Social Sciences

김나연 (NaYon Kim)**
정은경 (EunKyung Chung)***

초 록

오늘날 점차 데이터 집약적으로 변모하는 학문 환경 속에서 데이터는 연구부산물이 아닌 연구성과물로써 학술 커뮤니케이션의 기반으로 자리 잡아가고 있다. 그러나 데이터 공급의 확대나 접근가능성의 확보만으로는 실제적인 데이터 재이용을 담보하는 데 한계가 있다. 이를 극복하기 위해서는 학술연구자의 데이터 재이용 행위와 데이터요구를 심층적으로 파악할 필요성이 있다. 따라서 본 연구는 연구자의 주요 데이터 재이용 행위와 데이터요구를 규명하고자 하였다. 이를 위해 한국사회과학자료원(KOSSDA)의 최근 3개년 데이터 재이용문헌 중 KCI 등재 논문의 저자를 연구대상으로 선정하고, 인터뷰를 수락한 연구자 12명과의 심층면담을 수행하였다. 심층면담 분석결과, 데이터를 재이용하는 요인은 개인적, 경제적, 기술적, 사회적 측면 모두에서 나타났으며, 데이터 재이용 목적에 따라 데이터 그 자체를 이용하거나 데이터가 지닌 맥락정보를 활용하였다. 웹 기반의 정보원으로부터 데이터를 주로 습득하였으나 비공식적인 커뮤니케이션을 통해 파악하는 경우도 있었다. 한편 데이터 재이용 시에 발생하는 학술연구자의 데이터요구를 살펴보면 생산 단위는 기관을, 언어는 영어를, 국가로는 미국을 선호하였다. 또한 조사원 기입식 대인면접 조사 방식으로 수집된 양적 데이터를 우선시하였다. 메타데이터와 식별정보를 충분히 포함한 원자료 수준의 데이터를 긍정적으로 인식하였으나, 접근 및 이용이 통제된 데이터는 데이터가 지닌 가치에 대한 확신을 갖기 어려워 부정적으로 받아들였다. 그러나 데이터의 규모나 최신성과 관련된 선호는 뚜렷하게 나타나지 않았는데 이는 선택 가능한 유사 데이터가 부재하였기 때문이었다.

ABSTRACT

In today's increasingly data-intensive academic environment, data is becoming the foundation of academic communication as a research outcome rather than a research by-product. However, there is a limit to guaranteeing actual data reuse only by expanding the data supply or securing accessibility. In order to overcome this, it is necessary to understand the data reuse behavior and data needs in-depth. Therefore, this study attempted to identify the major data reuse behavior and data needs among researchers. To this end, the authors of KCI papers among the data reuse documents of the Korea Social Science Data Archive (KOSSDA) for the past 3 years were targeted. An in-depth interview was conducted with 12 researchers who accepted the interview. As a result, factors considered when reusing data were personal, economic, technical, and social aspects, and it was found that the data itself was used or contextual information of the data was used depending on the purpose of data reuse. The path to acquiring data is a web-based source of information, and a path through informal communication can also be found. In terms of the data needs, it was found that they prefer English, the United States, and institutional producers. Also they have a clear preference for quantitative data from an interviewer-filled interpersonal interview survey method, rich metadata along with raw data, and data that contains identification information. However, due to the lack of confidence in the value, it is negative for the use of data with controlled access and use, and it is difficult to confirm a clear preference because there is no similar data available for selection in terms of size and freshness.

키워드: 데이터요구, 데이터 재이용 행위, 한국사회과학자료원
data needs, data reuse behavior, Korea Social Science Data Archive, KOSSDA

---

* 본 논문은 김나연의 석사학위논문 「국내 사회과학 학술연구자의 데이터 재이용 행위와 데이터요구 연구」(2019) 내용 일부를 수정・보완한 것임.

** 이화여자대학교 일반대학원 문헌정보학과 석사(nykim105@naver.com) (제1저자)
*** 이화여자대학교 사회과학대학 문헌정보학과 교수(echung@ewha.ac.kr) (교신저자)

■ 논문접수일자 : 2020년 11월 18일 ■ 최초심사일자 : 2020년 12월 3일 ■ 게재확정일자 : 2020년 12월 12일
■ 정보관리학회지, 37(4), 1-26, 2020. http://dx.doi.org/10.3743/KOSIM.2020.37.4.001

1. 서 론

학술 커뮤니티는 지식 발전이라는 목표를 공유하는 개인과 공동체로서 앞선 연구자의 업적을 기반으로 삼는다. 그 과정에서 데이터는 지식의 기본 단위로 중요하게 여겨지는데(Mooney & Newton, 2012), 실험결과의 복제나 관찰연구의 수행, 주장에 대한 검증이 데이터 없이는 불가능하기 때문이다. 이에 Piwowar(2008)은 선행연구로부터 생산된 원 데이터를 재이용할 수 있다면 지식 발전은 보다 빠르고 효율적으로 성취될 수 있다고 보았다. 오늘날 학문이 점차 데이터 집약적으로 변모함에 따라 데이터는 연구 부산물이 아닌 연구 성과물로써 인정받고 학술 커뮤니케이션의 기반으로 자리 잡아가고 있다. 이처럼 데이터가 과학적 발견과 기술적 혁신의 핵심으로 다루어짐에 따라 국제기구, 연구비지원기관, 정부 등에 의해 오픈데이터 개념이 적극적으로 수용되고 있다. 특히 OECD는 2000년대 초 공적영역에서 생산된 데이터의 개방을 주장한 이래 최근에는 오픈데이터 기반의 오픈사이언스를 실현하기 위한 노력을 본격화하였다. 우리나라 또한 공공데이터 및 연구데이터의 개방과 공유가 활성화될 수 있는 환경을 마련하기 위해 노력하고 있다. 대표적으로 한국정보화진흥원은 공공데이터활용지원센터를 설치하고 통합제공시스템인 '공공데이터포털'[1]을 통해 현재 약 5만 5천여 건의 파일데이터, 오픈API, 표준데이터를 제공하고 있다. 한국과학기술정보연구원은 국가연구데이터플랫폼 '데이터온'[2] 서비스를 실시하였으며, 현재 약 340건의 국내 데이터, 해외 데이터, 이미지를 제공하고 있다. 이처럼 데이터 개방 및 공유를 확산시키기 위하여 정부 차원에서 법제도적 토대를 마련하고 기술적 환경을 구축하는데 주력하고 있다.

그러나 데이터 공급의 확대나 접근가능성의 확보만으로는 데이터 재이용을 촉진하기에는 한계가 있으며(Faniel & Jacobsen, 2010) 데이터 재이용 행위, 대상, 주체에 대한 심도 있는 이해가 요구된다. 만약 이와 같은 고려가 충분히 이루어지지 않는다면 Zimmerman(2007)의 주장처럼 데이터 공유나 관련기술이 활용되지 않을 뿐 아니라 해당 분야에 투자된 자원 또한 뚜렷한 성과 없이 낭비될 것이다. 이러한 필요성에도 불구하고 해외에 비하여 국내에서는 데이터 재이용 행위와 그를 유발하는 데이터요구를 심층적으로 파악한 연구는 찾아보기 어렵다. 또한 자연과학 및 공학 분야에서의 데이터 재이용을 다루고 있는 연구는 상당수 있지만 사회과학 분야는 그렇지 못한 편이다. 따라서 우리나라 사회과학 분야의 학술환경이 지닌 특수성을 반영한 연구가 수행되어야 할 필요성이 크다. 이에 본 연구는 디지털 정보자원으로서 데이터가 지니고 있는 잠재성과 데이터 재이용을 통해 창출 가능한 가치에 주목하여, 국내 사회과학 학술연구자를 대상으로 데이터 재이용 행위를 파악하고 핵심적인 데이터요구를 도출하고자 하였다. 이를 위해 한국사회과학자료원(KOSSDA)의 최근 3개년 데이터 재이용문헌

1) https://www.data.go.kr (2020.10.18. 접속)
2) https://dataon.kisti.re.kr (2020.10.18. 접속)

중 KCI 등재 논문의 저자를 연구대상으로 선정하고, 인터뷰를 수락한 연구자 12명과의 심층면담을 수행하였다. 이 연구결과가 이용자 중심의 데이터 리포지토리 정책 전반을 수립하여 시행하고 효율적인 데이터의 수집과 이용을 촉진하고자 하는 노력에 도움이 되기를 기대한다.

2. 관련 연구

2.1 데이터 재이용

데이터 재이용은 데이터 재이용의 목적과 재이용된 데이터의 종수를 기준으로 구분될 수 있다. 우선 데이터를 재이용하는 목적에는 연구대상으로의 재이용과 배경지식으로의 재이용이 있다. 연구대상으로 재이용되는 데이터는 연구의 핵심으로써 주요하게 다루어지며(Wallis, Rolando, & Borgman, 2013), 배경지식으로 재이용되는 데이터는 연구 설계나 데이터 처리 및 분석 시 맥락정보나 보정값을 제공한다(Wynholds, Wallis, Borgman, Sands, & Traweek, 2012). Pasquetto(2018)에 따르면 연구대상으로 데이터를 재이용하는 경우 데이터가 형성하는 상관관계와 인과관계를 추론하여 새로운 지식을 생산하는데 목적이 있으며, 배경지식으로 데이터를 재이용하는 경우 데이터를 통한 비교와 해석을 통해 지식의 맥락화를 달성하고자 한다. 다음으로 재이용된 데이터의 종수에 따라 독립적 재이용과 데이터 합성으로 구분된다. 단일 데이터를 재이용하는 독립적 재이용에는 재현연구가 있으며(Pasquetto, Randles, & Borgman, 2017), 복수 데이터를 재이용하는 데이터 합성에는 비교연구가 있다(Sun & Khoo, 2017).

The Evolutionary Informatics(EvoIO) Working Group(2011)은 데이터 재이용이 수행하는 6가지 주요 기능을 파악하였다. 첫 번째로 출판된 연구결과를 검증하기 위하여 연구를 반복 수행하는 연구 복제(study replication)가 있다. 이는 연구결과 조작이나 왜곡처럼 연구부정행위가 의심되는 경우 행해진다. 두 번째로 동일한 연구결과를 도출하는 데이터를 추가적으로 수집하는 합성(aggregation)이 있다. 이때 수집대상이 되는 데이터는 특정 형식과 유형을 갖춘 경우로 한정된다. 세 번째로 여러 연구에서 이루어진 개별적인 분석을 결합함으로써 기존 연구의 범위나 영향력을 확대하는 메타분석(Meta-analysis)이 있다. 네 번째로 기존과는 다른 연구목적을 가지고 연구결과를 이용하는 재목적화(Re-purposing)가 있다. 다섯 번째로 타 영역이나 다른 유형의 연구에서 생산된 데이터를 합치는 통합(Integration)이 있다. 이는 합성과 유사하나 수집대상에 한계를 두지 않는다는 차이가 있다. 여섯 번째로 통합에서 보다 나아가 개념적 참신성이나 창의성이 가미된 연구결과를 도출하는 종합(Synthesis)이 있다.

이러한 학술연구자의 데이터 재이용에 관한 연구는 연구대상에 따라 크게 두 가지 유형으로 구분될 수 있다. 우선 인용 데이터나 데이터 재이용문헌을 분석하여 재이용 데이터의 특성을 규명하고 데이터 중심의 학술 활동을 통해 구축된 지적구조를 탐색하는 연구가 있으며, 다음으로 데이터 재이용을 경험한 학술연구자를 연구대상으로 데이터 재이용 행위에 영향을 미치는 요인을 규명하고자 한 연구가 있다.

첫 번째 유형의 연구로 조재인(2016)은 2006

년~2015년 간 축적된 DCI 연구데이터 가운데 인용빈도 상위 500위 데이터를 연구대상으로 주제, 유형, 형식, 조사방법론을 파악하였다. 주제 분야는 경제학, 사회학, 인구학, 보건관리/정책학, 가족학에 속한 데이터의 비중이 높게 나타났다. 유형은 데이터의 조사방법, 연구방법론 등을 설명하는 데이터스터디가 대부분을 차지하였다. 형식은 서베이데이터가, 조사방법론으로는 인터뷰가 가장 많았다. 한편 데이터 주제와 유형이 피인용도에 미치는 영향을 통계적으로 분석한 결과 유의미한 상관관계가 확인되었다. 정은경(2018)은 Inter-university Consortium for Political and Social Research(ICPSR) 기탁 데이터를 재이용하여 2017년 발간된 570건에 대해 저자, 형태 및 주제를 살펴보고 제목 키워드에 대해 동시출현단어분석을 수행하였다. 저자의 소속은 건강과학 분야의 미국 대학 또는 연구기관인 경우가 많았다. 형태는 학술지가 대부분으로 나타났으며 주제 분야는 사회과학, 의학, 심리학이 높은 비중을 차지하였다. 이와 같은 경향은 제목 키워드 동시출현단어분석 결과에서도 유사하게 확인되었다. Lin & Lai(2018)는 Taiwan Social Science Citation Index(TSSCI) 내 경제학, 정치학, 사회학, 교육학, 심리학 분야 57개 학술지에서 2001년~2015년 간 발간된 데이터 재이용문헌 1,484건과 해당 문헌이 재이용한 데이터 2,990건을 분석하였다. 데이터 재이용이 가장 빈번한 학문분야는 경제학으로 나타났다. 데이터 출처는 정부기관인 경우가 많았으며 형식으로는 비즈니스데이터가 높은 비중을 차지했다. 한편 통계적 분석을 실시한 결과 데이터 출처 및 형식이 학문분야별로 상이한 차이를 보였다.

두 번째 유형의 연구로 Niu(2009)은 데이터 도큐멘테이션이 재이용 행위에 미치는 영향을 분석하였다. 데이터 도큐멘테이션 수준을 충족도, 편의도, 정확도로 측정하였다. 설문조사와 심층면담 결과를 바탕으로 데이터 재이용을 촉진하기 위한 방안을 다음과 같이 도출하였다. 첫째 학문적 훈련과 전문 경험을 통해 데이터 재이용자의 흡수역량을 높여야 한다. 둘째 데이터 생산자와 재이용자 간 커뮤니케이션 경로를 마련해야 한다. 이를 통해 협업을 활성화하고 외부정보에 대한 접근이나 암묵적 지식의 습득을 용이하게 할 수 있다. 셋째 데이터아카이브는 데이터 생산자가 데이터 도큐멘테이션 관련 설명문을 숙지하고 이행할 수 있도록 여러 도구와 방안을 활용하여야 한다. 넷째 데이터 생산자에게 인센티브를 제공하여 데이터 도큐멘테이션에 대한 동기를 유발하여야 한다. Faniel, Kriesberg, & Yakel(2012)은 심층면담을 통해 초보 사회과학 연구자가 데이터를 재이용하는 과정에서 내리는 주요 의사결정을 살펴보았다. 우선 데이터 신뢰도를 판단하기 위해 데이터 수집방법과 코딩 절차를 주요하게 참조하였다. 다음으로 학술공동체의 수용 여부에 따라 데이터 재이용 여부를 결정하였으며, 이 과정에서 데이터 생산자의 출판물이나 선행연구를 검토하였다. 마지막으로 데이터를 매칭하거나 합성하고자 할 때는 지도교수로부터 의견을 구하여 확신을 얻고자 하였다. Yoon(2014)은 사회과학 연구자의 질적 데이터 재이용 경험을 탐구하였다. 데이터 재이용자는 데이터의 신뢰도를 보장하고 구체적인 맥락정보를 마련하고자 사적 관계를 기반으로 질적 데이터를 획득하였다. 그러나 접근이 제한되거나 맥락정보가 누

락된 경우 또는 데이터 재이용과 관련된 학문분야별 관행이 엄격하거나 학술공동체의 수용도가 낮을 때 어려움이 초래되었다. Curty(2016)는 데이터 리포지토리를 이용한 경험이 있는 사회과학 연구자로부터 데이터 재이용 인식과 경험에 영향을 미치는 요인 25개를 도출하고 6개 영역으로 구분하였다. 구체적으로 살펴보면 첫째 인지된 이익 영역에서는 지식 확장, 절약, 선 보증이, 둘째 인지된 어려움 영역에서는 저평가될 두려움, 윤리규범을 위반할 두려움, 오차현상, 미확인된 오류에 대한 취약성이, 셋째 인지된 노력 영역에서는 오래된 데이터를 통한 혁신, 데이터에 대한 접근권한 획득, 데이터 탐색 과정, 연구문제와의 대응, 재이용 준비, 원 연구에 대한 이해가, 넷째 재이용성 평가 영역에서는 데이터 도큐멘테이션, 데이터 적합성, 데이터 생산자의 신뢰성, 데이터 품질, 연구의 견고성이, 다섯째 활성화 요인 영역에서는 데이터 도큐멘테이션의 효용도, 데이터 리포지토리 유용성, 데이터 생산자에 대한 접근가능성, 관련 지원 제공 여부, 훈련과 전문지식이, 여섯째 사회적 요인 영역에서는 학문분야별 수용도, 동료 연구자 협력이 파악되었다. Faniel, Kriesberg, & Yakel(2016)은 데이터 리포지토리의 유용성을 평가하는 차원에서 데이터 품질에 따른 데이터 재이용자의 만족도를 분석하였다. 데이터 품질을 판단하는 4개 속성 가운데 적합성을 제외한 완전성, 접근가능성, 신뢰성이 데이터 재이용 만족도와 연관이 있었다. 한편 간접적 요인으로 데이터 도큐멘테이션 품질이 데이터 생산자의 명성보다 영향력이 높았다. Yoon(2016)은 데이터 재이용을 저해하는 요인을 밝히고자 사회과학 연구자의 데이터 재이용 실패 경험에 주목하였다. 데이터 재이용의 실패는 데이터에 대한 설명이 잘못되었거나 불완전한 경우, 데이터에 대한 접근이 어려운 경우 또는 데이터 도큐멘테이션의 효용도가 낮은 경우 발생하였다. 또는 데이터 값이 부재하거나 오류가 있을 때나 원 연구의 데이터 조작, 클리닝 및 분석에 문제가 있을 때도 재이용에 실패하였다. 따라서 성공적인 재이용을 위해서는 접근성과 상호운용성을 보장하고 충분한 도큐멘테이션과 더불어 관련 기관, 학술공동체, 데이터 생산자로부터의 지원이 요구되었다. Yoon(2017)은 사회과학 연구자가 데이터를 재이용하면서 데이터에 대한 신뢰를 형성하는 과정을 단계별로 고찰하였다. 초기 단계에서는 데이터를 탐색하면서 데이터의 적합성 및 편의성을 바탕으로 신뢰를 마련하였다. 데이터를 획득하고 조사하는 잠정적 단계에서는 데이터의 타당성, 신뢰성, 견고성과 같은 내재적 속성이나 도큐멘테이션 수준을 토대로 신뢰 여부를 결정하였다. 데이터를 분석하는 최종 단계에서 문제가 발생하는 경우 데이터 생산자나 학술공동체의 지원이 신뢰도를 회복하는데 영향을 미쳤다. Yoon & Kim(2017)은 앞선 귀납적 연구들을 토대로 가설을 설정하고 이를 검증함으로써 사회과학 분야 내 데이터 재이용 행위를 연역적으로 설명하고자 하였다. 이를 위해 계획된 행동이론과 기술수용이론을 적용한 모형을 마련하였다. 설문조사 분석결과, 데이터 재이용에 대한 인식이 긍정적일수록 데이터를 재이용하려는 의도가 향상되었다. 인지된 이익이 큰 경우, 학문분야의 규범이 친화적인 경우 또는 데이터 리포지토리 유용성이 갖추어진 경우 데이터 재이용을 긍정적으로 인식하였다.

2.2 데이터의 재이용성과 재이용 행위

데이터 재이용이란 재이용이라는 연구방법을 활용하여 데이터라는 정보원을 학술적 목적으로 탐구한 행위로 이차분석의 일종이다. 다만 데이터에 한정된 이차분석을 표현하기 위해 데이터 재이용이라는 용어가 도입되었다(Zimmerman, 2007). 대부분의 연구에서 따르고 있는 Zimmerman(2003)의 개념 정의를 살펴보면 데이터 재이용이란 특정 목적을 가지고 수집된 데이터를 새로운 연구문제를 탐구하는데 활용하는 행위이며 데이터에 대한 이차적 이용을 의미한다. Law(2005), Curty & Qin(2014)은 연구방법을 데이터 재이용을 판단하는 기준으로 제안하고 기존 연구와 연구목적이 동일하더라도 새로운 연구방법을 시도하는 경우 데이터 재이용으로 보았다. Pasquetto, Randles, & Borgman(2017)은 연구자를 중심으로 데이터 재이용의 개념을 확립하였다. 따라서 리포지토리에 기탁된 데이터가 탐색을 통해 데이터 생산자가 아닌 새로운 연구자에 의해 이차적으로 이용되는 경우로 데이터 재이용을 정의하였다.

데이터 재이용성(reusability)이란 데이터가 재이용되기 위해 갖추어야 할 일정 조건을 의미하며, 잠재적 재이용자는 데이터가 지닌 속성을 토대로 데이터 재이용을 결정한다(Curty, 2016). 우선 Faniel & Jacobsen(2010)은 데이터 재이용성으로 적합성, 이해가능성, 신뢰성을 제시하였다. 각 속성을 살펴보면, 첫째 적합성은 연구자가 다루고자하는 문제에 데이터가 적합한가이다. 이는 연구자가 설계한 연구모형이나 연구분야의 전문지식을 기반으로 판단될 수 있다. 둘째 이해가능성은 데이터에 대한 맥락정보가 충분히 제공되고 있는가이다. 표본집단 설정 및 검증법, 데이터 측정 및 처리법, 가설과 실행에 대한 세부적인 맥락정보 뿐만 아니라 학술 커뮤니티 내에서 공유되는 이해기반이 함께 제공되어야 한다. 셋째 신뢰성은 확실성과 타당성으로 구성된다. 맥락정보가 제공하는 데이터 측정법을 통해 동일한 데이터가 지속적으로 측정된다면 확실성이 보장된다. 한편 연구 진행과정에서 발생한 문제의 해결과정을 살펴봄으로써 타당성이 판단된다. 이 과정에서 연구자는 연구분야의 전문지식과 데이터에 대한 도큐멘테이션을 바탕으로 누가 신뢰할 만한지, 무엇을 신뢰할지 결정하게 된다. 다음으로 Sun & Khoo(2017)는 주제적 적합성, 데이터 품질, 유용성을 데이터 재이용성으로 이해하였다. 주제적 적합성은 데이터가 연구목적에 적합한지, 원하는 변수를 포함하고 있는지, 적절한 방법으로 수집되었는지, 충분한 표본 수로부터 수집되었는지를 통해 판단된다. 데이터 품질은 연구방법론의 확실성과 타당성에 의해 결정된다. 재수행이 가능하도록 정확하고 지속적인 방식으로 데이터가 수집 및 코딩된 경우 확실성이 충족되며, 데이터의 변수가 연구문제를 잘 반영하고 있는 경우 타당성이 만족된다. 유용성은 데이터가 새로운 결과나 통찰을 제공하는지 혹은 데이터 재이용자가 지닌 연구 상의 필요를 충족시키는지를 나타낸다. 끝으로 Curty(2016)는 데이터 도큐멘테이션, 데이터 적합성, 데이터 생산자 및 출처에 대한 신뢰성, 데이터 품질, 연구의 견고성을 데이터 재이용성으로 보았다. 첫째 데이터 도큐멘테이션은 완전성과 명확성을 기준으로 판단되는데 코드북이나 데이터 사전, 데이터 수집 절차 보고서, 데이터 수집도구, 관련 출판물, 사용자 안내

서, 통계 매뉴얼, 데이터 추출 소프트웨어, 생명윤리위원회 문서 등을 토대로 삼는다. 둘째 데이터 적합성은 연구자의 목적에 데이터가 부합하는가를 나타내며, 주제나 데이터 분석 수준, 데이터 유형 등이 고려된다. 셋째 데이터 생산자 및 출처에 대한 신뢰성은 데이터 재이용성을 판단하는 초기 단계에 큰 영향을 미치는 속성이다. 넷째 데이터 품질은 견고성과 완전성을 기준으로 측정된다. 견고성은 데이터가 정확하게 추출된 경우에, 완전성은 부재한 데이터가 없거나 최소인 경우에 보장된다. 다섯째 연구의 견고성은 원 연구의 연구설계와 실행을 기준으로 판단되는데, 연구의 목적, 대상, 방법론이 연구 설계와 실행을 통해 연구결과물로 전환되는 과정이 얼마나 적합했는지를 기준으로 한다.

위에서 제시된 데이터 재이용성을 평가하는 기준이 되는 데이터 속성을 종합하면, 주제적 적합성, 이해가능성, 생산자 또는 데이터 품질에 대한 신뢰성, 이용가능성이 있다.

3. 연구방법

국내 대표 사회과학 데이터 아카이브인 한국사회과학자료원(Korea Social Science Data Archive, KOSSDA) 내 '자료이용문헌DB'를 활용하여 최근 5개년(2014년~2018년) 간 생산된 자료이용문헌 총 1,546건을 파악하였다. 이 가운데 국내 학술연구자가 출간한 KCI 등재 학술논문 총 626건과 재이용된 데이터 19종의 현황을 파악하였다. 아울러 연구자의 데이터 재이용 행위와 데이터요구를 심도 있게 살펴보기 위하여 심층면담을 수행하였다. 데이터를 재이용한 경험이 있는 연구자를 선별하기 위하여 최근 3개년(2016년~2018년) 간 생산된 자료이용문헌 총 395건을 파악한 후, 동일 저자에 의해 생산된 경우는 중복 처리하여 총 304건의 교신저자 혹은 주저자의 이메일 주소를 확보하였다. 이 가운데 2회 이상 논문을 출간한 학술연구자 54명에게 2019년 4월 10일 심층면담 요청문을 발송하였으며 수락의사를 밝힌 12명을 〈표 1〉과 같이 최종 참여자로 확보하였다. 경력기간은 데이터를 재이용하여 연구성과물을 출간한 경험을 기준으로 산출하였으며 타 기초정보에 비해 고른 분포를 보였다. 따라서 경력기간이 15년 이상인 그룹A, 10년 이상 15년 미만인 그룹B, 5년 이상 10년 미만인 그룹C로 구분하여 비식별화 하였다. 심층면담은 3개 영역 21개 문항으로 구성된 반구조화된 질문지를 바탕으로 2019년 4월 15일부터 5월 7일까지 평균적으로 약 1시간동안 이루어졌다.

데이터 분석은 크게 2단계로 이루어졌다. 첫 번째 단계에서는 데이터와 데이터를 재이용한 학술논문의 현황을 파악하고자 하였다. 이를 위해 최근 5개년(2014년~2018년) 간 데이터를 재이용하여 생산된 문헌 중에서 KCI 등재지에 게재된 논문 총 626건과 해당 논문에서 재이용된 데이터를 연도별, 주제별로 분석하였다. 두 번째 단계에서는 데이터를 재이용한 경험이 풍부한 연구자 중 심층면담에 동의한 12명의 연구자와의 인터뷰를 수행하였다. 심층면담 내용은 전사 후 근거이론에서 개발된 반복적 비교분석법을 적용하여 개방 코딩, 분석 코딩(범주화), 선택・이론 코딩(범주 확인) 순으로 분석하였다. 심층면담을 위해서 사용된 반구조화된 질문항목은 〈표 2〉와 같이 크게 세 영역으로

〈표 1〉 심층면담 참여자 기초정보

구분	연령	성별	최종학력	전공분야	소속 및 직책	경력기간
A1	50대	남	박사	사회학	사립대학교 교수	20년
A2	40대	남	박사	경제학	정부출연연구기관 연구위원	17년
A3	50대	남	박사	공중보건학	국립대학교 교수	15년
B1	40대	여	박사	사회학	국립대학교 교수	14년
B2	50대	남	박사	정치학	사립대학교 교수	12년
B3	50대	남	박사	정치학	교장	12년
B4	40대	남	박사	경영학	사립대학교 교수	10년
C1	40대	남	박사	사회학	국립대학교 교수	9년
C2	50대	남	박사	경제학	정부출연연구기관 연구위원	8년
C3	30대	남	박사	정치학	대학교부설연구소 연구원	8년
C4	40대	남	박사	사회학	사립대학교 강사	5년
C5	30대	여	석사	사회복지학	정부출연연구기관 연구위원	5년

〈표 2〉 심층면담을 위한 반구조화된 질문항목

영역	세부 항목
데이터 재이용 행위(4)	- 데이터 재이용 요인 - 데이터 재이용 유형 - 데이터 파악 및 습득 경로 - 학술연구활동 주기 내 데이터 재이용 과정
데이터요구(7)	- 생산자 - 생산국가 및 언어 - 유형 및 수집방법 - 규모 - 처리수준 - 접근 및 이용통제 - 최신성
연구참여자 기초정보(5)	- 전공분야(최종학력 기준) - 관심 연구분야 및 주제 - 소속 및 직책 - 데이터 재이용 경력기간 - 데이터 재이용 빈도(최근 3년 기준)

구성되었다. 데이터 재이용 행위 영역은 4개의 세부 항목이 포함되어 있으며, 데이터요구 항목은 7가지의 세부항목으로 구성되었다. 이 외에도 연구참여자의 인구통계적인 기초 항목이 포함되었다.

4. 분석결과

4.1 데이터와 재이용 학술논문 현황

분석대상인 626건의 학술논문에서 재이용된

데이터 19종을 연도별로 살펴보면, 〈표 3〉과 같다. 빈도수 상위 3순위는 한국노동패널조사, 한국종합사회조사, 근로환경조사로 전체의 과반 수준에 해당한다. 한편 재이용 데이터로부터 생산된 데이터 재이용문헌은 평균 33건이었으며 총 6종의 데이터(한국노동패널조사, 한국종합사회조사, 근로환경조사, 사업체패널조사, 국민노후보장패널조사, 전국범죄피해조사)가 평균 이상의 재이용 빈도수를 보였다. 연도별 증감 추세는 확인되지 않았으나 특정 데이터가 지속적으로 반복하여 재이용 되는 경향이 파악되었다.

KOSSDA 데이터 주제 분류를 기준으로 한 주제별 재이용 데이터 현황은 〈표 4〉와 같다. 대분류 기준으로 7개 주제영역, 소분류 기준 12개 주제영역에 분포되어 나타났다. 우선 대분류 수준에서 분포도 상위 3순위는 경제, 일반, 사회와문화 영역으로 확인되었으며 전체의 약 91%를 차지하였다. 소분류 수준에서 분포도 상위 3순위는 고용과노동, 일반, 산업과기업 영역으로 파악되었다. 각 주제영역에서 데이터가 재이용된 빈도는 약 52회로 총 3개 주제영역(고용과노동, 일반, 산업과기업 영역)이 평균 이상의 재이용 빈도를 보였다. 이를 통해 특정 주제영역에 속하는 데이터가 집중적으로 재이용되는 현상을 찾아볼 수 있다.

〈표 3〉 연도별 재이용 데이터 현황

재이용 데이터명	연도별 빈도수					합계(비율)
	2014	2015	2016	2017	2018	
한국노동패널조사	37	42	41	40	11	171(27.3%)
한국종합사회조사	35	29	46	37	10	157(25.0%)
근로환경조사	11	15	21	16	14	77(12.3%)
사업체패널조사	15	12	9	12	11	59(9.4%)
국민노후보장패널조사	0	0	11	13	15	39(6.2%)
전국범죄피해조사	4	3	8	11	9	35(5.6%)
한국아동・청소년행복지수조사	3	2	6	1	5	17(2.7%)
통일의식조사	3	3	6	4	0	16(2.6%)
한국청소년패널조사	0	0	0	1	15	16(2.6%)
한국기업혁신조사	3	1	4	0	4	12(1.9%)
국민여가활동조사	1	1	2	1	1	6(1.0%)
외래관광객실태조사	0	2	3	0	1	6(1.0%)
문화향수실태조사	2	2	1	0	0	5(0.8%)
문화예술인실태조사	0	1	1	1	0	3(0.5%)
공직부패의실태에관한설문조사	0	0	1	1	0	2(0.3%)
정부역할과삶의질에대한국민인식조사	0	0	0	2	0	2(0.3%)
한국의범죄피해에대한조사	2	0	0	0	0	2(0.3%)
다문화가족지원사업평가조사	1	0	0	0	0	1(0.2%)
초・중학교영어교실활동에관한교사인터뷰	0	1	0	0	0	1(0.2%)

〈표 4〉 주제별 재이용 데이터 현황

KOSSDA 주제분류		재이용 데이터명(19)	빈도(비율)
대분류(7)	소분류(12)		
경제	고용과 노동	근로환경조사	248(39.6%)
		한국노동패널조사	
	사업과 기업	사업체패널조사	71(11.3%)
		한국기업혁신조사	
	소 계		319(50.9%)
일반	일반	한국종합사회조사	157(25.1%)
	소 계		157(25.1%)
사회와 문화	노인과 고령화	국민노후보장패널조사	39(6.2%)
	청소년과 사회화	한국아동・청소년행복지수조사	32(5.1%)
		한국청소년패널조사	
	문화와 여가	국민여가활동조사	20(3.2%)
		문화예술인실태조사	
		문화향수실태조사	
		외래관광객실태조사	
	가족과 젠더	다문화가족지원사업평가조사	1(0.2%)
	교육	초・중학교영어교실활동에관한교사인터뷰	1(0.2%)
	소 계		93(14.9%)
사회문제와 복지	범죄와 안전	전국범죄피해조사	37(5.9%)
		한국의범죄피해에대한조사	
	소 계		37(5.9%)
정치, 통일 및 국제관계	통일과 국제관계	통일의식조사	16(2.6%)
	소 계		16(2.6%)
정부와 시민사회	정부운영과 행정	공직부패의실태에관한설문조사	2(0.3%)
	소 계		2(0.3%)
미분류	미분류	정부역할과삶의질에대한국민인식조사	2(0.3%)
	소 계		2(0.3%)
합 계			626(100%)

KCI 학술지 주제 분류를 기준으로 한 데이터 재이용 학술논문의 주제별 현황은 〈표 5〉와 같다. 대분류 기준 8개 학문분야, 중분류 기준 41개 학문분야에 분포되어 나타났다. 사회과학 분야의 데이터라는 특성을 반영한 결과로 사회과학 분야의 학술논문이 501건(80.2%)으로 가장 큰 비중을 차지하였다. 사회과학 분야를 중분류 수준에서 살펴보면 사회학, 사회과학일반, 경제학, 경영학 등의 순으로 나타났다. 사회과학 이외의 분야로는 의약학 대분류 중에서 예방의학 분야와 복합학 대분류 중에서 학제간연구 분야에서 데이터 재이용 학술논문이 다수 발간되었다. 각 학문분야에서 생산된 데이터 재이용문헌은 평균 15건으로 총 12개 학문분야(사회학, 사회과학일반, 경제학, 경영학, 사회복지학, 행정학, 기타사회과학, 국제/지역개발, 정치외교학, 교육학, 예방의학, 학제간연구)가 평균 이상의 데이터 재이용문헌 건수를 보였다.

〈표 5〉 데이터 재이용 학술논문의 주제별 현황

KCI 학술지 주제분류		데이터 재이용 학술논문	
대분류(8)	중분류(41)	건수	비율(%)
사회과학	사회학	86	13.7
	사회과학일반	77	12.3
	경제학	65	10.4
	경영학	53	8.5
	사회복지학	50	8.0
	행정학	45	7.2
	기타사회과학	22	3.5
	국제/지역개발	21	3.4
	정치외교학	21	3.4
	교육학	19	3.0
	법학	11	1.8
	정책학	10	1.6
	관광학	8	1.3
	지역학	5	0.8
	무역학	2	0.3
	미지정	2	0.3
	지리학	2	0.3
	신문방송학	1	0.2
	회계학	1	0.2
	소 계	501	80.2
의약학	예방의학	36	5.8
	간호학	10	1.6
	의학일반	2	0.3
	직업치료학	1	0.2
	소 계	49	7.9
복합학	학제간연구	27	4.3
	여성학	6	1
	과학기술학	3	0.5
	감성과학	2	0.3
	기술정책	2	0.3
	문헌정보학	1	0.2
	소 계	41	6.6
공학	공학일반	6	1.0
	산업공학	3	0.5
	안전공학	2	0.3
	컴퓨터공학	2	0.3
	소 계	13	2.1

KCI 학술지 주제분류		데이터 재이용 학술논문	
대분류(8)	중분류(41)	건수	비율(%)
자연과학	통계학	8	1.3
	생활과학	3	0.5
	자연과학일반	1	0.2
	소 계	12	2
인문학	기타인문학	2	0.3
	역사학	2	0.3
	종교학	2	0.3
	영어와문학	1	0.2
	소 계	7	1.1
예술체육학	체육	2	0.3
	소 계	2	0.3
농수해양학	조경학	1	0.2
	소 계	1	0.2
합 계		626	100

4.2 데이터 재이용 행위 분석

4.2.1 데이터 재이용 요인

Curty(2016)는 데이터 재이용 행위를 절약이라는 경제적 차원의 행위로만 이해해온 기존 연구들의 한계를 지적하고 개인적, 기술적, 사회적 차원에서도 데이터가 재이용이 고려된다고 보았다. 이를 토대로 심층면담 참여자가 밝힌 데이터 재이용 요인을 구분하였다.

첫 번째, 개인적 차원에서 파악된 데이터 재이용 요인은 연구상의 용이성과 데이터 접근성이다. "*한 번 쓴 데이터는 다시 쓰기에 너무 유용하니까요. … 연구에 용이해지죠.*"(참여자A2)처럼 연구 상의 용이성은 이미 재이용한 데이터를 다시 재이용하고자 할 때 발생하였다. 재이용 데이터에 대한 높은 이해도를 바탕으로 적은 수준의 노력만으로도 생산적인 연구를 수행할 수 있기 때문이다. 이를 Niu(2009)은 데이터에 대한 메타데이터가 데이터를 재이용 하는 학술연구자 개인에게 내재화됨으로써 배경지식이 되는 과정으로 이해하였다. 또한 "*자료가 아무래도 구하기가 쉬우니까요. … 접근이 쉬워서*"(참여자C5)와 같은 데이터 접근성은 데이터의 생산 및 제공이 공적 이익을 목표로 하는 기관에 의해 이루어질 때 확보되었다. 사적 이윤의 추구를 목적으로 하는 민간과 달리 공공은 데이터를 통해 사회현상을 이해하고 사회문제 해결방안을 마련하고자 접근과 이용을 적극적으로 보장하기 때문이다.

두 번째, 경제적 차원에서 확인된 데이터 재이용 요인은 연구상의 효율성이다. 효율성은 양적 개념인 능률성과 질적 개념인 효과성을 합한 복합적 개념으로, 최소의 투입으로 최대의 산출을 얻되 기대 목표에 대한 성취를 달성할 때 보장된다. 인력, 시간, 도구 등을 투입으로, 연구성과물 생산건수를 산출로 환산하면 효율성을 진단할 수 있다. 학술연구자들은 데이터를 재이용함으로써 "*비용이죠. … 돈과 시간*"(참여자C2)으로 인

한 부담을 감소시켜 능률성을 높이고, *"개인이 감당할 수 있는 금액으로 낮춰버리면 데이터 퀄리티가 떨어져요. … 그런 데이터는 사실 전혀 의미가 없어요."*(참여자C3)와 같이 데이터의 품질을 확보하여 효과성을 강화하였다.

세 번째, 기술적 차원에서 발생하는 데이터 재이용 요인은 정보기술 환경의 변화와 데이터 준비다. Law(2005)가 지적한 바와 같이 오늘날 데이터 공유가 보다 쉬워지고, 복잡한 연구분석을 빠르게 수행할 수 있을 뿐 아니라 대규모의 데이터 리포지토리가 갖추어짐에 따라 데이터 재이용에 대한 학술연구자의 역량이 강화된 것이다. *"어차피 컴퓨터로 다 하는데요. … 윗세대 연구자들이랑 (차이가 있죠.) … 다양한 분석과 데이터를 접하게 되죠, 환경적으로요."*(참여자A1)에서 나타나듯이 데이터 재이용과 관련된 정보기술의 급속한 발달과 보급을 통해 데이터 재이용의 진입장벽이 낮아졌다. 또는 *"통계분석을 할 수 있게 편리하게 되어있는 부분이 많아요. 데이터 준비라고 하는데 … 클리닝, 코딩, 가중치를 계산해서 넣어주는 등이요."*(참여자A3)처럼 데이터가 재이용에 적합한 형태로 학술연구자에게 제공되기도 한다. 네 번째, 사회적 차원에서 발생한 데이터 요인으로는 학술적 풍토의 변화와 연구분야 및 주제의 특성이 있다. *"우리나라도 연구중심으로 많이 바뀌어서요. … 새로운 데이터를 꼭 수집할 필요가 있느냐, 기존에 있는 걸 재분석하는 게 그게 훨씬 더 좋은 경우가 많다는 것을 (이야기 하죠)."*(참여자A1)를 통해 확인되는 바와 같이 오늘날 학술적 풍토가 점차 연구 중심적으로 변화함에 따라 데이터 재이용을 통한 학술연구 또한 활성화되어 학술공동체로부터 그 가치를 인정받고 있다.

한편 *"전공(경영학)의 관심사가 거시적인 측면에 있기 때문"*(참여자B4)이거나 *"경제학자들은 대부분이 있는 자료를 쓰죠."*(참여자C2)와 같이 특정 연구분야 및 주제의 경우에 학술공동체 내 데이터 재이용이 보편화되어 있어 그 수용성이 상대적으로 높은 경우도 있었다.

4.2.2 데이터 재이용 유형

Pasquetto(2018)은 데이터 재이용의 목적에 따라 연구대상으로써 데이터 재이용과 배경지식으로써 데이터 재이용을 구분하였다. 심층면담 참여자 대부분은 재이용 데이터를 연구대상으로만 활용하였으나 일부는 *"양쪽(연구목적, 배경지식으로 활용) 다 하게 되고요."*(참여자A3)와 같이 재이용 데이터를 연구대상과 배경지식으로 함께 활용하는 것으로 확인되었다. 아울러 데이터 재이용의 목적에 따라 재이용하는 데이터의 처리수준이 상이하게 나타났는데 연구대상인 경우 원 데이터를, 배경지식인 경우 생산기관이 집계했거나 분석한 요약본이나 보고서를 선호하였다. 현황이나 추세를 확인하기 위해 배경지식으로써 데이터를 활용하는 과정에서 비교가능성이나 신뢰성이 확보될 경우 자연스럽게 연구대상으로도 재이용한다고 밝힌 경우도 있었다. *"기관마다 … 보고서를 작성하기도 하고 언론기사 같은 것들을 만들기도 하잖아요? … 먼저 보고 나서 연구나 교육에 이용할까를 정하지(요)."*(참여자C1), *"비교가능성, 신뢰성이 충분히 확보된 데이터면 배경지식으로도 당연히 선호가 될 거고 … 분석 대상으로서의 데이터로서도 선호"*(참여자C3)된다는 점을 통해 데이터 재이용이 학술연구자의

연구과정 내에서 개별적으로 이루어지는 단절적 행위가 아닌 일련의 연속적 과정임이 드러난다.

데이터 정보원을 기준으로 단일 또는 복합으로 재이용된 데이터의 종수를 구분할 수 있다. 단일이란 한 개의 데이터를 재이용한 경우를 포함하여 두 개 이상의 데이터를 재이용하였으나 해당 데이터들이 동종(homogeneity)인 경우가 해당된다. 복합이란 두 개 이상의 데이터를 재이용하였으며 해당 데이터들이 이종(heterogeneity)인 경우가 해당된다. 심층면담 결과 단일 데이터를 재이용한 경우가 다수였으나 한 개의 데이터보다는 동종 데이터를 두 개 이상 재이용하였다. 한 개의 데이터는 주로 기존연구를 동일하게 반복함으로써 연구결과를 검증하고자 하는 목적으로 재이용되었다. 두 개 이상의 동종 데이터는 크게 두 가지 목적으로 재이용되었는데 장기간의 변화와 추세 및 인과관계를 도출하거나 또는 국가적 특수성 혹은 국가 간 동질성을 파악하였다. 이는 "*다른 사람들의 연구를 검증하기 위한 합리적 수단으로 (데이터 재이용)한 적이 많아요. … 시점이 다른 데이터도 분석을 하고요. … 지역이 다른 경우도 해보고요.*"(참여자A1)에서 확인된다. 다음으로 복합 데이터를 재이용하는 경우에는 상이한 데이터 정보원으로부터 생산된 이종 데이터를 상호결합하거나 합성하여 연구목적에 부합하는 데이터셋을 직접 구성하였다. 상호결합은 연구대상의 단위가 "*분석대상이 국가 단위*"(참여자C3)이거나 "*지역별로 나누어져*"(참여자C4) 있을 때 이루어졌으며 합성은 "*예를 들어서 … 호흡기 질환이면 보건 데이터도 있어야 하지만 환경 데이터도 있어야 하잖아요. 다른 분야의 것을 (합성하죠).*"(참여자C5)와 같이 타 학문분야에서 생산된 데이터를 기반으로 변수를 확장시키고자 할 때 수행되었다.

4.2.3 데이터 파악 및 습득 경로

심층면담 참여자 대부분은 웹 기반의 정보원을 통해 재이용 데이터를 파악하거나 습득하였다. 주요 정보원으로 생산 및 제공기관 웹사이트, 검색엔진 사이트, 학술 데이터베이스, 데이터 아카이브가 파악되었으며 상황이나 목적에 따라 보다 적합한 정보원이 선택되었다. 재이용하고자 하는 데이터에 대한 이해도가 높은 경우 생산 및 제공기관 웹사이트를 통해 재이용 데이터를 직접 파악한 뒤 습득하였다. 이는 "*직접 기관 홈페이지에 가서 찾아보는 것 같아요. … 사전지식이 있어서.*"(참여자C2)에서 드러난다. 그러나 생소한 연구분야 및 주제를 다루는 경우에는 "*데이터 보다는 연구를 찾죠. … 논문들을 보면서 어떤 데이터를 이용했는지를 보고(요).*"(참여자A2)처럼 학술 데이터베이스를 통해 관련 선행연구를 살펴보았다. 즉 타 학술연구자에게 널리 이용되었거나 잘 알려져 있는 데이터를 재이용함으로써 연구과정에서 발생할 수 있는 부담이나 위험을 최소화하고자 하였다. 연구문제 해결에 적합한 데이터를 판단할 때 학술공동체로부터의 수용 여부가 영향을 크게 미쳤다. 더 나아가 "*데이터를 어디서 찾았는지 … 보거든요. 아마 그런 과정에서 결국에는 남들이 많이 쓰는 데이터를 쓰게 되는 경우가 생기게 되는 것 같아요.*"(참여자C3)와 같이 학술논문의 인용과 마찬가지로 일정한 경향성이 발생되었다. 재이용되어 학술공동체로부터 신뢰성과 타당성을 인정받은 바 있는 데

이터는 타 데이터에 비해 재이용되기 쉬워지는 것이다.

한편 *"주제어로 검색하는 건 … 어떤 데이터가 있다는 … 그것조차 모를 때죠. … 어디에 가야 데이터를 얻을 수 있는지조차도 모를 때는 … 구글에서."*(참여자B2)에서 확인되듯이 데이터나 생산 및 제공기관의 존재 여부 자체를 파악할 수 없는 경우 일부 심층면담 참여자는 특정 검색엔진 사이트에서 주제어 검색을 수행하였다. 웹 기반의 정보원이 아닌 학술공동체 내 커뮤니케이션을 통해 재이용 데이터를 파악 또는 습득한 경우도 있었다. 특히 개인 연구자가 생산한 연구데이터는 데이터 리포지토리와 같은 공식적 데이터 제공기관보다는 학술공동체 내 비공식적 커뮤니케이션을 통해 제공되었다. *"연구자끼리 알음알음 서로 물어서"*(참여자A3), *"직장 내에서 같이 일하는 연구원이라든지 … 같은 연구실에 있는 분이나 교수님(이죠)."*(참여자C5), *"연구논문이나 프로젝트에 참여해있을 때"*(참여자C4)처럼 정보획득은 의도적으로 이루어지기도 하였지만 우연적인 경우도 있었다.

4.2.4 학술연구 활동주기와 데이터 재이용

McCall & Appelbaum(1991)은 일반적인 연구설계 및 과정은 순차적이며 선형적으로 이루어지는데 반해 데이터 재이용에 해당하는 이차분석은 연구방법 특성상 이미 데이터가 수집되어있는 상태이므로 연구과정이 재구조화되며 순환적 성격을 지니게 된다고 보았다. 이와 같은 특성을 반영하여 데이터 중심의 학술연구 활동 주기를 파악하고자 하였다. 정동열, 조찬식(2018)에 따르면 과학적 조사연구의 일반 과정은 일반적 문제제기, 문헌연구, 본 연구문제 기술, 연구방법 설정, 자료 수집, 자료분석 및 처리, 가설검증 결과 제시, 연구결과 작성이라는 8단계로 구성된다. 연구에 필요한 자료원 혹은 조사대상자나 조사대상물에 대한 표집은 연구문제에 대한 전체적인 설계가 완료된 이후 '자료수집' 단계에서 이루어진다. 그러나 데이터를 재이용 하는 연구의 경우 재이용 데이터의 파악 및 습득이 여러 단계에서 나타났다.

우선 연구의 목적이나 주제를 설정하는 '일반적인 문제제기' 단계에서 재이용 데이터를 파악 및 습득하는 경우가 있었다. *"가설까지 가지 않고요. … 이미 연구질문 시기에 데이터가 결정"*(참여자B4) 되는 등 연구질문 혹은 주제분야를 선정하는 과정에서 데이터에 대한 사전지식을 바탕으로 재이용하고자 하는 데이터가 확정되었다. 다음으로 개념적 틀을 도출하거나 가설을 설정하는 '본 연구문제 기술' 단계에서 재이용하고자 하는 데이터를 파악 및 습득하기도 하였다. 이때 *"가용할 데이터가 마땅치 않다고 했을 때는 다시 연구관심을 조금씩 수정을 하죠. … 순환이 되어요."*(참여자B1)에서 나타나듯이 데이터에 의해 연구문제가 수정 및 보완되는 상호작용이 이루어지는데 이와 같은 연구문제 재공식화 현상은 McCall & Appelbaum(1991)도 파악한 바 있다. 일반적인 과학적 조사연구 과정과 마찬가지로 '자료수집' 단계에서 재이용하려는 데이터를 파악 및 습득하는 경우도 있었다. *"연구문제가 확실해지면 데이터를 보는 편이에요. … 가설까지 설정해놓고."*(참여자C1)처럼 연구문제에 대한 정확한 가설을 설정하고 연구방법론 및 결과에 대한 이론적 근거를 제시할 수 있는 상태에서 데이터를 파악하거나 습득하

고자 하였다. 그러나 데이터를 파악하더라도 접근 및 이용이 통제된 경우 습득으로 이어질 수 없다는 점에서 일부 심층면담 참여자는 "*공적 접근이 가능한지 … 이용권한이 있는지*"(참여자A1) 확인하는 등 부여된 권한의 범위를 중요시하고, 더 나아가 습득을 위해서 요구되는 절차나 도구 등에 대한 충분한 수준의 메타데이터를 요구하였다.

재이용 데이터의 파악 및 습득 이후에는 '데이터 이해', '데이터 선정', '데이터 재현', '데이터 재구성'이 주요 과정으로 파악되었다. Niu & Hedstrom(2008) 또한 데이터 획득, 데이터 검증, 데이터 조작 및 분석을 통해 데이터가 재이용되며 그와 같은 과정 전반에서 충분한 수준의 도큐멘테이션이 필요하다고 강조한 바 있다.

우선 '데이터 이해'는 변수 등에 해당하는 데이터 자체와 자료수집기간, 조사지역, 분석단위, 조사대상, 자료수집방법, 표본추출방법, 표본크기, 가중치 등과 같은 메타데이터를 이해하는 과정이다. 이를 통해 학술연구자는 데이터가 지니고 있는 잠재적인 오류를 식별할 수 있으며 데이터에 대한 잘못된 해석이나 오용을 방지할 수 있게 된다. 이는 "*데이터셋을 이해를 해야 되니까 … 데이터와 관련된 정보들(을 보죠.)*"(참여자B2), "*변수들을 하나씩 요약을 해보고요. … 결측처리를 해놓았는데 혹시 안 되어 있는 건 없는지 … '모르겠다/무응답'은 어떻게 처리했는지 … 척도는 어떻게 되어있나(를 보죠).*"(참여자C3)를 통해 확인된다.

이와 같이 데이터에 대한 충분한 이해를 바탕으로 학술연구자는 연구목적 및 가설검증에 적합하도록 '데이터 선정'을 하게 된다. 이때 "*데이터 자체의 품질을 검증할 필요가 없는 … 일단 유명한 데이터를 보죠.*"(참여자A1), "*데이터가 얼마나 공신력이 있는가를 파악해야죠.*"(참여자A2), "*제가 보고 싶어 하는 요인들이 최대한 많이 들어가 있는 데이터들을 선택하게 될 테고요.*"(참여자C1)와 같이 학술공동체 내에서 널리 수용된 공신력을 갖춘 신뢰성이 높은 데이터를 재이용함으로써 일정 수준의 데이터 품질을 담보하고자 하였다. 또한 연구목적 및 가설검증에 가장 유용한 주제적 적합성이 높은 데이터를 재이용하고자 하였다.

다음으로 '데이터 재현'은 "*보고서를 보면 … 빈도나 평균이라든지 기본적으로 있잖아요, … 그런 것들이 재현이 되는지를 봐야 되겠죠. (제가) 볼 변수들이 기존 보고서에 있는 전체 샘플을 가지고 (비교했을 때) 차이가 없는지요.*"(참여자C1)처럼 데이터에 대한 기초정보를 담고 있는 집계나 분석 보고서를 참고하여 빈도나 평균 등 기술통계적 사항을 검증하는 과정이다. Niu & Hedstrom(2008)은 해당 과정을 파일검증과 표본검증으로 세분화하였다. 파일검증이란 도큐멘테이션에 명시되어 있는 내용과 실제 파일이 일치하는지를 확인하는 과정으로, 변수가 데이터에 실재하는가나 각 변수에 대한 요약통계가 실제 데이터를 통해 재현되는가 등을 확인한다. 표본검증에는 도큐멘테이션에 설명된 절차대로 데이터가 표집되었는지 점검하고, 결측치가 데이터에 미치는 영향의 정도와 범위를 확인하는 등의 과정이 포함된다.

끝으로 '데이터 재구성'은 학술연구자가 연구목적에 부합하게 데이터를 구성하는 과정으로 "*나에게 필요한 데이터셋으로 재구성*"(참여자B2)하거나 또는 "*데이터를 조작해야죠. … 모델링에 맞게 변수를 구성*"(참여자C2)한

다. Niu & Hedstrom(2008)에 따르면 데이터가 포함하는 변수를 바탕으로 한 새로운 변수 구성, 데이터의 재부호화, 여러 파일의 합성 등이 포함된다. 이후 데이터 분석이 이루어졌는데 이때 *"다른 데이터에서 비슷한 문항이 있다면 당연히 비교를 하게 되죠. … 결과가 얼마나 신뢰성이 있을까? (데이터) 결합은 해석에 있어서 … 삼각법(triangulation) 같은 거죠."*(참여자C1)처럼 연구결과를 검증하기 위한 목적으로 유사한 데이터를 결합한 다음 해석하는 경우도 있었다.

4.3 데이터요구 분석

데이터에 대한 정보요구를 의미하는 데이터요구는 학술연구자가 학술적 목적을 가지고 데이터를 정보원으로 재이용하고자 할 때 발생한다. 학술연구자의 데이터요구를 보다 심층적으로 이해하고자 생산자, 생산국가 및 언어, 유형과 수집방법, 규모, 처리수준, 접근 및 이용 통제, 최신성이라는 7가지 측면에서 살펴보았다.

4.3.1 생산 맥락과 언어

데이터 재이용에 있어 심층면담 참여자 대부분은 개인 단위의 데이터 생산자보다는 기관 단위의 데이터 생산자를 선호하였다. 특히 행정기관이나 정부출연연구기관, 대학부설 전문연구소, 국제기구에서 생산되는 데이터를 재이용하고자 하였다. 이는 해당 생산기관에서 제공되는 데이터가 *"연구분야와 적합한"*(참여자A3) 동시에 *"(해당 기관에서) 밖에 구할 수가 없기"*(참여자B2) 때문이었다. 즉 높은 주제적 적합성을 지닌 데이터가 생산되는 유일한 기관이라는 점이 크게 영향을 미쳤다. 한편 *"접근성이 좋죠. … 기업 데이터는 접근이 불가능하니까요."*(참여자A2)처럼 민간에서 생산되는 기업 데이터에 비해 접근과 이용이 용이하기 때문이기도 하였다. 다음으로 *"문화적 비교라든지 국가 간 비교를 한다면"*(참여자C1)과 같이 비교연구를 수행하는데 있어 적합한 비교가능성이 높은 데이터를 축적하고 있기에 해당 기관을 선호하였다. 또는 선호 기관의 데이터 생산 경험이 장기간 축적됨으로써 데이터 생산과정 및 품질에 대한 신뢰감을 주기도 하였다. 그리고 선호 기관의 명성을 바탕으로 데이터에 대한 신뢰성 및 타당성이 확보되므로 학술 커뮤니티 내에서 수용되기 위해 기울여야 할 부가적인 노력이 적다는 점도 주요한 선호 이유로 파악되었다. 이는 *"정교화 되어있죠, 수십 년 동안 했으니까요. … 정확하죠."*(참여자A2), *"제일 오래되었어요. 체계가 잡힌 것 같아서 선호하고 있어요. … 데이터의 퀄리티도 괜찮은 것 같아요."*(참여자C5), *"모든 사람이 인정하는 거죠. 데이터에 대한 추가적인 설명이 필요가 없어요."*(참여자A2), *"그 쪽(학술지)에서도 어느 정도 신뢰성이나 타당성을 인정해주기 때문에"*(참여자C4)를 통해 확인할 수 있다.

데이터 재이용 의사가 있는 생산국가 및 언어는 다음과 같았다. 우선 선호하는 데이터 생산국가로는 국가 수준에서는 미국, 일본, 중국, 영국, 대륙 수준에서는 유럽이 언급되었다. 정부 간 국제기구인 경제협력개발기구(OECD) 회원국을 특히 선호하는 경우도 있었다. 다음으로 데이터를 표현하는 언어는 영어, 일본어, 중국어가 선호되었다. 이와 같은 생산국가 및 언어에 대한 선호도는 심층면담 참여자의 개인

역량인 외국어 구사능력에 크게 영향을 받았다. 따라서 대부분의 심층면담 참여자가 *"편하게 쓰는 편인"*(참여자B3) 영어가 자국어인 국가나 *"영어로 번역이 잘 되는"*(참여자A1) 언어를 사용하는 대륙이 높은 선호도를 보였다. 특히 미국에서 생산된 데이터에 대한 선호는 개인역량 뿐만 아니라 *"그 쪽(미국) 것을 써야 그 쪽(미국)에서 게재될 확률이 높기 때문에"*(참여자C4)처럼 사회적 요인도 영향을 미쳤다. 최근 우리나라는 개인 학술연구자의 연구 성과를 판단하는 주요 지표로써 SCI급 학술지 게재 논문 수를 택하고 있다. 해당 학술지 상당수가 미국에서 발간된다는 점을 고려한다면 미국에 대한 선호는 우리나라 학술 풍토와 긴밀한 연관이 있어 보인다. 그러나 대륙 수준에서 유럽을 선호한 경우 *"전반적인 트렌드를 … 보는 편인 것 같다."*(참여자C3)는 점에서 재이용 목적상에 있어 차이를 보였다. 아울러 국제기구 회원국 여부가 데이터 선호 기준이 되는 이유는 비회원국은 *"제3국가라고 하면, 아예 수집할 수 있는 인프라 자체도 없는 데가 많아서요."*(참여자C5)와 같이 데이터 생산을 위한 여건이 마련되지 않아 관련 데이터가 존재하지 않는 경우가 대다수이며 간혹 생산된 데이터가 있다 하더라도 단발성에 그치거나 데이터의 품질에 대한 신뢰가 보장되지 않기 때문이었다. 영어가 자국어가 아닌 국가를 선호하는 경우 생산된 데이터가 영어로 제공되는가가 데이터 재이용에 영향을 미쳤다. *"일본자료는 제대로 보지를 못했네요. … 영어로 제공되었으면 잘 검색이 되었을 텐데."*(참여자B3)와 *"일본어를 못하니까, 영어로 제공 안 할 때는 거의 포기하게 되죠. … 언어가 되게 중요하죠."*(참여자C1)에서 드러나듯이 데이터를 검색하거나 이해하는 과정에서 어려움이 발생하였다. 반면 국외 데이터에 대한 요구나 선호가 크지 않은 경우도 있었다. *"뛰어난 데이터 품질을 가지고 있기 때문에"*(참여자B4) 국내 데이터를 국외 데이터보다 선호하거나 혹은 *"제가 관심 있는 … 가족주의는 국제비교가 어려워요. … 한국의 독특한 (것이에요)."*(참여자B1)처럼 주요 연구분야 및 주제가 우리나라의 특수성을 반영하는 경우나 연구수행의 범위를 우리나라에 한하는 경우에는 국내 데이터만을 재이용하는 것으로 나타났다. 위 내용들을 종합해볼 때 국외 데이터 재이용에 있어 보장되어야 할 데이터 속성은 표준화로 나타났다. *"국제 표준으로 되어 있잖아요, … 국제 틀에 매칭 시켜주니까."*(참여자A2), *"OECD에서 표준화를 … 지침 같은 게 있어요."*(참여자C5) 등과 같이 데이터가 표준화되어 있는 경우 각기 다른 국가에서 생산된 데이터를 마치 *"하나의 데이터"*(참여자A2)처럼 재이용할 수 있게 되고 국외 데이터를 재이용 하는 주목적인 국가 및 문화 간 비교가 용이하기 때문이다.

4.3.2 데이터 유형과 수집방법

데이터 유형 및 수집방법과 관련한 데이터 재이용 의사를 살펴본 결과 데이터 유형의 경우 심층면담 참여자 전원이 양적 데이터를 재이용하고자 하였으며 대부분 조사자료를 선호하였다. 그러나 일부 연구기관에 근무하는 심층면담 참여자의 경우 기업 회계장부 데이터(참여자A2)나 행정 데이터(참여자C5)를 선호하기도 하였는데 이는 해당 데이터에 대한 접근가능성이 타 심층면담 참여자에 비해 높기 때문으로

이해할 수 있다.

데이터 수집방법의 경우 조사원 기입식 대인면접 조사와 자동응답시스템(이하 ARS) 전화조사가 선호되었다. 우선 조사원 기입식 대인면접 조사를 선호하는 이유는 *"(면접에) 성실하게 되고"*(참여자A1), *"(응답자가) 설문자의 의도를 정확하게 파악하고 응답을 하는지를 계속적으로 커뮤니케이션하기"*(참여자A2) 때문으로 나타났다. 다음으로 ARS 전화조사를 선호하는 이유는 *"(응답자가) 편하게 대답하기"*(참여자C3) 때문으로 파악되었다. 일부 심층면담 참여자는 해당 데이터 수집 과정에서 조사원의 자질이나 역량이 데이터 품질에 미칠 수 있는 영향력을 *"사실은 (조사하는) 사람이 더 중요하죠. 방법보다도."*(참여자A3), *"자료 조사하시는 분의 태도에 따라서 응답이 되게 많이 달라지잖아요."*(참여자C5)와 같이 강조하였다. 데이터 수집방법은 선호 보다는 비선호를 결정하는데 영향을 미치기도 하였다. 우편이나 온라인을 통하는 경우 *"접근성에 있어 격차가 있기 때문에"*(참여자B1) 또는 수집된 데이터의 *"대표성이 떨어지고 … 심사과정에서 고초를 겪기 때문에"*(참여자C4) 기피되는 양상이 나타났다. 그러나 온라인 및 모바일을 통한 데이터 수집방법과 관련하여 상반된 의견도 있었다. 오늘날 대인면접 조사가 점차 어려워지는 한편 온라인 및 모바일 환경이 급속도로 발달 및 수용되고 있기에 오히려 데이터의 대표성을 확보하는 차원에서 온라인 및 모바일을 통한 데이터 수집방법이 선호될 수 있는 가능성이 드러났다. 이는 *"요즘에는 (데이터 수집을) 웹 링크로 하는 게 점점 늘어나고 (있어요). … 시대적 변화의 문제죠."*(참여자A1), *"앞으로 면대면은 점점 줄어들 수밖에 … 이건 트렌드이기 때문이에요. 앞으로 스마트폰을 사용해서 한다면 오히려 노인 인구들도 적극적으로 참여할 수도 있을 것 같아요."*(참여자C1)에서 확인된다. 추가적으로 확장자에 대한 선호를 살펴본 결과 학술연구자가 재이용 데이터를 통해 수행하려는 통계분석의 수준 및 관련 소프트웨어를 다루는 능력에 영향을 받았다. 심층면담 참여자 대부분은 회귀분석, 요인분석 등과 같은 중・고급 수준의 통계분석을 수행하고자 하였으며 데이터 관련 소프트웨어를 다루는 능력 또한 뛰어났다. 따라서 통계분석 소프트웨어가 가독할 수 있는 확장자인 '.csv'나 '.txt'가 선호되었지만 통계분석 소프트웨어를 통해 *"(확장자) 변환이 가능하므로"*(참여자C2) 영향을 크게 미치지는 못하였다. 선호하는 통계분석 소프트웨어는 STATA, SPSS, SAS, R 순으로 파악되었다.

4.3.3 규모와 처리수준

학술연구자가 재이용 데이터에 대해 통계분석을 수행하는 경우 *"별(astroid)이 잘 뜨려면"*(참여자C5), 즉 통계적으로 유의미한 결과를 얻기 위해서는 데이터 규모가 중요하게 고려되었다. 연구주제나 연구방법론에 따라 상이하지만 *"많으면 많을수록 … 거대한 데이터가 좋죠."*(참여자B4)와 같이 큰 규모의 데이터가 선호되었다. 그러나 *"출간되어있는 데이터는 어느 정도 다 기준을 맞추고 있어요."*(참여자C3)처럼 현재까지 재이용한 데이터의 경우 규모가 충분했기 때문에 재이용 여부를 결정하는데 큰 영향을 미치지는 못한 것으로 나타났다. 한편 데이터 규모로 인해 데이터 재이용에 어려움을

겪은 경우는 "*(데이터의) 전체 규모가 작은 게 아니고 내가 보고 싶은 사이즈를 추출을 하니까 너무 적었기 때문*"(참여자A2)이었다. 이 경우 "*그 자체로 인정을 해주어야 하는데 논문의 심사위원들이 너무 적은 거 아니냐고 (하죠).*"(참여자B4)와 같이 학술 커뮤니티 내에서의 수용 여부가 중요한 영향을 미쳤다.

데이터 처리수준은 데이터 집계 및 분석 정도, 관련 메타데이터 포함 정도, 식별정보 포함 정도를 기준으로 구분되며(Jones, Alexander, Bennett, Bishop, Budden, Cox, & Hardy, 2018), 그에 따른 재이용 의사는 다음과 같았다. 첫 번째, 집계 및 분석 정도에 따라 데이터를 원자료, 표본자료, 분석자료로 구분하였을 때 데이터 재이용 목적에 따라 상이한 선호가 나타났다. 우선 데이터를 연구대상으로 재이용하는 경우 심층면담 참여자 대부분이 원자료를 선호하였다. "*이제 컴퓨팅 파워 자체가 너무 좋아서 표본자료를 분석할 이유가 없어서*"(참여자A1) 혹은 "*할 수 있는 게 제일 많고*"(참여자B2), "*(원자료 이외에는) 핸들링 하는 데에 어려움을 겪을 수 있기*"(참여자C4) 때문에 원자료를 재이용하고자 하였다. 즉 컴퓨터의 연산・처리능력 발전이라는 기술 환경적 변화와 더불어 데이터 분석 과정에 있어 활용가능성 및 편리성을 최대한 확보하고자 원자료가 선호되었다. 한편 "*표본자료가 … 일단 물리적으로도 분석하기에 시간이 덜 걸리는 것도 있고(요).*"(참여자C5)와 같이 경제적인 차원에서 효율성을 제고하고자 분석자료를 선호하는 경우도 하였다. 다음으로 데이터를 연구대상으로 재이용하지는 않지만 데이터 재이용 여부를 결정하거나 데이터 재이용 과정 중에 도움을 얻고자 하는 경우 일정 수준으로 가공된 자료를 선호하였다. "*일단 아이디어가 떠오르면 … 전체적인 아웃라인을 보기 위해 정제된 데이터를 보죠.*"(참여자A2), "*기관에서 집계한 거는 쓸 때도 있죠. 예를 들면 연구를 하다가 특정한 내용에 대한 현황을 대강 보고 싶을 때 (쓰죠.)*"(참여자C3)와 같이 나타났다. 두 번째, 관련 메타데이터 포함 정도는 데이터를 재이용하는 과정에서 중요한 역할을 수행하였다. 이는 "*데이터 자체의 질을 평가하기 위해서 … 대부분의 연구자가 그것(메타데이터)을 보는 이유죠.*"(참여자A1)나 "*그런 자료들(메타데이터)이 주는 정보들에 대한 무게감, 신뢰감 이런 것들도 중요한 것 같아요.*"(참여자C1)와 같이 메타데이터를 통해 데이터의 품질이나 신뢰성을 판단할 수 있고 데이터 재이용 여부에 대한 의사결정 근거가 되기 때문이었다. 이에 재이용 경험이 있는 데이터를 기준으로 메타데이터 포함 정도가 충분하였는지 추가적으로 파악하였다. 그 결과 심층면담 참여자 대부분이 메타데이터가 "*대체로 잘 되어있어서 (데이터 재이용에) 큰 문제가 없었던*"(참여자B2) 것으로 확인되었다. 그러나 "*메타데이터가 충분하지 않은 거 같아요. … 변수들이 어떤 과정을 통해서 클리닝이 되면서 어떤 데이터값은 날아가고 어떻게 보정했다는 정보가요.*"(참여자B1)와 같이 특정 데이터 생산기관에서 제공되는 메타데이터의 경우 보충되어야 할 필요가 제시되었다. 이처럼 메타데이터를 만족스러운 수준으로 제공받지 못한 경우 두 가지 극복 방안을 통해 어려움을 극복하고자 하였다. 우선 메일이나 전화로 데이터 생산자로부터 직접적으로 정보를 획득하였다. 다음으로 "*해당 데이터를 썼던 논문들을 정리해서*"

(참여자C4) 참고하는 등 데이터를 재이용한 선행연구를 살펴봄으로써 정보를 간접적으로 파악하기도 하였다. 이와 같은 직·간접적 극복방안은 McCall & Applebaum(1991)과 Niu(2009)에서도 확인된 바 있다. 아울러 Niu & Hedstrom(2008)의 연구에서 확인된 바와 같이 메타데이터는 충분히 제공되어야 함과 동시에 접근성이 확보되어야 하였다. 접근 상에 어려움이 발생할 경우 *"그걸(메타데이터를) 빨리 못 보면 성급한 마음에 '아휴, 다른 데이터를 찾지 뭐.' 그런 경우도 있고요."*(참여자A1)와 같이 데이터를 재이용하고자 하는 의향을 감소시켰다. 세 번째, 식별정보 포함 정도에 대해서는 「개인정보보호법」이 데이터 재이용과 긴밀한 연관이 있었다. *"개인정보보호법이 강화되면서 데이터 재이용도 상당히 어려워지고 있는"*(참여자A2) 상황이며 *"데이터 활용하는 측면에서 … 개인정보를 어디까지 보호할 거냐"*(참여자C3)를 결정하기 때문이었다. 법적 한도 내에서 인구사회학적인 특성을 파악할 수 있는 식별정보가 포함된 데이터를 선호하였는데 구체적으로 나이, 성별, 직업, 교육수준, 소득수준, 결혼여부, 거주지 등이 언급되었으며 현재 거주지와 관련된 지역 식별정보가 부족한 것으로 파악되었다. 식별정보와 관련하여 어려움이 발생하는 경우 식별정보가 포함된 데이터를 제공하는 데이터센터를 활용하거나 혹은 관련 기관에 필요성을 적극적으로 표현하여 *"(해당 생산 기관에서) 별도 조사를 새로 했어요."*(참여자C2)와 같이 대안을 마련하였다.

4.3.4 접근과 최신성

접근 및 이용이 통제된 데이터에 대한 재이용 의사는 다음과 같았다. 데이터에 대한 접근이나 이용이 어려운 경우 *"그거 아니면 연구가 안 되는 것도 아니고요. … 저는 그냥 (재이용을) 생각 안 할 것 같아요."*(참여자C3)처럼 데이터 재이용을 고려하지 않는 경우가 있었다. 그러나 데이터에 대한 접근이나 이용이 어렵더라도 *"데이터만 좋으면 충분히"*(참여자C4) 데이터 재이용을 고려하는 것으로 나타났다. 이때 금전적 투자는 개인적으로 진행하는 연구보다는 연구프로젝트에서 지원을 받은 경우에 주로 고려되었다. 또한 *"연구관심에 따라서"*(참여자B4) 혹은 데이터가 지닌 *"가치에 따라서"*(참여자C3) 금전적 투자 수준이 결정되었다. 반면에 데이터 재이용을 위한 시간적 투자는 연구의 성격이나 우리나라 학술 풍토에 영향을 받았다. *"한국에서는 … (학술논문이) 대량생산이 많이 되어야 해요. 최신 데이터를 접근 해가지고 빨리 쓰는 게 중요한 풍토가 있는 거죠. … 뭔가 더 트렌드에 대해서 보여주고 그런 게 중요하게 되면 … 참을성 있게 못 기다리지 않을까요?"*(참여자C1)에서 잘 드러난다. 한편 데이터의 소유권이나 저작권이 분명하지 않은 경우에도 *"저자한테 물으면 기관에 (신청)하라고 하고 기관한테 물으면 저자한테 (요청)하라고 하고."*(참여자B3)와 같이 데이터 접근 및 이용에 한계가 있었다. 연구프로젝트를 통해 데이터가 생산되는 경우 연구의 뢰 및 수행 단계에서 데이터 재이용과 관련된 법적 고려가 필요한 것으로 파악되었다.

재이용에 선호되는 데이터 대부분이 기관 차원에서 생산된 조사자료이었으므로 최신성과 관련하여 학술연구자 개인의 선호가 반영될 수 있는 정도에 한계가 있었다. 따라서 조사주기 및 공표주기에 의해 재이용하는 데이터의 최신 정도

가 결정되었다. 그러나 선호 기한을 구체적으로 언급한 경우도 있었는데, 심층면담 참여자 절반이 5년 이내 생산된 데이터를 재이용하고자 하였으며 10년 이내 생산된 데이터에 대한 재이용 의사를 밝힌 경우도 있었다. 이와 같이 특정 기한 내에 생산된 데이터를 선호하는 주된 이유는 "*사회과학은 급변하니까요, 시간이 지나면 쓸모없는 자료도 있고요.*"(참여자B3)나 "*논문을 투고하다가 보면 … (심사의견이) 최신 데이터를 사용하지 않았다는 게 있어요. 그런 것들을 미리 방지하기 위해서 최신 데이터를 쓰죠.*"(참여자C4)에서 나타나듯이 시의성과 학술 커뮤니티 내의 수용 여부로 파악되었다. 또한 "*추세를 보고 싶다고 하면은 굉장히 오래 전부터 보고 싶을 것 같아요. … 최근에 생긴 현상이거나 굳이 과거를 참고할 필요가 없는 현상이라고 하면 최근 것만 봐도 되겠죠.*"(참여자C3)와 같이 연구목적이나 연구분야 및 주제에 따라 선호 기한이 상이하게 나타났다.

5. 논의와 결론

본 연구는 데이터를 재이용한 문헌과 재이용 데이터를 분석하고, 데이터 재이용 경험이 있는 학술연구자를 연구대상으로 한 심층면담을 실시하여 데이터 재이용 현황과 사회과학 연구자의 데이터 재이용과 데이터요구를 파악하고자 하였다. 특히 데이터 재이용 문헌과 재이용 데이터를 주제별로 살펴보았을 때 드러나는 특징이 주목할 만하다. 데이터 재이용 문헌은 대분류 기준 사회과학, 의약학, 복합학 분야, 중분류 기준 사회학, 사회과학일반, 경제학이 높은 비중을 보였다. 재이용 데이터는 대분류 기준 경제, 일반, 사회와문화 영역, 소분류 기준 고용과노동, 일반, 산업과기업 영역에 주로 분포되어있었다. 이러한 경향은 국외 데이터 재이용문헌을 연구대상으로 연구를 진행한 정은경(2018), Lin & Lai(2018)에서도 유사하게 나타난다. 따라서 이는 국내 학술연구자가 지닌 특수한 데이터 재이용 성향이기 보다는 사회과학 분야 내에서 발생하는 보편적인 현상으로 이해할 수 있다.

심층면담 결과를 통해 우선 데이터 재이용 요인 및 유형, 데이터 파악 및 습득 경로, 학술연구 활동 주기 내 데이터 재이용 과정을 살펴봄으로써 데이터 재이용 행위의 특성을 도출하였다. 첫째, 데이터 재이용이 고려되는 요인은 개인적, 경제적, 기술적, 사회적 차원에서 영향을 받았다. 개인적 차원에서는 연구상의 용이성과 데이터 접근성으로 나타났으며 경제적 차원에서는 연구상의 효율성으로 확인되었다. 기술적 차원에서는 정보기술 환경의 변화, 사회적 차원에서는 학술적 풍토의 변화와 연구분야 및 주제의 특성에 의해 데이터가 재이용되었다. 둘째, 데이터 재이용 유형의 경우 재이용의 목적에 따라 구분되었는데 주로 연구대상으로 데이터를 재이용하였다. 그러나 데이터가 배경지식으로서 재이용되는 경우 연구대상으로서도 재이용됨으로써 연속적인 재이용이 이루어졌다. 이러한 점을 고려할 때 원 데이터를 제공하는 동시에 배경지식으로 재이용될 수 있는 높은 신뢰성을 갖춘 집계 및 분석자료를 함께 제공해야 할 필요가 있는 것으로 나타났다. 한편 데이터를 재이용하고자 하는 목적에 따라 요구하는 데이터의 처리수준이나 종수가 상이하게 나타났다. 대부분 두 개 이상의 동종 데이터를

재이용하여 장기간의 변화나 추세, 인과관계를 살펴보거나 국가 간 비교를 통해 특수성 및 동질성을 도출하였다. 따라서 정부기관이나 대학부설연구소 등에서 표준화를 토대로 비교가능성을 갖춘 데이터를 지속적으로 생산한다면 재이용이 활성화되리라 기대된다. 셋째, 데이터를 파악하고 습득하는 주요 경로는 웹 기반의 정보원과 학술공동체 내 비공식적 커뮤니케이션으로 확인되었다. 따라서 생산기관마다 산발적으로 생산 및 공유되고 있는 데이터를 효과적으로 파악하고 습득할 수 있도록 데이터 리포지토리 또는 레지스트리를 구축하여야 한다. 또한 온・오프라인 커뮤니티를 마련하여 데이터 재이용 협업을 촉진할 필요가 있다. 넷째, 학술연구 활동 주기 내에서 데이터 재이용 과정은 일반적인 과학적 조사연구 과정과는 차이를 보였다. 데이터 파악 및 습득이 이루어지는 단계가 개인 연구자마다 상이하였으며 다른 연구단계와 상호작용하며 순환하는 특성을 보였다. 한편 데이터 속성을 기반으로 한 데이터 이해 및 선정, 데이터에 대한 검증, 데이터 재구성은 공통적인 연구단계로 도출되었다.

다음으로 데이터요구는 생산자, 생산국가 및 언어, 유형 및 수집방법, 규모, 처리수준, 접근 및 이용통제, 최신성이라는 7가지 측면에서 분석되었다. 첫째, 데이터의 주제적 적합성, 비교가능성과 신뢰성으로 인해 개인보다는 기관 단위의 생산자가 선호되었다. 둘째, 미국에서 생산되거나 혹은 영어로 기술된 데이터를 선호하였는데 이는 데이터의 이해가능성이 보장되기 때문이었다. 셋째, 대표성, 신뢰성과 같은 데이터의 품질로 인해 양적 데이터 가운데 조사자료 특히 조사원 기입식 대인면접 조사를 통해 생산된 데이터를 선호하였다. 넷째, 활용가능성으로 인해 원자료가 선호되었으며 가능한 풍부한 메타데이터를 얻고자 여러 방식의 노력을 기울이는 것으로 확인되었다. 또한 법적 한도 내에서 인구사회학적 특성을 파악할 수 있는 수준으로 식별정보가 포함된 데이터를 선호하였다. 다섯째, 접근 및 이용이 통제된 데이터는 선호도가 낮았으나 그 가치나 필요도에 따라 일정 수준의 금전적・시간적 투자를 통해 재이용하고자 하였다. 여섯째, 규모와 최신성과 관련해서는 뚜렷한 데이터요구를 확인하기가 어려웠는데 그 이유는 선호가 발생할 수 있을 만큼 다양한 데이터가 제공되고 있지 않았기 때문이다. 이처럼 선택의 폭이 제한된 상태에서 학술연구자는 제공되는 자료를 최대한 재이용하고 있었다. 이러한 연구결과는 향후 데이터 생산, 접근, 이용 활성화를 위한 다양한 정책에 반영될 수 있으리라 기대된다.

참 고 문 헌

정동열, 조찬식 (2018). 문헌정보학조사연구법. 서울: 한국도서관협회.

정은경 (2018). ICPSR 데이터 재이용 저작물 분석을 통한 사회과학 분야의 지적구조 분석. 한국문헌정보학회지, 52(1), 341-357. https://doi.org/10.4275/KSLIS.2018.52.1.341

조재인 (2016). Data Citation Index를 기반으로 한 연구데이터 인용에 관한 연구. 한국문헌정보학회지, 50(1), 189-207. https://doi.org/10.4275/KSLIS.2016.50.1.189

Curty, R. G. (2016). Factors influencing research data reuse in the social sciences: An exploratory study. IJDC, 11(1), 96-117. https://doi.org/10.2218/ijdc.v11i1.401

Curty, R. G., & Qin, J. (2014). Towards a model for research data reuse behavior. Proceedings of the American Society for Information Science and Technology, 51(1), 1-4. https://doi.org/10.1002/meet.2014.14505101072

Faniel, I. M., & Jacobsen, T. E. (2010). Reusing scientific data: How earthquake engineering researchers assess the reusability of colleagues' data. Computer Supported Cooperative Work, 19(3-4), 355-375. https://doi.org/10.1007/s10606-010-9117-8

Faniel, I. M., Kriesberg, A., & Yakel, E. (2012). Data reuse and sensemaking among novice social scientists. Proceedings of the American Society for Information Science and Technology, 49(1), 1-10. https://doi.org/10.1002/meet.14504901068

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. Journal of the Association for Information Science and Technology, 67(6), 1404-1416. https://doi.org/10.1002/asi.23480

Jones, K., Alexander, S. M., Bennett, N., Bishop, L., Budden, A., Cox, M., ... & Hardy, D. (2018). Qualitative data sharing and re-use for socio-environmental systems research: A synthesis of opportunities, challenges, resources and approaches. (SESYNC White Paper) https://doi.org/10.13016/M2WH2DG59

Law, M. (2005). Reduce, reuse, recycle: Issues in the secondary use of research data. IASSIST Quarterly, 29(1), 5-10. https://doi.org/10.29173/iq599

Lin, C. S., & Lai, C. Y. (2018). The reuse of quantitative data in social sciences in Taiwan: 2001-2015. Journal of Educational Media & Library Sciences, 55(1), 39-69. https://doi.org/10.6120/JoEMLS.2018.551/0039.RS.AM

McCall, R. B., & Appelbaum, M. I. (1991). Some issues of conducting secondary analyses. Developmental Psychology, 27(6), 911. https://doi.org/10.1037/0012-1649.27.6.911

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. Journal of Librarianship & Scholarly Communication, 1(1), 1-16. https://doi.org/10.7710/2162-3309.1035

Niu, J. (2009). Overcoming inadequate documentation. Proceedings of the American Society for Information Science and Technology, 46(1), 1-14. https://doi.org/10.1002/meet.2009.145046024

Niu, J., & Hedstrom, M. (2008). Documentation evaluation model for social science data. Proceedings of the American Society for Information Science and Technology, 45(1), 11-11. https://doi.org/10.1002/meet.2008.1450450223

Pasquetto, I. V. (2018). From open data to knowledge production: Biomedical data sharing and unpredictable data reuses. Doctoral dissertation, University of California, Los Angeles, LA, USA.

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. Data Science Journal, 16(8), 1-9. http://doi.org/10.5334/dsj-2017-008

Piwowar, H. A. (2008). Proposed foundations for evaluating data sharing and reuse in the biomedical literature. Bulletin of IEEE Technical Committee on Digital Libraries, 4(2). Retrieved from http://www.ieee-tcdl.org/Bulletin/v4n2/piwowar/piwowar.html

Sun, G., & Khoo, C. S. G. (2017). Social science research data curation: Issues of reuse. Libellarium: Journal for the Research of Writing, Books, and Cultural Heritage Institutions, 9(2). http://doi.org/10.15291/libellarium.v9i2.291

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PloS One, 8(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Wynholds, L. A., Wallis, J. C., Borgman, C. L., Sands, A., & Traweek, S. (2012). Data, data use, and scientific inquiry: Two case studies of data practices. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, Washington, DC. https://doi.org/10.1145/2232817.2232822

Yoon, A. (2014). 'Making a square fit into a circle': Researchers' experiences reusing qualitative data. Proceedings of the American Society for Information Science and Technology, 51(1), 1-4. https://doi.org/10.1002/meet.2014.14505101140

Yoon, A. (2016). Red flags in data: Learning from failed data reuse experiences. In proceedings of the 79th ASIS&T annual meeting: creating knowledge, enhancing lives through information & technology, copenhagen, denmark. https://doi.org/10.1002/pra2.2016.14505301126

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

Yoon, A., & Kim, Y. (2017). Social scientists' data reuse behaviors: Exploring the roles of attitudinal beliefs, attitudes, norms, and data repositories. Library & Information Science Research, 39(3), 224-233. https://doi.org/10.1016/j.lisr.2017.07.008

Zimmerman, A. (2003). Data sharing and secondary use of scientific data: Experiences of ecologists.

Doctoral dissertation, University of Michigan, Ann Arbor, USA.

Zimmerman, A. (2007). Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. International Journal on Digital Libraries, 7(1-2), 5-16. https://doi.org/10.1007/s00799-007-0015-8

[웹사이트]

EvoIO Working Group (2011). Reuse Cases. Retrieved from http://www.evoio.org/wiki/Reuse_Cases

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Jane (2016). Study about research data citation based on DCI(Data Citation Index). Journal of the Korean Society for Library and Information Science, 50(1), 189-207. https://doi.org/10.4275/KSLIS.2016.50.1.189

Chung, EunKyung (2018). An investigation on intellectual structure of social sciences research by analysing the publications of ICPSR data reuse. Journal of the Korean Society for Library and Information Science, 52(1), 341-357. https://doi.org/10.4275/KSLIS.2018.52.1.341

Jeong, Dong Youl, & Cho, Chan-Sik (2018). Research methods in library and information science. Seoul: Korean Library Association.