생명과학분야 학술논문의 전문에 나타난 연구데이터의 공유, 재이용, 인용 분석*

Data Sharing, Reuse, and Citation in Biological Research: An Analysis of Full-Text Literature

박 형 주 (Hyoungjoo Park)**

목 차

초 록

본 연구의 목적은 생명과학 분야 학술논문의 전문(full-text)에 나타난 연구데이터의 공유, 재이용, 인용의 관례을 분석하는 것이다. 연구데이터 공유의 역사가 상대적으로 오래된 생명과학 분야를 대상으로 하였다. 데이터의 수집은 Clarivate Analytics사의 Data Citation Index(DCI)와 Web of Science(WoS)의 연구분야(research area)를 대상으로 하였다. 학술논문의 전문을 분석하기 위하여 연구데이터의 공유, 재이용, 인용을 지시하는 단어 및 구문을 이용하여 반자동 텍스트 검색 기술(semi-automatic text searching techniques)을 활용하고 내용분석(content-analysis)을 실시하였다. 생명과학분야는 공식적인 데이터 인용(formal data citation)이 4.62%로 비공식적인 데이터 인용(informal data citation) 95.38%로 보다 널리 퍼져 있었다. 즉, 생명과학분야에서의 데이터 공유자는 학술논문에서 저자가 받는 학술 크레딧을 연구데이터의 공유에서는 받지 못하고 있음을 확인할 수 있었다. 학술논문에서의 비공식적인 데이터 인용의 위치가 많은 빈도수는 본문, 보충자료, 참고문헌, 감사의글(acknowledgements), 각주, 초록 순이었다. 학술논문에서는 데이터의 재이용(47.8%)이 데이터의 공유(22.1%)보다 2배 이상 많이 출현하였다. 본 연구의 공헌은 데이터 공유, 재이용, 인용의 현상을 학술논문의 전문을 통해서 분석함으로써 연구데이터 인용의 실제적인 현상을 파악했다는 점이다.

ABSTRACT

This study aims to examine data sharing, reuse, and citation practices in academic literature within the biological sciences, a field with a well-established tradition of data sharing. Data were sourced from Clarivate Analytics' Data Citation Index (DCI) and the Web of Science (WoS). Semi-automatic text-searching techniques were used to identify terms and phrases related to data sharing, reuse, and citation, followed by an in-depth content analysis of full texts. The findings reveal that informal data citations (4.62%) are substantially more prevalent than formal data citations (95.38%) in biological science literature. Informal citations most frequently appear in the main text, followed by supplementary materials, references, acknowledgments, footnotes, and abstracts. Data reuse (47.8%) was observed to occur approximately twice as often as data sharing (22.1%) in academic literature. This prevalence of informal data citation suggests that data sharers often do not receive the same academic recognition as bibliographic authors. By analyzing data sharing, reuse, and citation practices in the full texts of academic literature, this study contributes valuable insights to the field of research data citation.

키워드: 연구데이터, 데이터 인용, 데이터 공유, 데이터 재이용, 생명과학
Research Data, Data Citation, Data Sharing, Data Reuse, Biological Sciences

* 이 연구는 충남대학교(교육・연구 및 학생지도비)에 의해 지원되었음.
** 충남대학교 문헌정보학과 조교수(hyoungjoo.park@cnu.ac.kr / ISNI 0000 0004 6442 7767)
논문접수일자: 2024년 10월 27일 최초심사일자: 2024년 11월 5일 게재확정일자: 2024년 11월 15일
한국문헌정보학회지, 58(4): 335-353, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.4.335

1. 서 론

연구데이터는 과학적 연구를 위한 1차 자료로 사용되는 사실적 자료이며 연구 결과를 검증하는데 필수적이므로(OECD, 2015) 새로운 연구의 토대를 마련하는 등 데이터의 보존과 재이용에 도움을 준다. 연구데이터는 학술 연구에서 중요한 역할을 하고 있으며, 주요 출판사와 주요 연구재단들이 새로운 데이터 정책을 도입하면서 그 중요성이 더욱 부각되고 있다. 과학기술정보통신부는 2019년 개정한 대통령령을 통하여 데이터관리계획(data management plan)의 수립을 통하여연구데이터의 체계적인 관리의 중요성을 강조하고 있다(국가연구개발혁신법. 법률 제18645호). 주요 연구재단인 미국 국립과학재단(National Science Foundation), 국립보건원(National Institutes of Health), 국립암센터(National Cancer Institute) 등은 연구데이터 공유를 위하여 일정 금액 이상의 연구제안서에 대하여 데이터 공유 계획서의 제출을 필수로 요구하고 있다. Springer Nature 출판사는 모든 Nature 저널과 약 1,600개의 Springer Nature 저널에 표준화된 연구데이터 공유 정책을 도입하였다(Springer Nature Group, n.d.). 2014년 PLoS저널은 새로운 규정을 시행하여, 연구자가 논문에 사용된 연구데이터를 공개하고 독자가 해당 데이터를 쉽게 찾을 수 있도록 하는 데이터 가용 문구(data availability statement)를 명시하도록 하였다. 데이터 가용 문구의 명시는 학술논문에 기술된 연구데이터의 위치와 데이터 공유와 관련된 제반 사항의 명확한 제시를 요구한다. 연구데이터의 공유를 통하여 데이터 공유자는 학술논문의 서지(bibliographic) 인용처럼 데이터 인용을 통하여 본인의 학술적인 노력에 대한 보상을 받기를 희망할 수 있다. 연구데이터의 공유는 재이용이 가능한 데이터에 대한 접근성을 높이는 필수적인 조건이며, 이를 통해 데이터 재이용의 활성화를 촉진하고, 연구의 완전성, 재현성 및 투명성을 향상시킬 수 있다(Curty et al., 2017). 연구의 초기 단계에서 연구자들은 기존의 연구데이터를 재이용할지 아니면 새로운 연구데이터를 직접 수집할지를 결정해야 한다. 연구데이터의 재이용은 이미 수집되거나 생성된 데이터를 원래의 목적 외의 새로운 연구와 분석 또는 응용에서 재이용하는 것이다. 데이터 재이용은 중복된 데이터의 수집을 줄이고, 연구의 효율성을 높이고, 연구 비용의 절감에 기여할 수 있다(Borgman, 2015). 최근 연구자들은 연구데이터를 재이용할 때의 이점을 더욱 인식하고 있으며, 이에 대해 긍정적인 평가를 하고 있다(Tenopir et al., 2020). 하지만, Tenopir et al.은 데이터 공유의 주요 장벽 중 하나는 연구 성과가 감소할 것이라는 우려라고 하였다. 데이터 공유는 학술논문 출판의 가시성을 높여 연구자들에게 추가적인 크레딧을 받을 수 있도록 하지만(Piwowar et al., 2007), 연구자들은 일반적으로 학술논문 출판보다 데이터 공유에서 받는 보상이 적다고 인식한다(Steel et al., 2019). 연구데이터 재이용의 활성화를 위하여 재이용 관례 및 데이터에 대한 심도있는 이해가 필요하다. 하지만 실제 연구데이터 재이용의 경우, 데이터 인용이 데이터 재이용의 중요한 지표임에도 불구하고, 저자들이 재이용된 데이터를 공식적으로 인용하지 않는 경우가 많아 이러한 데이터 재이용을 정확히 측정하기는 쉽지 않다(Park & Wolfram, 2017). 데

이터 인용은 연구에 활용된 데이터의 출처를 명시하고, 해당 데이터를 학술적으로 인용하는 것이다(Altman & King, 2007). 데이터 인용을 통하여 연구에 활용된 데이터에 대한 정확한 정보 및 출처를 알 수 있고, 데이터를 추적 및 색인하고 검증할 수 있다. 데이터 인용을 통하여 연구 과정에서 사용된 데이터에 대한 투명성이 강화되고, 데이터의 공유 및 재이용의 관례를 살펴볼 수 있으며, 데이터 공유자는 학문적 공로를 인정받을 수 있다. 하지만 공식적인 데이터 인용(formal data citation)은 학술커뮤니케이션의 관례가 아니다(Park et al., 2018). 학술논문에서의 데이터 공유, 재이용, 인용의 관례는 활발하게 일어나지 않으므로(Park & Wolfram, 2017; Park et al., 2018) 공유 역사가 긴 학문 분야를 별도로 연구하면 연구데이터의 관례를 이해하는 데 도움이 된다. 생명과학분야는 다른 학문 분야보다 연구데이터 공유 역사가 길어서 데이터 공유가 활성화 되어있으므로 연구데이터의 관행을 심도있게 포괄적으로 살펴볼 수 있다. 미국 국립보건원은 2003년부터 일정 금액 이상의 연구제안서를 제출할 때 데이터 공유 계획서를 필수로 요구(National Institutes of Health, 2003)해온데 반해, 국립과학재단은 2011년부터 데이터 공유 계획서를 요구했으므로(National Science Foundation, 2011), 생명과학분야의 연구데이터의 필수적인 공유 역사가 다른 과학 분야보다 약 10여년 더 빠르기 때문이다. 연구 질문은 다음과 같다.

- 연구 질문: 생명과학 분야에서 학술논문의 전문에 나타난 데이터 공유, 재이용, 인용 관례는 어떠한가?

2. 선행연구

선행연구는 데이터 공유, 데이터 재이용, 데이터 인용으로 구성되어 있다.

2.1 데이터 공유

데이터 공유는 연구자나 기관이 자발적으로 또는 기관의 규범에 따라 원시 또는 사전 처리된 데이터 혹은 1차 연구 데이터를 공개하는 행위를 의미한다(Curty, 2015). 연구데이터는 데이터 생애주기동안 생성, 수집, 처리된 후 데이터 리포지토리에 보관되어 공유 및 재이용이 가능하다. 한 연구에 따르면, 데이터의 공유는 대중의 신뢰를 증진하는 데 기여하며, 미국 성인의 57%가 연구자들이 데이터를 공유할 경우 연구 결과에 대한 신뢰를 보일 것이라는 결과가 보고되었다(Funk et al., 2019). 데이터 공유자는 파일 형식의 지속성과 시간이 경과되어도 호환 가능성을 보장해야 하는데, 이는 운영체제와 기술이 정기적으로 변화하기 때문일 수 있다. 많은 연구자가 서로 다른 운영 체제와 기술을 활용하여 연구 데이터를 저장하고 있다(Corti et al., 2014).

데이터 공유는 더 많은 연구자가 공유된 데이터를 통해 연구자로서 보상을 받을 수 있는 기회를 제공한다. 선행연구에 따르면, 연구자들은 연구 데이터를 학술 저널에 공유하기보다는 보류하는 경향이 있는 것으로 나타났다(Boulton et al., 2012; Cohen, 1995; Piwowar, 2011). 연구자는 데이터 공유에 대한 보상이 낮거나 전혀 없다고 인식할 경우 데이터를 공유하지 않는 경향이 있으나(Sterling & Weinkam, 1990),

연구자의 인식과 보상 체계는 데이터 공유 행동을 촉진하는 중요한 요소가 된다(Kling & Spector, 2003). Tenopir et al.(2011)의 연구에 따르면, 생물학 분야 연구자의 49%는 데이터 공유에 동의하거나 어느 정도 동의하며, 생물학이 사회과학보다 데이터 공유에 대한 동의 비율이 약 두 배 높은 것으로 나타났다.

2.2 데이터 재이용

데이터 재이용은 과학자들이 기존의 연구데이터를 활용하여 데이터 리포지토리에서 얻거나, 이메일 등의 개인적인 의사소통 채널을 통해 획득한 다른 기존의 연구데이터나 새로 수집한 연구데이터를 결합하여 이전의 연구 결과를 재현하는 과정이다(King, 1995). 데이터 재이용은 원본 데이터를 새로운 연구 문제를 해결하기 위해 2차적으로 활용하는 제반 활동이며, 일반적으로 기존 데이터의 2차 분석으로 설명되는 다양한 차원과 사례를 포함한다(Curty, 2015). 예를 들어, GenBank, Sequence Read Archive(SRA), UniProtKB/TrEMBL에 기탁된 연구데이터의 수는 시간이 지남에 따라 기하급수적으로 증가하고 있으며(Sielemann et al., 2020), 이러한 대규모의 데이터베이스를 유지하고 업그레이드하는 데는 많은 노력이 필요하므로 데이터 재이용을 통해 중복을 줄이는 것이 유익할 수 있다. 연구자들은 공개된 데이터의 재이용을 긍정적으로 평가한다(Tenopir et al., 2020).

데이터 재이용의 주요 도전 과제는, 연구자가 서로 다른 출처에서 얻은 데이터세트를 비교하고 통합하는 데 어려움이 있을 수 있다는 점이다(Pasquetto et al., 2017). 데이터의 재이용을 위해서는 재이용할 데이터가 신뢰할 수 있고 재이용하기에 적합한 품질을 가지고 있어야 하기 때문이다. 예를 들어, 메타게놈학에서 독립적으로 수행된 연구를 비교할 때, 워크플로우의 차이, 기록되지 않은 변수, 통일되지 않은 표현 형식 등의 문제(Ten Hoopen et al., 2017)는 데이터 재이용의 도전 과제이다. Sielemann et al.(2020)은 다른 데이터베이스에서 온 데이터세트 간의 유효한 비교나 데이터베이스 통합을 위해서는 연구데이터의 유형 및 분야에 적합한 특정 조건을 충족해야 한다고 하였다. 공개된 데이터를 재분석하는 등의 데이터 재이용은 연구데이터의 품질, 통합, 비정규화 등의 문제를 초래할 수 있으며, 이러한 문제들은 큐레이션과 자가 수정으로 해결될 수 있지만 실질적인 시행은 어려울 수 있다. 또한, 연구데이터의 동료심사(peer review)는 아직은 학술 커뮤니케이션의 관례가 아니다. 동료심사를 받은 연구데이터가 통제된 환경에서 재이용될 경우, 데이터 재이용의 위험이 일부 완화될 수 있다(Spertus, 2012). 데이터 재이용에 대한 연구자의 인식 및 경험은 인프라스트럭처, 학문 분야, 사회적 인식, 요구사항 등의 다양한 요인에 따라 다르게 나타난다(Faniel & Zimmerman, 2011).

학술논문에서의 데이터 재이용을 자동으로 측정하는 것은 어려움이 있다. 연구데이터의 공유 및 재이용은 여러 도전 과제에 직면해 있으며, 연구데이터의 재이용은 간소화될 필요가 있다. 연구자는 학술논문을 살펴본 후 재이용된 데이터를 식별하기 위하여 학술논문을 하나하나 읽는 수작업을 수행해야 하며, 이는 연구데이터의 재이용을 방해하는 주요 요인 중 하

나일 수 있다. 연구데이터의 재이용을 위해서는 데이터를 보존하는 문서화와 문맥(context)에 대한 이해가 필수적이다(Faniel et al., 2019). 또다른 도전 과제는, 연구자들이 학술논문에서 재이용된 연구데이터를 식별하기 위해서는 학술논문을 수동으로 하나하나 읽어야 한다는 점이다. 예를 들어, 생물의학 분야에서는 자연어 처리 기술(natural language processing)을 활용하여 초록을 분석한 연구가 있었는데, 이는 초록이 대체로 표준화된 양식을 가지고 있기 때문에 가능하였다(Lin, 2009). 연구데이터의 공유 및 재이용을 지시하는 용어를 사용하여 반자동 텍스트 검색 기술을 활용하여 연구데이터의 재이용을 탐색한 연구도 있었다(Park et al., 2018). 그러나 데이터 재이용은 주로 학술논문의 연구방법론 섹션에 출현하는 경향이 있으므로, 자동으로 연구데이터의 인용을 색인하는 것은 어려움이 있다(Park & Wolfram, 2017). 현재로서는 연구데이터의 재이용을 측정하기 위한 적절한 자동화 도구가 부족한 상황이다.

2.3 데이터 인용

데이터 인용은 학술연구에서 활용된 데이터에 대한 출처를 명시하고, 해당 데이터를 학술적으로 인용하는 것이다(Altman & King, 2007). Altman, King은 데이터 인용의 구성요소는 일반적으로 저자, 제목, 출판일자, 고유식별자(identifier) 등이 있다고 하였다. 데이터 인용은 연구데이터에 대한 공로를 인정하고 과학실험과 발견의 재현성을 촉진하기 위해 필수적이다(Piwowar, 2011; Silvello, 2018). 데이터 인용은 데이터 생성자(data creator)의 공로를 인정하며, 저자에게 데이터 공유에 대한 인센티브를 제공할 수 있다. 그러나 많은 과학자들은 데이터 인용이 데이터 공유에 강력한 인센티브를 제공함에도 불구하고, 데이터 공유에서 충분한 인정과 보상을 받지 못한다고 주장한다(Dorta-González et al., 2021). 하지만, 학술논문과 연구데이터를 함께 제공하는 학술 논문은, 서지 인용 횟수가 25% 증가하였다(Colavizza et al., 2020). 대다수의 연구자들은 오픈액세스를 통해 자신의 학술논문과 데이터를 공유할 의향이 있다고 보고하고 있다(Tenopir et al., 2020).

데이터 인용은 학술논문과 관련된 연구데이터에 대한 인용을 의미하며, 이는 학술 커뮤니케이션에서 중요한 요소로 자리 잡고 있다. 전통적으로 학술 커뮤니케이션에서의 보상 시스템은 동료 심사를 거친 학술지와 그 학술지에 출판된 논문의 영향력에 기반해 왔다. 그러나 최근에는 연구데이터와 같은 비전통적인 과학적 작업물도 주요 연구 재단과 영향력 있는 저널에 의해 출판물로서 인정받고 있다. 미국 국립과학재단, 국립보건원, 그리고 네이처(Nature)와 PLoS 저널 등이 그 예이다. 비록 데이터 공유가 자동으로 데이터 인용으로 이어지지는 않지만(Silvello, 2018), 데이터는 인용 가능하며 인용되어야 한다(Lawrence et al., 2011). 적절한 데이터 인용은 데이터 공유와 재이용을 촉진하기 위한 중요한 인센티브를 제공한다(Mooney & Newton, 2012). 데이터 인용은 출판된 데이터를 접근, 위치 확인, 인증, 식별, 해석하여 출처를 명확히 하고 기여를 인정하는 역할을 하는 것으로 확인되었다(CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013).

전통적인 서지 인용 절차는 데이터 인용에 그대로 적용하기 어려우며, 새로운 방법론과 솔루션이 필요하다. 데이터 공유는 인용 이점을 제공한다(Piwowar & Chapman, 2010; Piwowar et al., 2007). 예를 들어 천체물리학(Drachen et al., 2016), 유전자 발현 연구(Piwowar & Vision, 2013) 등에서 데이터 공유로부터 더 높은 학술논문의 피인용 횟수를 얻는데 기여하였다. 또한, 이미지 처리 논문에서는 계산 코드 공유와 인용 수 간의 긍정적인 연관이 발견되었다(Vandewalle, 2012). Drachen et al.(2016)은 천체물리학 분야의 주요 저널에서 데이터와 연결된 학술논문의 인용이 25% 증가했음을 확인하였다.

데이터 인용에 있어 학문 분야 간 차이도 중요하다. 다양한 학술 커뮤니티는 데이터 공유, 재이용, 관리에 대한 태도가 다르며, 이에 따라 연구데이터의 크기와 내용도 분야별로 다르다(Digital Curation Centre, 2010; Palmer & Cragin, 2008). 천문학, 유전체학, 지구과학과 같은 분야에서는 데이터 공유가 일반적이며, 공유된 데이터를 재이용하는 연구가 증가하고 있다(Pierce et al., 2019; Tenopir et al., 2020). 다른 연구 분야는 고유한 데이터 요구를 가지고 있으며, 이러한 차이로 인하여 데이터 인용 관례도 달라질 수 있다.

Data Citation Index(DCI)는 연구데이터를 추적하고 색인할 수 있는 학술 데이터베이스로서 연구데이터의 공유, 재이용, 인용을 측정할 수 있도록 돕는다. DCI는 2012년 Thomson Reuters에 의해 출시된 이후, 2016년에 Clarivate Analytics사에 인수되었으며, 현재는 구독(subscription)에 기반한 서비스를 제공하고 있다. DCI는 2024년 6월 현재 전 세계 약 450개의 데이터 리포지토리에서 1,490만개 이상의 연구데이터, 164만개 이상의 데이터 스터디(data studies), 49만개 이상의 소프트웨어(software)를 추적하고 있다(Clarivate Analytics, 2024). DCI는 WoS가 학술논문, 학술발표집, 도서 등을 색인하는 것과 유사한 형태로 데이터세트를 추적하고 색인한다. DCI 데이터의 인용 레코드는 WoS의 연관 문헌과 연결되어 색인되고 있어서, 연구데이터의 인용을 직접적으로 색인할 수 있다. DCI는 다양한 학문 분야에서 연구데이터의 인용을 추적하는 데 사용되었으며, 유전학 및 인문학 등에서 연구데이터가 어떻게 인용되는지를 분석하는 데 유용하다(Park & Wolfram, 2017; Robinson-García et al., 2016). DCI를 분석하는 것은 연구데이터의 관례을 탐색하기에 적절하다. DCI는 WoS의 학술논문과 연관데이터(associated data)로 추적되고 있으므로, 연구데이터의 공유 및 재이용 관례 및 자기 인용(self-citation) 관례를 살펴보기에 적절하다. 조재인(2016)은 2006년부터 2015년까지의 DCI 연구데이터 중 인용 빈도 상위 500위 데이터를 분석하여, DCI 데이터의 주요 주제 분야와 주요 데이터 유형을 분석하였다.

선행연구를 종합하면, 데이터 공유, 재이용, 인용의 관례를 데이터 레벨(data-level)에서 분석한 연구는 활발히 수행되었지만, 학술논문 레벨(article-level)에서의 데이터 공유, 재이용, 인용의 관례를 분석한 연구는 활발히 진행되지 않았음을 확인할 수 있었다. 따라서 본 연구는 학술논문의 전문에 나타난 데이터 공유, 재이용, 인용을 분석하고자 한다.

3. 연구방법론

3.1 데이터 수집

본 연구는 Clarivate Analytics사의 DCI를 연구데이터의 공유, 재이용, 인용을 식별하기 위한 출발점으로 활용하였다. DCI의 기능(feature)을 사용하면 전 세계 약 450여개의 데이터 리포지토리에서 수집된 데이터세트, 데이터 연구, 소프트웨어의 인용 이력을 단일 접근점(single access point)에서 확인할 수 있다. DCI의 데이터 인용 정보에는 저자, 제목, 출판 연도, 출판사, 판 또는 버전, 인용 이력, 접근 정보(예: URL 또는 DOI)와 같은 구성 요소가 있었다.

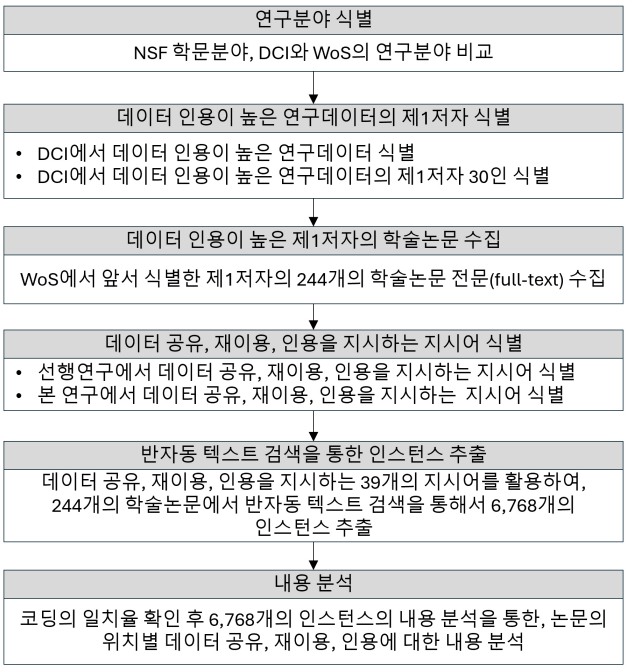
본 연구는 생명과학 분야를 대상으로 한정하였는데, 생명과학 분야가 데이터 공유 역사가 더 길기 때문이었다. 미국 국립과학재단의 데이터공유계획서 제출 요구가 2011년부터 시작되었고(National Science Foundation, 2011), 미국 국립보건원은 2003년부터 연구데이터의 공유계획서의 제출을 요구했으므로(National Institutes of Health, 2003), 생명과학분야가 다른 STEM(Science, Technology, Engineering, Mathematics) 학문분야보다 약 10년 정도 데이터 공유의 역사가 길기 때문이었다. 즉, 생명과학 분야를 연구하면 데이터 공유, 재이용, 인용의 관례를 다른 학문 분야보다 더 포괄적으로 살펴볼 수 있다.

〈표 1〉은 본 연구에서 활용한 미국국립과학재단(National Science Foundation)의 주요 학문분야(major discipline)와 WoS All Collections 및 DCI의 연구분야(research area)를 보여준다. 본 연구는 WoS와 DCI의 연구분야를 분석의 대상으로 하였다. 다학제 분야(interdisciplinary)는 범위가 광범위하고 특정 학문 분야에 포함되기 어려우므로 다학제 분야는 데이터 수집 대상에서 제외하였다. 연구 모집단은 DCI에서 인용된 연구데이터와 WoS의 학술논문들로 설정하였다. WoS의 학술 데이터베이스에는, WoS Core Collection(WoS 핵심 컬렉션), DCI(데이터 인용 색인), Biosis Citation Index(생명과학 인용 색인), Scientific Electronic Library Online Citation Index(과학 전자 도서관 온라인 인용 색인), Chinese Science Citation Database(중국 과학 인용 색인), Russian Science Citation Index(러시아 과학 인용 색인)의 데이터베이스가 포함되어 있었다.

〈그림 1〉은 본 연구에서 활용한 연구 방법론

〈표 1〉 미국 국립과학재단의 주요 학문분야와 WoS 및 DCI의 연구분야 비교

미국 국립과학재단의 주요 학문분야	WoS All Collections 및 DCI의 연구분야
생명 과학	Genetics and Heredity(유전자 및 유전), Biochemistry & Molecular Biology(생화학 및 분자생물학), Biotechnology & Applied Microbiology(생명공학 및 응용미생물학), Cell Biology(세포생물학), Developmental Biology(발생생물학), Evolutionary Biology(진화생물학), Marine & Freshwater Biology(해양 및 담수생물학), Mathematical & Computational Biology(수리 및 계산생물학), Microbiology(미생물학), Plant Sciences(식물과학), Reproductive Biology(생식식물학), Environmental Sciences & Ecology(환경과학 및 생태학), Biodiversity & Conservation(생물다양성 및 보전), Research & Experimental Medicine(연구 및 실험의학)



〈그림 1〉 연구 방법론 도식화

을 도식화한 것이다.

DCI의 생명과학분야에서의 데이터인용이 높은 저자들의 학술논문의 표본 추출을 위하여 다음의 방법을 활용하였다. 이 방법론을 활용한 이유는, 학술논문에서의 데이터 공유, 재이용, 인용의 관례는 아직 활발하지 않으므로(Park & Wolfram, 2017; Park et al., 2018), 데이터 공유, 재이용 인용 관례를 최대한 많이 식별하기 위함이었다. 첫째, DCI에서의 데이터인용 횟수가 가장 높은 30인의 저자를 식별하였다. 〈표 1〉에서 식별한 DCI의 생명과학분야 연구분야를 기준으로 DCI에서 데이터인용 횟수가 가장 높은 순으로 정렬을 하였다. 구체적으로 DCI의 'most highly cited' 기능을 활용하였다. 이를 통하여 데이터인용 횟수가 가장 높은 상위 30개의 연구데이터의 제1저자를 식별할 수 있었다. 30개의 연구데이터 중에서 제1저자가 동일한 저자가 발견된 경우, 다음 순위 연구데이터의 제1저자를 식별하였다. 둘째, 식별된 30명의 제1저자의 이름을 Web of Science 데이터베이스를 검색하여 저자들의 학술 출판물에 대한 정보를 획득하였다. 해당 저자의 모든 학술 출판물과 관련된 인용 정보를 다운로드 받은 후 마이크로소프트 엑셀에 저장하였다. 체계적 무작위 표본 추출(systematic random sampling) 기법을 활용하여, 각 저자의 10번째 학술논문을 식별한 후, 학술논문의 전문(full text)을 수집하여 다운로드 받았다. 구체적으로 1번째, 11번째, 21번째와 같이 10번째 학술논문의 순서로 논문의 전문을 수집하였다. 체계적 무작위 표본 추출을 실시한 이유는, 단순 무작위 표본 추출(simple random sampling)을 실시할 경우,

학술논문의 수집이 1번째, 11번째, 12번째, 13번째 논문처럼 편향되게 수집될 가능성을 배제하기 위해서였다 최종적으로 총 244개의 학술논문이 수집되었다.

3.2 데이터 분석

데이터의 분석을 위하여 반자동 텍스트 검색 기법(semi-automatic text searching techniques)을 활용하였다. 본 연구는 244개의 학술논문을 분석하였으므로 사람이 수동으로 일일이 244개의 학술논문을 읽어서 데이터 공유, 재이용, 인용을 식별하기가 쉽지 않고, 수동으로 읽을 때의 사람의 피로도로 인하여 정확한 식별이 어려울 수 있고, 많은 시간이 소요될 수 있기 때문이었다. 보다 정확하고 신속하게 연구데이터의 관례를 식별하기 위하여, 데이터 공유, 재이용, 인용을 지시하는 용어 및 구문(indicating terms and phrases)을 식별한 후, 지시 용어 및 구문이 출현하는 전문을 기반으로 반자동 텍스트 검색을 실시하였다. 이를 통하여 데이터 공유, 재이용, 인용의 사례를 식별하고 수집할 수 있었다.

데이터 공유, 재이용, 인용을 지시하는 용어를 식별하기 위하여 다음의 방법을 활용하였다. Park, Wolfram (2017), Park et al.(2018)이 식별한 지시어를 바탕으로 하였다. Park, Wolfram (2017)은 DCI에서 데이터 인용이 가장 활발한 분야인 유전자학(Genetics and Heredity)에서 데이터 공유, 재이용을 지시하는 지시어를 식별하였고, Park et al.(2018)은 생명의학 분야에서 데이터 공유, 재이용, 인용을 지시하는 지시어를 식별하였다. 본 연구는 생명과학 분야의 연구이므로 기존 연구의 지시어를 활용은 하였으나, 누락되거나 새롭게 발견된 지시 용어를 추가로 식별하기 위하여 총 5개의 생명과학 분야 학술논문을 추가로 검토하였다. 본 연구는 5개의 생명과학분야의 학술 논문을 샘플로 식별한 후, 해당 학술논문을 사람이 읽어서 논문에 나타난 데이터 공유, 재이용, 인용 현상을 지시하는 지시용어 및 지시 구문을 추가로 식별하였다. 5개의 학술논문은 앞서 식별된 학술논문에서 무작위 추출법(random sampling)으로 식별하였다. 〈표 2〉는 본 연구에서 활용한 데이터 공유, 재이용, 인용을 지시하는 지시어 및 지시 구문의 목록이다. 지시어는 데이터 공유, 재이용, 인용을 구분하지 않았는데, 동일한 지시어가 동시에 데이터 공유와 데이터 재이용을 포함하고 있는 경우가 있기 때문이었다. 'data(데이터)'라는 용어도 지시어로 식별되었으나, 데이터라는 용어는 학술논문에서 데이터 공유 및 재이용을 식별하는 데 좋은 지시어라기 보다는 학술 논문에서 일반적으로 사용되는 단어이므로 지시어에서 제외하였다.

〈표 3〉은 본 연구에서 활용한 생명과학분야 논문의 총 수와 데이터 공유, 재이용, 인용을 지시하는 학술논문의 문장의 총 인스턴스 수를 보여준다. 각 인스턴스는 〈표 2〉의 지시어/구문이 학술논문의 전문(full-text)에서 출현한 경우, R 프로그래밍 언어를 활용하여 250자의 텍스트를 인스턴스로 추출하였다. 이 후 모든 인스턴스를 마이크로소프트 엑셀에 저장하였다.

데이터 분석을 위하여 두 명의 코더(coder)가 코딩에 참여하였다. 코딩에 경험이 있는 두 명의 코더는, 전체 인스턴스의 10%에 해당하는 677개의 인스턴스를 내용분석(content analysis)

〈표 2〉 생명과학 분야에서 데이터 공유, 재이용, 인용을 지시하는 용어 및 구문

생명과학 분야에서 연구데이터의 공유, 재이용, 인용을 지시하는 용어 및 구문		
.com	data sets (데이터 세트)	obtained from (~로부터 획득된)
.edu	database (데이터베이스)	project website (프로젝트 웹사이트)
.gov	dataset (데이터 세트)	provided by (~에 의해 제공된)
.org	deposited (기탁된)	publicly available (공개적으로 가능한)
accession (접근)	donated by (~에 의해 기부된)	purchased from (~로부터 구매된)
acquire (획득하다)	donated from (~로부터 기부된)	repository (리포지토리)
available (가능한)	downloaded (다운로드된)	repository numbers (리포지토리 숫자)
benefited (이점의)	ftp://	samples (샘플)
browser (브라우저)	gift (선물)	stored (보관된)
commercial (상업의)	Inc. (회사)	suppl (보충의)
Corp. (회사)	National Institutes of Health (국립 보건원)	supplemental (보충의)
data availability (데이터 가용)	NIH (국립 보건원)	supplemental material (보충 자료)
data center (데이터 센터)	NOAA (해양 대기청)	survey (설문조사)

〈표 3〉 분석에 활용된 생명과학분야의 논문 및 인스턴스

학문 분야	분석된 논문의 총 수	학술논문에 나타난 총 인스턴스 수
생명과학	244개	6,768개

을 통하여 데이터 공유, 재이용, 인용을 코더 1과 코더 2가 개별로 코딩하였다. 이 후 코더 1과 코더 2 두 명의 코딩의 결과를 비교한 결과, 코딩의 일치율(agreement rate)이 95.2%를 달성하였음을 확인하였다. 95.2%의 일치율은 나머지 인스턴스를 코딩하기에 신뢰도가 높은 점수이므로 남은 인스턴스를 코더 1이 코딩하였다.

4. 결 과

〈표 4〉는 본 연구에서 분석한 생명과학분야 학술논문에 출현한 데이터 공유, 재이용, 인용을 지시하는 지시어/구문의 출현 빈도수 및 퍼센트를 보여준다. 퍼센트는 전체 출현 인스턴스인 6,768개의 퍼센트이다. 출현 빈도수 상위 10위는 samples(1,425회), available(970회), database(623회), suppl(531회), .org(513회), dataset(375회), NIH(314회), National Institutes of Health(212회), obtained from(204회), data sets(191회)이며, 데이터 공유, 재이용, 인용을 식별하는 재현률 및 정확률이 우수한 지시어/구문이라고 할 수 있다.

〈표 5〉는 생명과학 분야의 공식적인 데이터 인용(formal data citation)과 비공식적인 데이터 인용(informal data citation)을 비교 분석한 표이다. 공식적인 데이터 인용은 WoS에 색인된 학술논문의 참고문헌 섹션에 데이터의 인용이 색인되는 경우이다. 본 연구는 생명과학 분야에서 비공식적인 데이터 인용이 95.38%로 공식적인 데이터 인용인 4.62%보다 널리 퍼져 있음을 확인하였다. 데이터 공유자는 자신의

〈표 4〉 생명과학 분야에서 데이터 공유, 재이용, 인용을 지시하는 용어 및 구문

순위	지시어/구문	빈도수	퍼센트
1	samples	1,425	21.05
2	available	970	14.33
3	database	623	9.2
4	suppl	531	7.84
5	.org	513	7.58
6	dataset	375	5.54
7	NIH	314	4.64
8	National Institutes of Health	212	3.13
9	obtained from	204	3.01
10	data sets	191	2.82
11	provided by	151	2.23
12	.com	146	2.16
13	.edu	144	2.13
14	Inc.	124	1.83
15	.gov	106	1.57
16	acquire	100	1.48
17	stored	88	1.3
18	survey	87	1.29
19	accession	61	0.9
20	publicly available	56	0.83
총합		6,421	94.86

〈표 5〉 공식적인 데이터 인용과 비공식적인 데이터 인용의 비교

데이터 인용의 종류	합계(퍼센트)
공식적인 데이터 인용	65 (4.62%)
비공식적인 데이터 인용	1,342 (95.38%)
총합	1,407 (100%)

데이터가 참고문헌 섹션에 나타날 경우 공식적인 인용으로 학술적 인정을 받을 가능성이 더 크다. 비공식 인용은 공유된 데이터가 감사의 글(acknowledgment)이나, 논문의 본문에 언급될 때 발생한다(Cronin, 1995; 2001). WoS와 같은 인용 데이터베이스는 위치 때문에 비공식적인 인용인 본문에 언급만 하고 참고문헌에는 인용이 있지 않은 연구데이터는 색인하지 않는다. 일부 학문 분야에서 공식 데이터 인용률이 낮은 이유는 연구자들이 데이터 공유로부터 인정을 받지 못하기 때문이다(Dorta -González et al., 2021).

〈표 6〉은 본 연구에서 식별한 학술논문에 나타난 데이터 공유, 재이용, 인용의 학술논문에서의 위치를 보여준다. 학술논문에서의 위치의 총 수는 내림차순으로 본문(main text)은 1,064건(75.62%), 보충자료(supplementary materials)는 188건(13.36%), 참고문헌은 65건(4.62%),

〈표 6〉 데이터 인용의 종류와 학술논문에서의 출현위치 비교

데이터 인용의 종류	학술논문에서의 출현위치	합계 (퍼센트)
공식적인 데이터 인용	참고문헌	65 (4.62%)
비공식적인 데이터 인용	각주	31 (2.2%)
	감사의글	46 (3.27%)
	본문	1,064 (75.62%)
	보충자료	188 (13.36%)
	초록	13 (0.92%)
총합		1,407 (100%)

감사의글(acknowledgements)은 46건(3.27%), 각주(footnotes)는 31건(2.2%), 초록은 13건(0.92%)이었다. 즉, 비공식적인 데이터 인용(95.38%)이 널리 퍼져있고, 대부분의 비공식적인 인용은 본문(main text)에서 주로 발견되었으며, 특히 방법론(research methods) 섹션에서 주로 발견되었다. 방법론 섹션에서는 데이터를 어떻게 수집하였는지에 대한 내용이 적혀 있었는데, 정확한 디지털 객체 식별자(Digital Object Identifier; DOI)나 URL(Uniform Resource Locator)이 적혀있지 않았다. 예를 들어, 본문의 방법론 섹션에 "Zenodo 데이터 리포지토리에서 다운로드 받았다"와 같은 표현으로, DOI 등의 연구데이터의 고유 식별자가 아닌, 데이터 리포지토리 전체를 언급만 하고 지나가고 있었다. 학술논문의 사례와 비교를 하자면, 저자가 자신이 참고한 학술논문을 인용하는 것이 아니라, 학술지 이름을 인용하는 것과 비슷한 상황인 것이다. 즉, 데이터 공유자는 공식적인 학술 크레딧을 받지 못하고 있다. 감사의글(acknowledgements) 섹션에서는 데이터를 특정 인물로부터 기증받았다는 표현으로 언급만 하고 지나가고 있어서 WoS 등의 학술 데이터베이스가 연구데이터의 재이용을 색인할 수 없는 형식으로 연구데이터의 재이용이 언급되고 있었다.

〈표 7〉은 생명과학 분야의 데이터 공유 및 재이용과 관련된 학술논문 내의 위치를 비교분석한 결과이다. 〈표 8〉에서의 '반복'이라는 단어의 뜻은, 동일한 현상이 같은 인스턴스에

〈표 7〉 데이터 재이용, 공유의 학술논문에서의 위치의 상세 비교

학술논문 내의 위치	데이터 재이용	데이터 재이용/반복	데이터 재이용/공유	데이터 공유	데이터 공유/반복	총합
초록	7	3	0	3	0	13
감사의글	24	9	0	11	2	46
각주	20	4	0	5	2	31
본문	529	274	8	169	84	1,064
보충자료	40	12	0	120	16	188
참고문헌	52	9	0	4	0	65
총합	672	311	8	311	105	1,407

서 반복적으로 동시에 발견된 경우를 의미하는데 예를 들어 '데이터 재이용/반복'의 경우, 동일한 인스턴스에서 데이터 재이용이 반복적으로 여러 번 발견된 경우를 의미하고, '데이터 공유/반복'의 경우 동일한 인스턴스에서 데이터 공유가 여러 번 발견된 인스턴스를 의미한다. 본 연구는 생명과학분야의 학술논문에서의 연구데이터 재이용이 연구데이터의 공유보다 널리 퍼져있음을 확인하였다. 데이터 재이용과 공유가 동시에 발견되는 인스턴스는 드물었다. 데이터 공유 및 재이용의 빈도수는 내림차순으로 본문(1,064회), 보충자료(188회), 참고문헌(65회), 감사의글(46회), 각주(31회), 초록(13회)이었다. 연구데이터의 공유 및 재이용은 대부분 본문(main text)에 주로 위치해 있었으며, 특히 연구방법론 섹션에서 발견되었다. 대부분의 경우 참고문헌에는 데이터의 공유 및 재이용이 위치해 있지 않았다. 이는 학술데이터베이스가 연구데이터를 추적 및 색인할 수 없음을 의미하며, 데이터 재이용이 공식적으로 색인되지 못하고 있음을 보여준다. 데이터의 재이용 및 공유는 동일한 텍스트내에 여러 번 위치하는 '반복'이 자주 발견되었다.

〈표 8〉은 생명과학분야의 학술논문에 나타난 비공식적인 데이터 인용의 데이터의 형태, 지시어/구문, 학술논문에서의 출현 위치, 실제 예시를 보여준다. 데이터 공유, 재이용은 논문의 본문, 보충자료, 감사의 글에서 언급만 하고 지나갈 뿐 참고문헌에는 연구데이터가 위치해 있지 않아서, 데이터 공유자는 인용 등의 학술 크레딧을 받을 수 없음을 확인할 수 있다. 마지막 예시인 'GenBank accession numbers for the sequences are GQ919302GQ920615.(해당 시퀀스의 GenBank 접근 번호는 GQ919302GQ920615 입니다.)' 문구의 경우, 본문에 데이터에 대한 고유하고 지속가능한(persistent) 식별번호를 명시하고 있지만, 이 역시 참고문헌에는 명시하지 않아서, 데이터 공유자는 인용 등의 학술 크레딧을 받을 수 없음을 확인할 수 있다.

〈표 8〉 학술논문에 나타난 데이터 공유, 재이용, 비공식적인 데이터 인용의 구체적인 예시

데이터 형태	지시어	출현 위치	예시
데이터 공유	downloaded	보충자료	Meta-analysis results can be **downloaded** from the SSGAC website. (메타분석 결과는 SSGAC 웹사이트에서 **다운로드**할 수 있습니다.)
데이터 재이용	publicly available	본문	We used the Blood eQTL browser, a **publicly available** database, to examine whether any lead variants, or their most correlated HapMap proxy (with R2 〉 0.8), were associated with expression levels of nearby genes in blood. (**공개적으로 이용 가능한** 데이터베이스인 Blood eQTL 브라우저를 사용하여 리드 변형이나 가장 상관관계가 높은 HapMap 프록시(R2 〉 0.8)가 혈액 내 근처 유전자의 발현 수준과 연관되어 있는지 조사했습니다.)
	samples	감사의글	We acknowledge use of DNA **samples** from the NIHR Cambridge BioResource. (우리는 NIHR 케임브리지 생물자원의 DNA **샘플**을 사용했습니다.)
비공식적인 데이터 인용	accession	본문	GenBank **accession numbers** for the sequences are GQ919302GQ920615. (해당 시퀀스의 GenBank **접근 번호**는 GQ919302GQ920615입니다.)

5. 논 의

본 연구는 데이터 공유, 재이용, 인용을 식별하기 위하여 반자동 텍스트 검색 기법을 활용하였는데, 현재까지는 공식적인 데이터 인용을 측정하기 위한 자동화되고 일반화할 수 있는 표준화된 방법론이 부족한 실정이다. 비공식적인 데이터 인용이 공식적인 데이터 인용보다 널리 퍼져있기 때문이다. 본 연구에서 활용한 반자동 텍스트 검색 기법은 연구데이터의 공유, 재이용, 인용의 식별에 대한 정확률(precision)이 상승할 수 있다는 점에서는, 학술논문을 사람이 수동으로 읽어서 데이터 공유, 재이용, 인용을 확인하는 방법보다는 효율적인 방법론으로 활용될 수 있다. 하지만 비공식적인 인용의 모든 인스턴스를 식별하지 못할 수 있는 한계를 가지고 있다. 공유된 데이터의 많은 재이용을 위하여 연구데이터를 인용할 수 있는 충분한 정보가 제공되어야 한다. 더욱 효과적으로 학술논문에서의 데이터 공유, 재이용, 인용을 측정할 수 있는 방법론이 개발된다면 더욱 큰 스케일의 연구가 수행될 수 있을 것이다.

현대 과학 연구에 있어서 연구데이터의 중요성이 더욱 증가하고 있지만, 연구데이터는 역사적으로는 주요한 연구 부산물(research products)로 여겨오지 않아왔다. 연구물을 인용하는 주요 원인 중 하나는 기존의 연구를 인정(acknowledge)하거나 크레딧(credit)을 줄 수 있기 때문이다. 하지만, 이러한 경우는 대부분 학술논문, 도서 등과 같은 전통적인 학술 출판물과 관련이 있다. 학술 출판물에서의 데이터 인용은 아직 관례가 아니다(Late & Kekäläinen, 2020; Park et al., 2018; Zuiderwijk et al., 2020). 데이터 인용은 footnotes, 표의 노트(notes on tables), 이미지(figures)에 데이터를 언급만 하고 지나가서는 좋은 데이터 인용이 아니다. 데이터 인용이 제대로 되려면, 최소한의 요소가 제공되어야 하는데, 그 예로는 저자, 데이터의 제목, 출판 년도, 버전 정보, 데이터 출판자(data publisher), 고유 식별자 등이 있다 (Jessop, 2022). 이는, 버전 정보를 제외하면, 일반적인 학술 출판물을 인용할 때 필요한 최소한의 요소와 비슷하다. 버전 정보는 새로운 데이터가 추가되거나, 현재의 데이터가 수정(correct)되었을 때 변경된다. 따라서, 연구데이터의 경우, 버전 정보는 중요하다.

본 연구는 데이터 공유 및 재이용이 반복적으로 하나의 인스턴스에서 발견됨을 확인하였다. 하지만, 데이터 공유와 재이용이 반복적으로 발생하더라도 인용의 관점에서는 한 번의 인용으로 측정될 뿐이다. 반복적인 데이터 공유와 재이용 현상은 해당 학술논문에서의 데이터의 중요성을 측정하는 데 활용될 수 있을 것이다. 전통적인 학술지와 연구데이터 간의 인용 관례에 대한 차이는 더 많은 연구를 필요로 한다.

6. 결 론

본 연구의 목적은 생명과학분야 학술논문에 나타난 연구데이터의 공유, 재이용, 인용 현상을 분석하는 것이다. 이를 위하여 미국 국립과학재단의 학문분야 분류코드, Clarivate Analytics사의 Web of Science의 연구분야(research area), Data Citation Index의 연구분야(research area)

를 표본으로 삼았다. 해당 학문 분야 중에서 Data Citation Index에서 공식적인 데이터 인용이 가장 높은 연구데이터를 식별하였다. 이 중 데이터 인용이 가장 높은 상위 30위 연구데이터의 제1저자를 식별한 후, 제1저자가 게재한 학술논문을 식별하였다. 체계적 무작위 표본 추출 기법을 활용하여 학술논문 244개의 전문을 수집하였다. 데이터 공유, 재이용, 인용을 지시하는 지시어 39개를 식별한 후, 이 지시어를 활용하여 244개의 학술논문에 나타난 6,768개의 인스턴스를 추출하였다. 2명의 코더가 내용분석을 통하여 학술논문에서의 출현 위치, 데이터 공유, 재이용, 인용을 코딩하였다.

본 연구는 생명과학 분야는 데이터 재이용(47.8%)이 데이터 공유(22.1%)보다 두 배 이상 높음을 확인하였다. 생명과학분야의 비공식적인 데이터 인용은 생명과학 분야는 95.38%임을 확인하여, 생명과학분야의 데이터 공유자는 인용 등의 공식적인 학술 크레딧을 받지 못하고 있음을 확인하였다. 95.38%의 연구자는 학술논문에서 데이터의 공유, 재이용, 인용을 본문 등에서 언급만 하고 참고문헌에는 명시하지 않고 있기 때문이었다. 생명과학분야의 학술논문에서는 연구데이터의 재이용이 연구데이터의 공유보다 더 많이 발견되었다. 데이터에 대한 직접적인 접근인 디지털 객체 식별자(Digital Object Identifier, DOI)를 제공하는 경우는 드물었다. 저자는 자신이 접근한 데이터 소스(data source)나 데이터 제공자(data provider)를 단순히 언급만 하고 지나치고 있었다. 본 연구의 한계는 연구의 범위가 생명과학분야로 한정되어 있으므로, 다른 학문 분야에도 본 연구의 결과를 일반화하기에는 어렵다는 점이다. 본 연구의 공헌은, 연구데이터 공유, 재이용, 인용에 대한 관례에 대한 이해를 제공하고, 대량의 인스턴스를 식별하기 위한 반자동 텍스트 검색 기법을 제안했다는 점이다. 향후 연구는 더 다양한 학문 분야를 대상으로 한 확장된 연구를 하고자 한다.

참 고 문 헌

국가연구개발혁신법. 법률 제18645호.

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

Altman, M. & King, G. (2007). A proposed standard for the scholarly citation of quantitative data. D-Lib Magazine, 13(3/4). https://doi.org/10.1045/march2007-altman

Borgman, C. L. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World. Cambridge, MA: MIT Press.

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, D. W., Laurie, G., O'Neill, B. O., Rawlins,

M., Thornton, D. J., Vallance, P., & Walport, M. (2012). Science as an Open Enterprise. Available: https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf

Clarivate Analytics (2024). Data Citation Index. Available: https://clarivate.com/products/scientific-and-academic-research/research-discovery-and-workflow-solutions/webofscience-platform/data-citation-index/

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013). Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Science Journal, 12, CIDCR1-CIDCR7. https://doi.org/10.2481/dsj.OSOM13-043

Cohen, J. (1995). Share and share alike isn't always the rule in science. Science, 268(5218), 1715-1718. https://doi.org/10.1126/science.7792594

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. PLoS ONE, 15(4), e0230416. https://doi.org/10.1371/journal.pone.0230416

Corti, L., Eynden, V., Bishop, L., & Woollard, M. (2014). Managing and Sharing Research Data: A Guide to Good Practice. Los Angeles, CA: SAGE.

Cronin, B. (1995). The Scholar's Courtesy: The Role of acknowledgement in the Primary Communication Process. London: Taylor Graham.

Cronin, B. (2001). Acknowledgement trends in the research literature of information science. Journal of Documentation, 57(3), 427-433. https://doi.org/10.1108/EUM0000000007089

Curty, R. (2015). Beyond "Data Thrifting": An Investigation of Factors Influencing Research Data Reuse in the Social Sciences. Doctoral dissertation, Syracuse University, United States.

Curty, R. G., Crowston, K., Specht, A., Grant, B. W., & Dalton, E. D. (2017). Attitudes and norms affecting scientists' data reuse. PLoS ONE, 12(12), e0189288. https://doi.org/10.1371/journal.pone.0189288

Digital Curation Centre (2010). Data Dimensions: Disciplinary Differences in Research Data Sharing, Reuse, and Long Term Viability. Available: https://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf

Dorta-González, P., González-Betancor, S. M., & Dorta-González, M. I. (2021). To what extent is researchers' data-sharing motivated by formal mechanisms of recognition and credit? Scientometrics, 126, 2209-2225. https://doi.org/10.1007/s11192-021-03869-3

Drachen, T. M., Ellegaard, O., Larsen, A. V., & Dorch, S. B. (2016). Sharing data increases citations. LIBER Quarterly, 26(2), 67-82. https://doi.org/10.18352/lq.10149

Faniel, I. M. & Zimmerman, A. (2011). Beyond the data deluge: a research agenda for large-

scale data sharing and reuse. The International Journal of Digital Curation, 6(1), 58-69. https://doi.org/10.2218/ijdc.v6i1.172

Faniel, I. M., Frank, R. D., & Yakel, E. (2019). Context from the data re-user's point of view. Journal of Documentation, 75(6), 1274-1297. https://doi.org/10.1108/JD-08-2018-0133

Funk, C., Hefferon, B., & Johnson, C. (2019). Trust and Mistrust in Americans' Views of Scientific Experts. Available: https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/08/PS_08.02.19_trust.in_.scientists_FULLREPORT.pdf

Jessop, P. (2022). Data Citation: A Guide to Best Practice. Publications Office of the European Union. https://doi.org/10.2830/59387

King, G. (1995). Replication, replication. PS: Political Science and Politics, 28(3), 444-452. https://doi.org/10.2307/420301

Kling, R. & Spector, L. (2003). Rewards for scholarly communication. In D. L. Andersen eds. Digital Scholarship in the Tenure, Promotion, and Review Process. Armonk, NY: ME Sharpe, Inc, 78-103.

Late, E. & Kekäläinen, J. (2020). Use and users of a social science research data archive. PLoS ONE, 15(8), e0233455. https://doi.org/10.1371/journal.pone.0233455

Lawrence, B., Jones, C., Mattew, B., Pepler, S., & Callaghan, S. (2011). Citation and peer review of data: moving towards formal data publication. International Journal of Digital Curation, 6, 4-37. https://doi.org/10.2218/ijdc.v6i2.205

Lin, J. (2009). Is searching full text more effective than searching abstracts? BMC Bioinformatics, 10(1), 46. https://doi.org/10.1186/1471-2105-10-46

Mooney, H. & Newton, M. (2012). The anatomy of a data citation: discovery, reuse, and credit. Librarianship and Scholarly Communication, 1(1), eP1035. https://doi.org/10.7710/2162-3309.1035

National Institutes of Health (2003). Final NIH Statement on Sharing Research Data. Available: https://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html

National Science Foundation (2011). Digital Research Data Sharing and Requirement. Available: https://www.nsf.gov/nsb/publications/2011/nsb1124.pdf

OECD (2015). Making Open Science a Reality. Paris: OECD Publishing. https://doi.org/10.1787/23074957

Palmer, C. L. & Cragin, M. H. (2008). Scholarship and disciplinary practice. Annual Review of Information Science and Technology, 42, 163-212.

Park, H. & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. Scientometrics, 111(1), 443-461. https://doi.org/10.1007/s11192-017-2240-2

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. Journal of the Association for Information Science and Technology, 69(11), 1346-1354. https://doi.org/10.1002/asi.24049

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

Piwowar, A. H., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. PLoS ONE, 2(3), e308. https://doi.org/10.1371/journal.pone.0000308

Piwowar, H. & Chapman, W. W. (2010). Public sharing of research datasets: a pilot study of associations. Journal of Informetrics, 42(2), 148-156. https://doi.org/10.1016/j.joi.2009.11.010

Piwowar, H. A. & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, e175. https://doi.org/10.7717/peerj.175

Piwowar, H. A. (2011). Who shares? who doesn't? factors associated with openly archiving raw research data. PLoS ONE, 6(7), e18657. https://doi.org/10.1371/journal.pone.0018657

Robinson-Garcia, N., Mongeon, P., Jeng, W., & Costas, R. (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. Journal of Informetrics, 11(3), 841-854. https://doi.org/10.1016/j.joi.2017.07.003

Sielemann, K., Hafner, A., & Pucker, B. (2020). The reuse of public datasets in the life sciences: Potential risks and rewards. PeerJ, 8, e9954. https://doi.org/10.7717/peerj.9954

Silvello, G. (2018). Theory and practice of data citation. Journal of the Association for Information Science and Technology, 69(1), pp. 6-20. https://doi.org/10.1002/asi.23917

Spertus, J. A. (2012). The double-edged sword of open access to research data. Circulation: Cardiovascular Quality and Outcomes, 5(2), 143-144. https://doi.org/10.1161/CIRCOUTCOMES.112.965814

Springer Nature Group. (n.d.). Research Data Policies. Available: https://www.springernature.com/gp/authors/research-data-policy

Steel, K. M., Thompson, H., & Wright, W. (2019). Opportunities for intra-university collaborations in the new research environment. Higher Education Research & Development, 38(3), 638-652.

https://doi.org/10.1080/07294360.2018.1549537

Sterlling, T. D. & Weinkam, J. J. (1990). Sharing scientific data. Communications of the ACM, 33(8), 112-119. https://doi.org/10.1145/79173.79182

Ten Hoopen, P., Finn, R. D., Bongo, L. A., Corre, E., Fosso, B., Meyer, F., Mitchell, A., Pelletier, E., Pesole, G., Santamaria, M., Willassen, N. P., & Cochrane, G. (2017). The metagenomic data life-cycle: Standards and best practices. GigaScience, 6(8), 87. https://doi.org/10.1093/gigascience/gix047

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. PLOS ONE, 6(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Tenopir, C., Rice, N. M., Allard, S., Baird, L., Borycz, J., Christian, L., Gant, 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

Vandewalle, P. (2012). Code sharing is associated with research impact in image processing. Computing in Science & Engineering, 14(4), 42-47. https://doi.org/10.1109/MCSE.2012.63

Zuiderwijk, A., Shinde, R., & Wei, J. (2020). What drives and inhibits researchers to share and use open research data? A systematic literature review to analyze factors influencing open research data adoption. PLoS One, 15(9), e0239283. https://doi.org/10.1371/journal.pone.0239283

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally 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

The National Research Development and Innovation Act. Act 18645.