STEM 분야의 연구데이터 분석*

- Data Citation Index를 중심으로 -

Analyzing STEM Research Data in the Data Citation Index

박 형 주 (Hyoungjoo Park)**

목 차

초 록

본 연구의 목적은 데이터 공유 및 인용이 높은 STEM(Science, Technology, Engineering, Mathematics) 분야 연구데이터의 유형과 데이터 리포지토리를 분석하는 것이다. STEM 분야의 식별은 미국국립과학재단(National Science Foundation, NSF)의 학문코드(discipline code)와 Data Citation Index의 연구분야(research area)를 비교하여 식별하였다. 식별된 STEM 학문 분야는 천체물리학, 생명과학, 화학, 컴퓨팅, 지구과학, 공학, 수학, 기술로 총 8개의 학문 분야였다. 본 연구는 전세계 630만개 이상의 STEM 분야 연구데이터를 수집 및 분석하였다. 본 연구는 STEM 분야에서 다양한 종류의 연구데이터 유형이 인용을 받음을 확인하였다. 학문 분야별로 주요한 데이터 리포지토리가 다양하였으며, 높은 데이터 공유가 높은 데이터 인용을 의미하지는 않았다. 데이터 인용을 받은 연구데이터의 유형은 STEM 학문 분야별로 다양하였다. 데이터 인용을 받은 연구데이터의 유형은 모두 양적 데이터였다. 데이터 인용을 받은 질적 데이터는 없었다. STEM 분야에서 공식적인 데이터 인용(formal data citation)은 학술 커뮤니티의 관례가 아니다. 본 연구의 공헌은 전세계 630만개 이상의 대량의 STEM 분야 연구데이터의 관례를 실질적으로 분석하였다는 것이다.

ABSTRACT

This study examines the types of research data and data repositories in STEM(Science, Technology, Engineering, and Mathematics) fields that demonstrate high levels of data sharing and citation. STEM fields were identified by aligning discipline codes from the National Science Foundation(NSF) with research areas listed in the Data Citation Index. The selected disciplines include astrophysics, biological sciences, chemistry, computing, earth sciences, engineering, mathematics, and technology, encompassing a total of eight fields. This study involved the collection and analysis of over 6.3 million STEM research data records. The findings indicate that various types of research data across STEM fields are cited; however, key data repositories differ by discipline. Notably, high levels of data sharing do not necessarily correspond with high data citation rates. The types of cited research data also vary across disciplines, with all cited data being quantitative—no qualitative data received citations. Despite the growing emphasis on open science, formal data citation remains uncommon in STEM fields. This study contributes to the literature by providing a comprehensive analysis of data-sharing and citation practices across more than 6.3 million STEM research data records worldwide.

키워드: 연구데이터, 데이터 인용, 데이터 리포지토리, 데이터 공유, Data Citation Index
Research Data, Data Citation, Data Repository, Data Sharing, Data Citation Index

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2022-NR068754).

** 충남대학교 문헌정보학과 조교수(hyoungjoo.park@cnu.ac.kr / ISNI 0000 0004 6442 7767)

논문접수일자: 2025년 1월 24일 최초심사일자: 2025년 2월 1일 게재확정일자: 2025년 2월 14일
한국문헌정보학회지, 59(1): 489-516, 2025. http://dx.doi.org/10.4275/KSLIS.2025.59.1.489

1. 서 론

연구데이터는 수치, 텍스트, 이미지, 음성 데이터 등 사실에 기반한 디지털 데이터이며 주로 과학분야의 연구를 위하여 사용되고 연구 결과를 검증하는 데 활용될 수 있다(OECD, 2007). 연구데이터는 관찰 데이터, 실험 데이터, 분석 데이터 등이 있다. 데이터 공유는 원시 또는 전처리된 데이터를 연구자 혹은 기관이 자발적으로 또는 규범에 따라서 공개하는 행위이다(Curty, 2015). 데이터 공유는 데이터 설명 및 교환을 위한 표준화 작업, 데이터 관리 및 큐레이션을 위한 사이버 인프라스트럭처 구축 등의 활동을 포함하며, 이러한 과정은 데이터 공유의 효과와 효율을 결정짓는 핵심 요소로 작용한다.

데이터는 신속하게 공유되어야 하며, 이는 생명을 구하는 데 기여할 수 있다. 데이터 공유는 과학적 지식의 발전을 가속화한다는 점에서, 팬데믹과 같은 공공 비상 상황에서 핵심적인 요소로 간주된다. 예를 들어, Zika 바이러스의 문제 해결을 위하여 데이터 리포지토리가 구축되어 데이터가 신속히 공개되었다(Bulter, 2016). Elsevier는 2020년 1월 이후 코로나19 정보센터(Elsevier's Novel Coronavirus Information Center)를 개설하여, 코로나19와 관련된 연구데이터를 PubMed Central 및 데이터 리포지토리를 통해 즉시 공개할 수 있도록 하였다(Elsevier, 2020). 하지만 많은 연구자들은 데이터 공유와 관련된 기술과 실질적인 능력을 충분히 갖추지 못한 상황이다(Perrier et al., 2020). 연구자들의 데이터의 공유에 대한 태도는 학문 분야별로 차이가 있다. 이는 각 학문분야의 고유한 연구 요구사항과 데이터 활용 방식에 기인한다(Borgerud & Borglund, 2020; Corti et al., 2014). 데이터 인용은 연구에서 활용된 연구데이터에 대한 출처를 명시하고 해당 데이터를 학술적으로 인용하는 것이다(Altman & King, 2007). 데이터 인용은 학술적 공로를 인정하고 저자에게 인센티브를 제공할 수 있다. 하지만 많은 과학자들은 데이터 공유에서 충분한 보상과 인정이 주어지지 못한다고 주장한다(Dorta-González et al., 2021).

연구데이터의 관례를 분석함에 있어서 학문 분야 간 차이는 중요하다. STEM(Science, Technology, Engineering, Mathematics) 분야는 데이터 공유가 개별 학문 수준에서 주로 연구되어 왔다. 그러나 현대 과학에서 연구실, 학과, 대학, 국가 간의 협업이 일반화됨에 따라, 학문 차이를 고려하지 않으면 전반적인 과학 분야의 데이터 공유 및 인용을 연구하기 어려울 수 있으므로 여러 학문을 살펴보는 것이 필요하다. 본 연구는 학문 분야를 STEM 분야로 한정하였는데, STEM 분야의 데이터는 수치데이터가 주로 활용되므로, 인간 연구를 수행하고 문맥(context)이 중요한 역할을 하는 사회과학 분야의 데이터와는 차이가 있으므로, STEM 분야의 연구데이터를 별도로 분석할 필요가 있기 때문이었다. 또한, STEM 분야의 연구데이터 공유가 상대적으로 오래되었기에(National Institues of Health, 2003; National Science Foundation, 2011) STEM 분야를 분석하면 연구데이터의 관례를 포괄적으로 살펴볼 수 있다. 각 학문 분야는 각각 다른 연구데이터 공유 관례가 존재하므로(Helbig et al., 2015; Mongeon et al., 2017) STEM 분야의 연구데이터를 학문 분야별로 분석하고 통합해서 분석하는 것이 필요하

였다. 연구 질문은 다음과 같다.

- 연구문제 1: STEM 분야에서 데이터 인용, 이용이 높은 연구데이터의 유형은 무엇인가?
- 연구문제 2: STEM 학문에서 데이터 공유 및 인용이 높은 데이터 리포지토리는 무엇인가?

2. 선행연구

데이터 공유는 연구자 또는 기관이 자발적으로 혹은 기관의 규범에 따라 원시 또는 전처리된 연구데이터를 공개하는 행위이다(Curty, 2015). Funk et al.(2019)은 데이터 공유는 대중의 신뢰를 증진시키는 효과가 있다고 보고하였다. 데이터 공유를 위해 연구자들은 시간이 지나도 지속 가능한 파일 형식과 기술 호환성을 보장해야 한다. 이는 운영체제와 기술이 지속적으로 변화하기 때문이다. 연구자들 간에 사용되는 운영체제와 기술은 매우 다양하여, 데이터 저장 방식에서도 차이가 있다(Corti et al., 2014). 데이터 공유는 연구자들이 타인이 공유한 데이터를 활용함으로써 보상을 받을 기회를 제공한다. 그러나 연구자들은 데이터를 학술지에 공유하기보다는 보유하려는 경향을 보인다(Boulton et al., 2012; Cohen, 1995; Piwowar, 2011). 연구자들은 데이터 공유로 인해 얻는 보상이 낮거나 없다고 인식하는 경우, 데이터를 공유하지 않는 경향이 강하다(Sterlling & Weinkam, 1990). 반면, 연구자들이 데이터 공유로 인한 보상을 명확히 인식할 경우, 데이터 공유 행동이 촉진된다(Kling & Spector, 2003). Tenopir et al.(2011)은 설문 응답자 204명 중 47명(23%)이 연구데이터의 손쉬운 접근 가능성에 동의하거나 부분적으로 동의한다고 보고하였다. 반면, 생물학 분야 연구자의 49%는 데이터 공유에 동의하거나 부분적으로 동의했으며, 이는 사회과학 분야 연구자보다 약 두 배 높은 수치라고 하였다. 미국 성인의 57%는 연구자가 데이터를 공유한다면 연구 결과를 신뢰할 수 있다고 하였다(Funk et al., 2019).

데이터 리포지토리는 연구데이터를 보관 및 공개하는 플랫폼으로, 연구 결과물로서 데이터를 저장하고 관리하는 역할을 한다. 데이터 리포지토리를 통한 데이터 공유는 서지 정보와 리포지토리 간 데이터 전송을 표준화하고 간소화하여 연구자들 간의 데이터 접근성을 향상시킬 수 있다. 데이터 리포지토리는 데이터 센터나 도서관 내에 구축될 수 있으며, 데이터 관리, 접근 및 지속 가능성을 지원하는 동시에 메타데이터를 데이터로 변환하는 기능도 제공한다. 데이터 리포지토리는 연구재단, 학술단체, 출판사, 정부, 전문 학회에 의해 운영된다. Springer Nature(2019)는 생물학, 재료과학, 화학, 물리학 등 다양한 학문 분야별로 데이터 리포지토리 추천 목록을 제공하고 있다. 데이터 제공자는 리포지토리에 데이터를 저장할 때 데이터를 충분히 설명할 필요가 있으며, 사용자가 저장된 데이터를 탐색하고 재이용 가능성을 평가할 수 있는 환경을 제공해야 한다. 사용자들이 데이터 리포지토리를 신뢰하는 데에는 개인적 경험, 리포지토리의 평판 및 관례, 동료 연구자들의 경험이 중요한 역할을 한다(Downs & Chen, 2006; Yakel et al., 2013; Yoon, 2014).

데이터 리포지토리 연구는 리포지토리의 공동저자 분석, 메타데이터 분석, 서비스 분석, 신뢰성 및 효율성에 관한 연구가 수행되었다. Burns et al.(2013)은 기관 리포지토리의 비용과 가치에 대한 설문조사 연구를 통하여, 기관 리포지토리의 비용과 가치에 대한 논의가 연구중심대학에서 논의될 필요가 있다고 하였다. Costa et al.(2016)은 유전체학 분야의 대규모 리포지토리인 GenBank의 공동저자 패턴 분석을 통해, 경력이 짧은 연구자는 공식적인 출판물 보다는 데이터셋을 통한 협업이나 데이터 공유를 하려는 경향이 더 강하다는 것을 밝혔다. Gries et al.(2018)은 환경분야 데이터 리포지토리의 상호운영성을 촉진하기 위해서는, 여러 리포지토리가 합의된 표준 및 모범 사례를 공유하는 커뮤니티 플랫폼으로서 역할을 해야 한다고 하였다. Oblasser et al.(2020)은 리포지토리에 색인된 각각의 레코드에 필터링 가능과 검색이 가능한 리포지토리의 레지스트리를 구축하기 위해서는, 각각의 색인된 리포지토리 레코드에 풍부한 메타데이터를 제공하는 것이 중요하다고 하였다. 특히, 기계로 실행이 가능한 데이터 관리계획의 내용과 데이터 정책 요소를 필터링 레지스트리에 제공하면, 상위의 리포지토리를 제안하는 데 효과적이라고 하였다. 김주섭 외(2023)는 생태분야 데이터 리포지토리의 운영 현황을 분석하고 EcoBank 서비스를 제안하였는데, 연구데이터 정책, 연구데이터 품질 검토, 연구데이터 관리 교육 및 워크샵(workshop)임을 밝혔다. 이혜림(2023)은 CoreTrustSeal에서 인증을 받은 데이터 리포지토리의 디지털 보존 정책을 비교 분석하여 정책 프레임워크에 필요한 구성요소를 추출하였다. 이혜림 외(2024)는 CoreTrustSeal(CTS) 인증 획득을 한 데이터 리포지토리의 신청서를 비교 분석하여, 리포지토리의 신뢰성과 효율성에 미치는 영향을 분석하였다.

데이터 인용은 학술연구에 활용된 데이터의 출처를 명시하고 데이터를 학술적으로 인용하는 것이다(Altman & King, 2007). 데이터 인용은 주요 연구 결과에 대한 보상과 인정을 위해 연구데이터에 대한 참고정보를 제공한다. 일부 저명한 주요 출판사는 데이터 인용을 권장한다. 예를 들어, Springer는 모든 Nature 저널과 약 1,600개의 Springer Nature 저널에 표준화된 연구데이터 정책을 채택해서 운영하고 있다(Springer Nature, n.d.). 고유하고 지속가능한 식별자가 직접적인 데이터 인용 및 인용의 접근성을 보장하지 않는다면, 데이터 인용에 있어 어려움이 발생할 수 있다. 데이터 인용을 위한 주요 메타데이터 요소는 저자, 제목, 출판일자, 고유식별자 등이 있다(Altman & King, 2007). 더블린코어의 15개 메타데이터 요소 중 생성자, 출판년도, 식별자, 주제 등은 데이터 큐레이션에 필요한 필수 필드로 널리 사용되고 있다. 더블린 코어 메타데이터 요소 집합은 Drayad Application Profile(Ball, 2009; Diamantopoulos et al., 2011)과 같은 다양한 애플리케이션 프로필을 개발하는 데 활용되어, 다양한 플랫폼 간의 상호운용성을 가능하게 한다.

Data Citation Index를 분석하는 것은 연구데이터의 관례를 포괄적으로 살펴보기에 적절할 수 있다. Data Citation Index는 연구데이터가 다양한 학문 분야에서 어떻게 공유, 추적, 그리고 색인되고 있는지 단일 접근점(single access point)에서 살펴볼 수 있도록 돕는 학술 데이터

베이스이다. Data Citation Index는 2012년 Thomson Reuters에 의해 처음 도입되었으며, 이후 Clarivate Analytics로 매각되었다. Data Citation Index는 Web of Science 데이터베이스의 하나로 구독(subscription)에 기반한 서비스를 제공한다. 2024년 11월 현재, Data Citation Index는 전세계 약 450개 이상의 데이터 리포지토리에서 1,500만 개 이상의 데이터세트, 170만 개 이상의 데이터 연구, 그리고 54만개 이상의 소프트웨어를 추적 및 색인한다(Clarivate Analytics, 2024). Data Citation Index를 활용하면 다양한 학문 분야의 연구데이터를 단일 접근점(single access point)에서 검색할 수 있으며, Data Citation Index의 '연관 데이터(associated data)' 기능을 통하여 Web of Science의 학술논문과 연구데이터를 연결하여 색인할 수 있다. Data Citation Index는 Web of Science가 학술논문, 학술발표집 등을 색인하는 것과 유사한 방식으로 연구데이터를 추적하고 색인한다. 조재인(2016)은 2006년부터 2015년까지 약 10년간의 Data Citation Index 데이터의 인용빈도 상위 500위의 데이터를 분석하여, Data Citation Index 데이터의 주요 주제와 데이터 유형을 분석하였다. Park, Wolfram(2018)은 Data Citation Index에서 색인하는 연구소프트웨어의 인용 관례를 분석하였는데, 소프트웨어는 많은 경우 인용되지 않고 있고, 특히 소프트웨어 재이용을 추적하는 것도 쉽지 않은 상황이라고 하였다. 또한, 고유하고 지속가능한 식별자인 Digital Object Identifier(DOI)의 사용 비율도 높지 않음을 확인하였다.

선행연구를 분석한 결과, 데이터 공유, 인용, 리포지토리에 대한 연구는 활발히 진행되어 왔으나, 데이터 리포지토리 별 혹은 연구데이터의 유형 별 데이터 인용, 이용, 그룹저자에 대한 세부연구는 활발히 진행되지 않아왔음을 확인할 수 있었다. 이에, 본 연구는 그 갭을 채우고자 한다.

3. 연구방법론

본 연구의 목적은 데이터 인용이 높은 STEM 분야 연구데이터의 유형과 데이터 리포지토리를 분석하는 것이다. 모집단은 Clarivate Analytics 사의 Data Citation Index에서 색인하는 연구데이터 레코드이다. Data Citation Index를 선택한 이유는 연구데이터의 공식적인 인용 관례를 포괄적으로 살펴볼 수 있기 때문이었다. Data Citation Index는 전세계 약 450개 이상의 데이터 리포지토리에서 약 1,500만개 이상의 데이터세트, 약 170만개 이상의 데이터 연구, 약 54만개 이상의 소프트웨어를 추적 및 색인(Clarivate Analytics, 2024)하므로 데이터 인용의 관례를 포괄적으로 살펴보기에 적절하다. 또한, 공식적인 데이터 인용(formal data citation)을 단일 접근점에서 분석할 수 있는 기능을 제공한다. 본 연구는 STEM 분야로 연구 범위를 한정하였다. STEM 분야의 연구데이터는 양적 데이터 등이 많이 활용이 되므로, 사람을 연구하고 문맥(context)이 중요한 사회과학 분야의 연구데이터와는 차이가 있으므로, STEM 분야를 별도로 연구할 필요가 있기 때문이었다.

〈표 1〉은 본 연구에서 식별된 STEM 분야를 보여준다. STEM 분야의 식별은 미국 국립과학재단(National Science Foundation, NSF)

〈표 1〉 NSF의 학문분야와 Data Citation Index의 연구분야 비교를 통한 STEM 분야 식별

식별된 STEM 분야	NSF의 학문분야 코드	Data Citation Index의 연구분야
Astrophysics(천체물리학)	Astronomy(천문학) Physics(물리학)	Astronomy & Astrophysics(천문학 및 천체물리학), Physics(물리학), Spectroscopy(분광학)
Biological sciences(생명과학)	Biological sciences(생명과학)	Genetics and Heredity(유전학 및 유전), Biochemistry & Molecular Biology(생화학 및 분자생물학), Biotechnology & Applied Microbiology(생명공학 및 응용미생물학), Cell Biology(세포생물학), Developmental Biology(발달생물학), Evolutionary Biology(진화생물학), Marine & Freshwater Biology(해양 및 담수생물학), Mathematical & Computational Biology(수학 및 계산생물학), Microbiology(미생물학), Plant Sciences(식물과학), Reproductive Biology(생식생물학), Environmental Sciences & Ecology(환경과학 및 생태학), Biodiversity & Conservation(생물다양성 및 보존), Research & Experimental Medicine(연구 및 실험의학)
Chemistry(화학)	Chemistry(화학)	Chemistry(화학), Crystallography(결정학)
Computing(컴퓨팅)	Computing(컴퓨팅)	Computer Science(컴퓨터 과학)
Earth sciences(지구과학)	Earth sciences(지구과학)	Geology(지질학), Oceanography(해양학), Geochemistry & Geophysics(지구화학 및 지구물리학), Meteorology & Atmospheric Sciences(기상학 및 대기과학), Water Resources(수자원)
Engineering(공학)	Engineering(공학)	Engineering(공학)
Mathematics(수학)	Mathematical sciences (수학과학)	Mathematics(수학)
Technology(기술)	-	Technology(기술)

의 학문분야 코드(discipline code), Data Citation Index의 연구분야(research area)를 기준으로 하였다. 식별된 8개의 STEM 분야는 천체물리학, 생명과학, 화학, 컴퓨팅, 지구과학, 공학, 수학, 기술 분야이다. 기술 분야는 국립과학재단의 학문분야 코드에는 존재하지 않지만, STEM이라는 단어 자체가 기술(technology)을 포함하고 있으므로 본 연구에 포함하였다. 국립과학재단의 학문분야에는 천문학과 물리학이 분리되어 있지만, 해외의 많은 대학들이 두 학과를 동일한 단과대학으로 운영하고 있으므로 본 연구는 천문학, 물리학을 하나의 분야인 천체물리학으로 묶었다. 학제적 분야(interdisciplinary area)는 포함하지 않았는데, 학제적 분야는 어떤 특정 학문 분야에 포함하기 어렵기 때문이었다(박형주, 2024). 생명과학 분야에서 활용된 Data Citation Index의 연구분야는 박형주(2024)의 연구에서 식별한 연구분야를 활용하였다.

〈표 2〉는 본 연구에서 수집한 STEM 분야별 연구데이터의 레코드 수와, 데이터 인용 총수를 보여준다. 데이터의 수집은 Data Citation Index에서 다운로드 받았다. Data Citation Index는 전세계 450개 이상의 데이터 리포지토리에서 연구데이터를 수집 및 색인하므로(Clarivate

〈표 2〉 본 연구에서 수집된 STEM 학문 분야의 연구데이터 수와 인용 수

학문 분야	수집된 연구데이터의 수	수집된 연구데이터의 인용 수
천체물리학	98,899	13,248
생명과학	100,000	153,715
화학	100,000	109,071
컴퓨팅	10,148	2,298
지구과학	99,493	67,029
공학	97,806	4,519
수학	24,593	140,808
기술	99,996	293,240
총합	630,935	783,928

Analytics, 2024), Data Citation Index를 분석하는 것은 연구데이터의 관례를 살펴보기에 적절하다. Data Citation Index는 하나의 검색어 당 최대 10만개의 레코드를 다운로드 받을 수 있는 기능을 제공한다. 따라서, Data Citation Index에서 데이터 인용이 높은 순으로 정렬한 후, STEM 학문분야 별로 최대 10만개의 연구데이터 레코드를 다운로드 받아서 마이크로소프트 엑셀에 저장하였다. 최종적으로 총 630,935개의 데이터가 수집되었다. 수집된 데이터는 데이터 명, 저자, 그룹 저자, 데이터 리포지토리 명, 데이터 유형, 2013년 이후 이용 횟수, 인용 횟수 등이 포함되어 있었다. 데이터는 2024년 12월에 수집되었다.

데이터의 분석은 연구데이터의 유형 별 데이터 인용, 데이터 리포지토리 별 데이터 공유 및 데이터 인용, Z-score, 그룹저자 등을 분석하였다. '데이터 리포지토리'는 Data Citation Index에서 색인하는, 연구데이터가 공유되어 있는 데이터 리포지토리이다. '데이터 인용'은 Data Citation Index에서 색인하는 공식적인 데이터 인용(formal data citation)이다. '그룹 저자(group author)'는 연구데이터를 공유한 저자가 개인 저자(individual author)가 아닌 그룹 저자인 경우이다. 그룹 저자는 Data Citation Index의 CA 필드태그(field tags)를 활용하여 식별하였다. '데이터 인용 퍼센트'는, '전체 데이터 인용 수' 대비 '데이터 인용'의 백분율이다. 구체적으로, 분자는 '개별 데이터 인용 수', 분모는 '전체 데이터 인용수'이다. '데이터 공유 별 인용 횟수'는, '데이터 유형' 별로 '데이터 공유'와 '데이터 인용'을 구한 값으로, 분자는 '데이터 인용수', 분모는 '데이터 공유수'이다. '데이터 이용수'는, 2013년 이후의 이용 횟수이며, Data Citation Index의 U2 필드태그로 식별하였다. '데이터 이용의 퍼센트'는, 전체 데이터 이용 수 대비 데이터 이용의 백분율이다. 'Z-score(표준점수)'는 데이터의 값이 평균에서 얼마나 떨어져 있는 지를 표준편차 단위로 나타낸 값이다. Z-score를 살펴보면 데이터 공유수와 인용수를 표준화할 수 있어서 연구 결과의 해석을 객관화할 수 있다. Z-score는 계산식을 활용하여 마이크로소프트 엑셀로 계산하였다. 수집된 데이터는 엑셀의 피벗(pivot) 테이블 기능을 활용하여 각 필드를 필터링한 후 행과 값을 구하여 데이터를 분석하였다.

4. 결 과

결과 섹션은 STEM 학문 분야별로 구성되어 있으며, 구체적으로 천체물리학 분야, 생명과학 분야, 화학 분야, 컴퓨팅 분야, 지구과학 분야, 공학 분야, 수학 분야, 기술 분야, STEM 분야 전체로 구성되어 있다. 〈표 2〉의 '수집된 데이터 수'는 '학문 분야별로 수집된 연구데이터의 총합'이고, 데이터 인용 및 이용의 총합은 '상위 10위 연구데이터의 총합'이므로, 두 개의 총합은 다르다.

4.1 천체물리학 분야

〈표 3〉은 천체물리학 분야의 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 총 98,899개의 연구데이터가 수집되었으며, 13,248번의 데이터 인용, 98,898번의 데이터 이용이 있었다. 천체물리학 분야의 연구데이터는 총 18개의 데이터 유형이 있었다. 데이터 인용을 많이 받은 상위 3위의 연구데이터의 유형은 '비어 있음'(6,115회, 46.16%), '질량 스펙트럼 데이터'(6,070회, 45.82%), '소프트웨어'(599회, 4.52%)였다. 상위 10위의 연구데이터가 전체 데이터 인용의 99.61%를 받고 있었다. 2위를 차지한 '질량 스펙트럼 데이터'는 질량분석기(Mass Spectrometer)를 통해 생성된 데이터로, 화학적, 생물학적 샘플에 포함된 분자(예: 단백질, 화합물, 이온)의 질량을 분석한 정보, 그 물질의 구성 성분, 분자의 질량과 그 상대적인 양을 분석한 데이터를 의미하며(Smith & Kelleher, 2018), 천문학 분야에서는 성간 물질 및 운석의 화학적 조성 연구에 활용된다. 4위를 차지한 '천문 데이터'는 천체의 위치와 형태를 관찰하기 위하여 촬영된 이미지 데이터 및 천체에서 방출되거나 흡수된 전자기파의 스펙트럼을 기록한 데이터 등이 있다. 천체물리학 분야의 연구데이터가 많이 큐레이션되는 이유는

〈표 3〉 천체물리학 분야의 데이터 공유별 인용수가 높은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	(비어 있음)	6,115	46.16	35,429	35.82
2	mass spectral data(질량 스펙트럼 데이터)	6,070	45.82	24,525	24.80
3	software(소프트웨어)	599	4.52	1,203	1.22
4	astronomical data(천문 데이터)	235	1.77	1,509	1.53
5	scattering data(산포 데이터)	69	0.52	68	0.07
6	mass spectrometry(질량 분석법)	29	0.22	29,101	29.43
7	software used in astronomy or astrophysics research (천문학 또는 천체물리학 연구에 사용되는 소프트웨어)	28	0.21	15	0.02
8	dataset(데이터 세트)	19	0.14	6	0.01
9	NMR results(NMR 결과)	18	0.14	15	0.02
10	experiment session(실험 세션)	15	0.11	3,188	3.22
총합		13,197	99.61	95,059	96

첫째, 표준화된 형식을 사용하기 때문일 수 있는데 예를 들어 나노미터, 전자볼트 등의 표준화된 표기법을 준수하고, 둘째, 높은 품질의 메타데이터를 보유하고 있기 때문일 수 있다. 요약하면, 천체물리학 분야는 연구데이터 유형을 설정하지 않고 공유 및 인용되는 경우가 46.16%였으며, 데이터 인용이 주로 일어나는 데이터 유형은 '질량 스펙트럼 데이터'였다. 천체물리학 분야의 리포지토리는 주로 물질에 대한 구성성분을 분석한 데이터들이 주로 발견되었다. 예를 들어, 행성과 행성 사이에 존재하는 행간 물질에 대한 데이터, 행성의 질량 분석 데이터, 스펙트럼 데이터 들이 주로 발견되었다.

〈표 4〉는 천체물리학 분야에서 데이터 공유별 인용을 많이 받은 상위 10위의 데이터 리포지토리를 보여준다. 천체물리학 분야는 데이터 공유가 주로 일어나는 데이터 리포지토리인 MassBank가 있지만, 높은 데이터 공유가 반드시 높은 데이터 인용으로 이어지지는 않았다. '고립된 은하의 성간 매질 분석' 및 '지구 행성 탐색' 리포지토리의 경우, 표준편차가 0이어서 z-score가 null이다. 데이터 리포지토리는 학문 특화된 리포지토리(discipline-specific repository)가 주로 발견되었다. MassBank의 그룹저자 수가 3,395로 많았다.

4.2 생명과학 분야

〈표 5〉는 생명과학 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 생명과학 분야의 연구데이터는 총 100,000개가 수집 및 분석되었으며, 총 데이터 인용 수

〈표 4〉 천문물리학 분야의 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유	데이터 인용	데이터 공유 별 인용	데이터 인용 z-score	그룹 저자 수
1	Planetary Data System(행성 데이터 시스템)	301	765	2.54	0	6
2	Analysis of the Interstellar Medium of Isolated Galaxies(AMIGA)(고립된 은하의 성간 매질 분석)	4	4	1	null	0
3	Earthbound Planet Search(지구 행성 탐색)	1	1	1	null	0
4	Small Angle Scattering Biological Data Bank(소각 산란 생물학 데이터 뱅크)	2,392	2,092	0.87	0	0
5	Astrophysics Source Code Library(천체물리학 소스 코드 라이브러리)	1,216	627	0.52	0	20
6	CSIRO Data Access Portal(CSIRO 데이터 액세스 포탈)	286	130	0.45	0	0
7	Legacy Archive for Microwave Background Data Analysis(LAMBDA)(마이크로파 배경 데이터 분석을 위한 레거시 아카이브)	7	2	0.29	0	7
8	MassBank	24,525	6,070	0.25	0	3,395
9	Oak Ridge Leadership Computing Facility(OLCF) Constellation Portal(오크리지 리더십 컴퓨팅 시설 별자리 포털)	224	57	0.25	0	1
10	German Astrophysical Virtual Observatory Data Center(독일 천체물리학 가상 관측소 데이터 센터)	21	4	0.19	0	7

〈표 5〉 생명과학 분야의 데이터 공유별 인용수가 높은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	(비어 있음)	54,227	35.28	858	56.41
2	quantitative trait locus map & information(양적 형질 유전자좌 지도 및 정보)	30,348	19.74	2	0.13
3	longitudinal(종단적)	17,930	11.66	41	2.70
4	image stored as floating point number (4 bytes) (부동 소수점 숫자(4바이트)로 저장된 이미지)	10,671	6.94	32	2.1
5	expression profiling by high throughput sequencing(고처리량 시퀀싱을 통한 발현 프로파일링)	3,107	2.02	65	4.27
6	longitudinal: cohort(종단적: 코호트)	2,830	1.84	6	0.39
7	specialized mix(특수 혼합)	2,462	1.6	126	8.28
8	nucleotide sequencing information(뉴클레오티드 시퀀싱 정보)	1,949	1.27	24	1.58
9	software(소프트웨어)	1,361	0.89	7	0.46
10	genome binding/occupancy profiling by high throughput sequencing; SRA(고처리량 시퀀싱을 통한 게놈 결합/점유 프로파일링; SRA)	1,332	0.87	0	0
총합		126,217	82.11	1,161	76.33

는 153,715개였다. 데이터 유형을 명시하지 않은 '비어있음'이 가장 높은 데이터 인용을 받고 있었다. 데이터 인용이 높은 상위 3개의 데이터 유형은 '비어 있음', '양적 형질 유전자좌 지도 및 정보', '종단적'이었다. 총 데이터 이용수는 1,521회였다. 생명과학 분야에서는 유전체 서열, 단백질 구조, 발현 패턴, 대사 경로 등 다양한 형태의 데이터가 생성된다. 이러한 데이터는 연구 목적에 따라 서로 다른 형식과 구조를 가진다. 생명과학 데이터는 다양한 출처와 형식으로부터 수집되므로, 이를 효과적으로 통합하고 분석하기 위한 시스템 및 알고리즘의 설계가 중요하기 때문이다(송영옥 외, 2010). 또한, 생물학 분야의 분석이 유전체 단위로 바뀜에 따라 방대한 양의 유전체 정보를 분석 및 해석하기 위한 학문인 유전체학이 발전하고 있으며, 유전체학 연구는 다양한 IT(information technology) 기술 분야와의 융복합 연구를 통해 생명공학 기술 패러다임을 변화시키고 있다(김용민, 2016). 요약하면, 생명과학 분야는 데이터 인용을 많이 받는 데이터 유형이 있었으며, 구체적으로 '양적 형질 유전자좌 지도 및 정보', '부동 소수점 숫자(4바이트)로 저장된 이미지'였다.

〈표 6〉은 생명과학 분야에서 데이터 공유, 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리를 보여준다. 생명과학 분야는 데이터 공유 별 인용이 높은 주요 데이터 리포지토리가 있었지만, z-score는 0으로 데이터 인용이 관례가 아니었다. 생명과학 분야의 데이터 공유 별 인용의 횟수는 STEM 분야 중에서 가장 높았다. '국제 토양 탄소 네트워크'는 1건의 데이터만 있어서 z-score가 null이다.

〈표 6〉 생명과학 분야의 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유	데이터 인용	데이터 공유 별 인용	데이터 인용 z-score	그룹 저자 수
1	Database of Genotypes and Phenotypes: dbGaP (유전형 및 표현형 데이터베이스: dbGaP)	626	26,816	42.84	0	0
2	European Centre for Medium-Range Weather Forecasts(유럽 중기 기상 예보센터)	75	2,829	37.72	0	66
3	NCAR Research Applications Laboratory(NCAR 연구 응용 연구실)	7	126	18	0	1
4	International Soil Carbon Network(국제 토양 탄소 네트워크)	1	13	13	null	0
5	UNITE	24	236	9.83	0	9
6	Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics(오크리지 국립 연구소 분산 활성 아카이브 생물지화학 역학센터)	255	2,305	9.04	0	10
7	Bioconductor	175	1,417	8.1	0	10
8	Broad Institute Genome Data Analysis Center (브로드 연구소 게놈 데이터 분석 센터)	8	60	7.5	0	8
9	OPA - Ocean Predictions and Applications (OPA-해양 예측 및 응용 프로그램)	30	198	6.6	0	0
10	Centre for Environmental Data Analysis(환경 데이터 분석 센터)	92	549	5.97	0	28

4.3 화학 분야

〈표 7〉은 화학 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 총 10만개의 데이터가 수집되었으며, 총 20개의 연구데이터 유형이 있었으며, 총 153,715회의 데이터 인용, 총 1,521회의 데이터 이용이 있었다. 데이터 인용을 많이 받은 연구데이터 상위 3위는 '결정구조'(73,836회, 73.84%), '결정구조: 결정학 정보'(21,477회, 21.48%), '부동 소수점 숫자(4바이트)로 저장된 이미지'(3,944회, 3.94%)였다. 상위 10위의 연구데이터 유형이 전체 데이터 인용의 99.97%를 받고 있었다. 1위를 차지한 결정 구조는 물질의 원자나 분자가 어떻게 배열되어 있는지를 나타내며, 원자, 이온 또는 분자가 3차원 공간에서 규칙적으로 배열된 구조를 의미한다. 2위를 차지한 결정학 정보는 결정체의 구조, 대칭성, 원자 배치, 물리적 및 화학적 특성을 기술하는 데이터를 의미한다. 3위를 차지한 '부동 소수점 숫자로 저장된 이미지'는 컴퓨터에서 실수를 표현하는 방식 중 하나로, 소수점이 고정되어 있지 않고 떠다니는 형태로 나타나며, 숫자의 크기에 따라 소수점이 이동할 수 있는 형식으로 컴퓨터가 실수를 표현할 때 유용하게 사용될 수 있는 데이터 유형이다. 요약하면, 화학 분야에서 데이터 인용이 주로 일어나는 데이터 유형은 '결정구조' 데이터였다.

〈표 8〉은 화학 분야에서 공유 별 데이터 인용을 많이 받은 상위 8위의 데이터 리포지토리를 보여준다. 화학 분야는 10만개 전체 데이터가 8개의 데이터 리포지토리에 모두 공유되

〈표 7〉 화학 분야의 데이터 공유별 인용수가 높은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	crystal structure(결정 구조)	73,836	73.84	155	47.55
2	crystal structure; crystallographic information(결정 구조; 결정학 정보)	21,477	21.48	135	41.41
3	image stored as floating point number(4 bytes)(부동 소수점 숫자(4바이트)로 저장된 이미지)	3,944	3.94	24	7.36
4	specialized mix; materials properties(특수 혼합; 재료 속성)	287	0.29	3	0.92
5	(비어 있음)	260	0.26	1	0.31
6	crystallographic data; crystal structure(결정학적 데이터; 결정 구조)	63	0.06	0	0
7	image stored as signed byte(부호화된 바이트로 저장된 이미지)	55	0.06	0	0
8	image stored as signed integer(2 bytes)(부호화된 정수(2바이트)로 저장된 이미지)	23	0.02	0	0
9	dataset(데이터 세트)	12	0.01	0	0
10	mass spectrometry data(질량 분석 데이터)	9	0.01	1	0.31
총합		99,966	99.97	319	97.86

〈표 8〉 화학 분야의 데이터 공유별 인용수가 높은 상위 8위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유 수	데이터 인용 수	데이터 공유별 인용 수	데이터 인용 z-score	그룹 저자 수
1	Cambridge Structural Database(케임브리지 구조 데이터베이스)	68,666	76,162	1.11	0	5
2	Crystallography Open Database(결정학 오픈 데이터베이스)	23,049	23,756	1.03	0	1
3	Electron Microscopy Data Bank(전자 현미경 데이터 뱅크)	5,641	6,294	1.12	0	13
4	Carbohydrate Structure Database(탄수화물 구조 데이터베이스)	2,517	2,528	1	0	0
5	The Materials Project(재료 프로젝트)	112	295	2.63	0	92
6	Mass Spectrometry Interactive Virtual Environment(질량 분석 대화형 가상 환경)	8	16	2	0	1
7	Chemotion	4	16	4	0	0
8	SBGrid Data Bank(SBGrid 데이터 뱅크)	3	4	1.33	0	0

어 있었다. 화학 분야에서 데이터 공유가 가장 높은 데이터 리포지토리의 순위는 Cambridge Structural Database(68,666회), Crystallography Open Database(23,049회), Carbohydrate Structure Database(2,517회)이었다. 1위인 '케임브리지 구조 데이터베이스'는 1920년대부터 구조데이터를 수집해 왔으며, 유기화합물, 금속 유기 화합물에 대한 결정학적 정보를 제공하여 연구자들이 화합물의 구조를 이해하고 분석하는 데 도움을 주는 리포지토리이다(Cambridge Crystallographic Data Centre, n.d.). 2위인 '결정한 오픈 데이터베이스'는 연구자들에게 무

료로 공개된 결정구조 정보를 제공하며, CIF (Crystallographic Information File)의 형태로 데이터가 제공된다(Vilnius University, n.d.). 3위인 '전자 현미경 데이터 뱅크'는 전자현미경으로 얻은 3차원 구조 데이터와 거대 분자 복합체 및 세부 하포 구조의 대표적 단층 촬영도에 대한 리포지토리이다(EMBL-EBI, 2025). 요약하면, 화학분야 데이터 리포지토리는 데이터 공유가 주로 일어나는 리포지토리는 있지만, 높은 데이터 공유가 높은 인용을 의미하지는 않았다. 데이터 인용 z-score는 모두 0으로 데이터 인용이 관례인 데이터 리포지토리는 없었다.

4.4 컴퓨팅 분야

〈표 9〉는 컴퓨팅 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 컴퓨팅 분야는 연구데이터의 공유가 활발하지 않았는데, Data Citation Index에서 색인되는 전체 연구데이터 수가 10,148개였다. 총 138개의 데이터 유형 중에서 2,298개의 데이터 인용, 204회의 데이터 이용이 있었다. 데이터 인용을 많이 받는 상위 3개의 연구데이터의 유형은 '비어 있음'(1,256회), '뉴런 또는 기타 전기적으로 흥분되는 세포; 컴퓨터 모델; 소프트웨어'(180회), '소프트웨어'(157회)였다. 상위 10개의 데이터 유형이 전체 데이터 유형의 89.64%를 차지하고 있었다. 컴퓨팅 분야의 데이터는 급속히 고성능화 되고 있는 슈퍼 컴퓨터와 함께 양자역학 기반의 제일원리 전산 프로그램의 발달이 머신 러닝과 연동되어, 광범위한 나노스케일 소재 후보군에 대한 고속 스크리닝 같은 새로운 연구개발 패러다임을 구축하고 있다(강준희, 한병찬, 2019). 빅데이터 시대에 접어들면서 컴퓨팅 분야에서는 방대한 양의 데이터를 효율적으로 저장, 처리, 분석하는 기술이 중

〈표 9〉 컴퓨팅 분야의 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	(비어 있음)	1,256	54.66	119	58.33
2	neuron or other electrically excitable cell; computer Model; software(뉴런 또는 기타 전기적으로 흥분되는 세포; 컴퓨터 모델; 소프트웨어)	180	7.83	0	0
3	software(소프트웨어)	157	6.83	19	9.31
4	dataset/neurophysiology(데이터 세트/신경 생리학)	137	5.96	8	3.92
5	realistic network; computer model; software(현실적인 네트워크; 컴퓨터 모델; 소프트웨어)	104	4.53	1	0.49
6	computer model; software(컴퓨터 모델; 소프트웨어)	80	3.48	3	1.47
7	code; software(코드; 소프트웨어)	58	2.52	21	10.29
8	software; model(소프트웨어; 모델)	43	1.87	6	2.94
9	synapse; computer model; software(시냅스; 컴퓨터 모델; 소프트웨어)	25	1.09	1	0.49
10	code(코드)	20	0.87	11	5.39
총합		2,060	89.64	189	92.63

요해졌다. 이는 분산 컴퓨팅, 클라우드 컴퓨팅 등의 기술 발전을 촉진하였다. 요약하면, 컴퓨팅 분야는 연구데이터 유형을 특정 짓지 않은 경우의 데이터 인용이 가장 높았으며, 상위 10위의 연구데이터 유형이 전체 데이터 인용의 약 90%를 차지하고 있었다. 컴퓨팅 분야에서는 구조적(structured), 반구조적(semi-structured) 데이터가 많이 사용되고 있었다. 인공지능을 학습할 수 있는 데이터가 많이 공유되어, 머신러닝의 학습데이터로 많이 활용되고 있다고 해석될 수 있다.

〈표 10〉은 컴퓨팅 분야에서 공유 별 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리를 보여준다. 총 6개의 데이터 리포지토리가 컴퓨팅 분야 전체의 데이터 공유를 받고 있었다. 'GroupLens 데이터 세트'는 z-score가 0.35로, STEM분야 전체에서 z-score가 0이 아닌 유일한 데이터 리포지토리였다. GroupLens는 머신러닝을 학습시킬 수 있는 방대한 데이터를 무료로 제공하고 있으므로, 인용 및 재이용이 높다고 해석될 수 있다. 요약하면, 컴퓨팅 분야의 데이터 리포지토리는 데이터 인용이 관례가 아니지만, 'GroupLens 데이터 세트'는 데이터 인용이 상대적으로 높은 데이터 리포지토리이다. 컴퓨팅 분야의 경우, 다른 학문분야와 다르게 기관 리포지토리(institutional repository), Zenodo와 Figshare 등 일반 리포지토리(general repository)가 상위에 랭크된 경우가 있었다. 인공지능 기술이 대두되면서, 컴퓨팅 분야에 신경과학과 관련된 데이터 리포지토리가 상위에 랭크하고 있는 것으로 해석될 수 있다.

4.5 지구과학 분야

〈표 11〉은 지구과학 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여

〈표 10〉 컴퓨팅 분야의 데이터 공유별 인용수가 높은 상위 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유 수	데이터 인용 수	데이터 공유별 인용 수	데이터 인용 z-score	그룹 저자 수
1	4TU.Centre for Research Data(4TU 연구데이터 센터)	3	16	5.33	0	0
2	GroupLens Datasets(GroupLens 데이터세트)	2	7	3.5	0.35	0
3	UCI Machine Learning Repository(UCI 머신러닝 리포지토리)	232	603	2.6	0	15
4	Mantid Project(Mantid 프로젝트)	42	46	1.1	0	1
5	Collaborative Research in Computational Neuroscience (CRCNS.org)(계산 신경 과학 협력 연구)	130	140	1.08	0	1
6	ModelDB	1,052	610	0.58	0	0
7	University of Southampton Institutional Repository (사우스햄튼 대학교 기관 리포지토리)	3,889	596	0.15	0	698
8	Knowledgebase of Interatomic Models(원자간 모델에 대한 지식기반)	1,218	153	0.13	0	0
9	Zenodo	353	28	0.08	0	37
10	Figshare	2,086	84	0.04	0	14

〈표 11〉 지구과학 분야의 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	(비어 있음)	54,555	81.39	1118	70.01
3	digital table(디지털 표)	3,132	4.67	80	5.01
2	specialized Mix(특수 믹스)	2,940	4.39	160	10.02
4	dataset(데이터세트)	1,689	2.52	44	2.76
5	digital(디지털)	1,644	2.45	11	0.69
6	data collection(데이터 컬렉션)	794	1.18	6	0.38
7	digital image(디지털 이미지)	418	0.62	5	0.31
8	geoscientific Information(지구과학 정보)	387	0.58	69	4.32
9	software(소프트웨어)	370	0.55	8	0.5
10	digital map(디지털 지도)	333	0.5	4	0.25
총합		66,262	98.85	1505	94.25

준다. 지구과학 분야는 총 99,493개의 연구데이터가 수집되었고, 총 67,029회의 데이터 인용이 있었다. 상위 10위의 데이터 유형이 전체 데이터 인용의 98.85%를 차지하고 있었다. 지구 과학 분야의 연구데이터는 총 54개의 유형이 있었다. 데이터 인용을 많이 받은 상위 3위의 연구데이터 유형은 '비어 있음'(54,555회, 81.39%), '디지털 표'(3,132회, 4.67%), '특수 믹스'(2,940회, 4.39%)였다. 2위를 차지한 '디지털 표'는 디지털 형태의 표 데이터로, 주로 수치나 속성 정보를 제공하는데, 예시로는 지진 관측소 데이터의 진폭, 시간, 위치를 나타내는 CSV(Comma-separated Value) 파일 등이 있다. 3위를 차지한 '특수 믹스'는 시간, 공간, 기후, 지질 등 여러 유형의 데이터를 통합하여 특정 목적에 맞게 결합한 데이터로, 예시로는 화산 분출 데이터를 대기 오염 데이터와 결합하여 화산재 확산 시뮬레이션으로 표현한 데이터가 있다(Wang et al., 2024). 6위를 차지한 '데이터 컬렉션'은 측정 장비나 센서를 통해 지속적으로 수집되는 데이터로, 예시로는 지구의 자기장 데이터를 측정하는 위성 관측 데이터 등이 있다. 요약하면, 지구과학 분야의 연구데이터는 대부분 데이터 유형을 특정 짓지 않고 공유 및 인용되고 있었다. 지구과학 분야의 데이터 리포지토리는, 일별, 지역별 데이터가 방대하고 장기간에 걸친 데이터라는 특징이 있다. 데이터 활용에 대한 메타데이터의 상세 설명이 제공된다면, 시각화 등의 활용에 있어서 좀 더 편리할 수 있다. 국제적인 기관이나 국가에서 리포지토리를 활용하는 경우라면, 신뢰할 수 있는 기관에서 제공할 경우 연구데이터가 좀 더 많이 활용되고 있는 특징이 있었다. '디지털 표' 등의 엑셀 혹은 스프레드시트 유형을 가진 정형화된 데이터가 많이 활용되고 있었다.

〈표 12〉는 지구과학 분야에서 공유 별 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리를 보여준다. 지구과학 분야는 총 99,493개의 연구데이터가 수집되었으며, 총 67,029회의 데이터 인용이 있었다. 데이터 공유 별 인용수가 높은 상위 3위의 데이터 리포지토리는 '유럽 중기 기상 예보 센터'(20.73회), 'TCCON 데

〈표 12〉 지구과학 분야의 데이터 공유별 인용수가 높은 상위 10위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유	데이터 인용	데이터 공유 별 인용 수	데이터 인용 z-score	그룹 저자 수
1	European Centre for Medium-Range Weather Forecasts(유럽 중기 기상 예보 센터)	136	2,819	20.73	0	122
2	TCCON Data Archive(TCCON데이터 아카이브)	17	310	18.24	0	0
3	NCAR Research Applications Laboratory(NCAR 연구 응용 연구실)	7	127	18.14	0	1
4	Computational and Information Systems Laboratory Research Data Archive(계산 및 정보 시스템 연구실 연구데이터 아카이브)	167	2,599	15.56	0	126
5	World Glacier Monitoring Service(WGMS)(세계 빙하 모니터링 서비스)	14	201	14.36	0	3
6	CaltechDATA	21	287	13.67	0	0
7	VLIZ - Flanders Marine Institute(VLIZ-플랜더스 해양 연구소)	196	1,594	8.13	0	134
8	KNB Data Repository(KNB 데이터 리포지토리)	1	8	8	null	0
9	Climate Hazards Group Data Archive(기후 위험 그룹 데이터 아카이브)	2	14	7	0	2
10	British Oceanographic Data Centre(영국 해양학 데이터 센터)	162	977	6.03	0	8

이터 아카이브'(18.24회), 'NCAR 연구 응용 연구실'(18.24회)이었다. 요약하면, 지구과학 분야 데이터 리포지토리는 데이터 공유별로 인용이 주로 일어나는 주요 데이터 리포지토리가 있었지만, 데이터 인용 z-score는 0으로 데이터 인용이 관례인 데이터 리포지토리는 없었다. 지구과학 분야는 타 학문 분야에 비하여 그룹저자가 많았다.

4.6 공학 분야

〈표 13〉은 공학 분야에서 공유 별 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 공학 분야는 총 97,806개의 연구데이터에서 4,519회의 데이터 인용이 있었다. 총 12개의 데이터 유형이 전체 연구데이터를 구성하고 있었다. 데이터 인용을 많이 받은 상위 3위의 연구데이터 유형은 '비어있음'(3,437회, 76.06%), '특수 믹스; 재료 속성'(562회, 12.44%), '특수 믹스'(298회, 6.59%)였다. 요약하면, 공학 분야의 연구데이터는 대부분 연구데이터의 유형을 특정 짓지 않고 공유 및 인용되고 있었으며, 상위 10위의 연구데이터 유형이 99.95%의 데이터 인용을 받고 있었다. 공학 분야는 타 STEM 분야에 비하여, 데이터 공유 대비 데이터 인용의 수가 가장 낮았다. 공학 분야는 재료와 관련된 연구데이터가 많이 인용이 되었는데, 그래픽카드, 반도체 등이 재료와 관련이 있어서, 상향산업이기 때문일 수 있다. 전통적으로 공학 분야는 건물, 다리 건축 등이 하향산업이라고 인식되는 경향이 있다. 하지만, 반도체, 인공지능 등의 발달로 인하여, 상위에 랭크된 연구데이터가 재료와 관련되었을 수 있다.

〈표 14〉는 공학 분야에서 공유 별 데이터 인

〈표 13〉 공학 분야의 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	(비어 있음)	3,437	76.06	301	51.02
2	specialized mix; materials properties(특수 믹스; 재료 속성)	562	12.44	174	29.49
3	specialized mix(특수 믹스)	298	6.59	109	18.47
4	materials testing report(재료 테스트 보고서)	151	3.34	0	0
5	numeric data(숫자 데이터)	36	0.8	2	0.34
6	test data(실험 데이터)	18	0.4	1	0.17
7	specialized mix; structure & properties(특수 믹수; 구조 및 속성)	10	0.22	1	0.17
8	direct download file(직접 다운로드 파일)	2	0.04	0	0
9	specialized mix; materials properties; materials computations (특수 믹스; 지료 속성; 재료 계산)	2	0.04	0	0
10	catalog(카탈로그)	1	0.02	0	0
총합		4,517	99.95	588	99.66

〈표 14〉 공학 분야의 데이터 공유별 인용수가 높은 상위 9위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유 수	데이터 인용 수	데이터 공유 별 인용 수	데이터 인용 z-score	그룹 저자 수
1	MATDAT Materials Properties Database (MATDAT 재료 속성 데이터베이스)	1,193	1,348	1.13	0	0
2	Prognostics Center of Excellence Data Repository (예후 센터 우수성 데이터 리포지토리)	3	3	1	null	0
3	NREL Data Catalog(NREL 데이터 카탈로그)	128	97	0.76	0	1
4	DesignSafeCI	1,533	697	0.45	0	2
5	IEEE DataPort	12,155	1,511	0.12	0	370
6	MatDB	2,078	19	0.01	0	1,008
7	Most Wiedzy Open Research Data Catalog(Most Wiedzy 오픈 연구 데이터 카탈로그)	2,129	26	0.01	0	2
8	The Materials Project(재료 프로젝트)	78,586	818	0.01	0	78,507
9	Universidad de Oviedo(UNIOVI), Spain; Queens University Belfast, Northern Ireland	1	0	0	null	0

용을 많이 받은 상위 9위의 데이터 리포지토리를 보여준다. 공학분야는 9개의 리포지토리에 모든 연구데이터가 공유되어 있었다. 공학 분야는 데이터 공유 별 인용 수는 높지 않았다. 공학 분야의 데이터 리포지토리는 인용 횟수가 상대적으로 높은 주요 리포지토리는 없었다. 공학분야는 데이터 인용 z-score는 모두 0으로, 데이터 인용이 관례인 데이터 리포지토리는 없었다. 학문 특화된 데이터 리포지토리가 많았다.

4.7 수학 분야

〈표 15〉는 수학 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여

〈표 15〉 수학 분야의 데이터 인용을 많이 받은 상위 연구데이터 유형

순위	연구데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	software(소프트웨어)	140,593	99.85	1,815	96.49
2	geoid undulation given on a grid(격자에 주어진 지오이드 기복)	215	0.15	66	3.51
총합		140,808	100	1,881	100

준다. 수학 분야는 총 24,593개의 연구데이터가 공유되어 있었으며, 총 140,808회의 데이터 인용이 있어서, 타 STEM 학문 분야에 비하여 데이터 인용이 굉장히 높았다. 데이터 인용을 받은 연구데이터는 2가지 유형만 있었는데, 1위는 '소프트웨어'(140,593회, 99.85%), '격자에 주어진 지오이드 기복'(215회, 0.15%)이었다. 특히, '소프트웨어'는 전체 데이터 인용의 99.85%를 차지하고 있었다. 요약하면, 수학 분야의 연구데이터는 데이터 유형이 2가지 밖에 없었으며, '소프트웨어' 유형이 전체 데이터 인용의 99.85%를 차지하고 있었다. 수학 분야는 소프트웨어 등 수학적 또는 공학적 계산을 위해 사용되는 데이터, 지오이드 등 위성데이터나 GPS(Global Positioning System)등을 통해서 수집이 되는 데이터 유형이 있었다. 즉, 다른 프로그래밍 데이터처럼 수학 또는 공학적인 계산을 위해 사용되고 있다고 할 수 있다.

〈표 16〉은 수학 분야에서 데이터 공유 별 인용을 많이 받은 상위의 데이터 리포지토리를 보여준다. 수학 분야는 총 3개의 데이터 리포지토리에 모든 연구데이터가 공유되고 있었다. 데이터 인용 z-score는 0으로 데이터 인용이 관례인 리포지토리는 없었다. CRAN은 수학적인 프로그래밍 코드를 재현할 수 있어서 상위 리포지토리로 랭크된 것으로 해석될 수 있다. '지오이드를 위한 국제 서비스'는 지오이드 모델의 연구나 개발을 담당하는데, 전세계적인 지역 모델을 제공하고, 중력장 연구 또는 지리정보시스템을 활용할 수 있도록 지원한다. 'SuiteSparse 매트릭스 컬렉션'은 희소 행렬을 포함한 리포지토리인데, 과학 혹은 수학적 계산을 할 때 연구자가 주로 활용하는 리포지토리이다. 요약하면, 수학적인 프로그래밍 코드의 재현, 측지학, 희소행렬 등 수학 또는 공학분야에서 주로 활용되는 데이터 리포지토리였다.

〈표 16〉 수학 분야의 데이터 공유별 인용수가 높은 상위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유	데이터 인용	데이터 공유 별 인용 수	데이터 인용 z-score	그룹 저자 수
1	Comprehensive R Archive Network(CRAN, 종합 R 아카이브 네트워크)	24,272	140,593	5.79	0	0
2	International Service for the Geoid(지오이드를 위한 국제 서비스)	96	215	2.24	0	0
3	SuiteSparse Matrix Collection(SuiteSparse 매트릭스 컬렉션)	225	0	0	null	0

4.8 기술 분야

〈표 17〉은 기술 분야에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 기술 분야는 총 99,996개의 연구데이터가 수집되었으며, 총 데이터 인용 수는 293,240회로, 데이터 인용이 높은 상위 10개의 연구데이터가 전체 데이터 인용의 98.97%를 차지하고 있었다. 총 974개의 데이터 유형이 있었으며, 데이터 인용을 많이 받은 상위 3위의 유형은 '소프트웨어'(201,749회, 68.8%), '비어 있음'(56,278회, 19.19%), 데이터세트(28,054, 9.57%)였다. 기술분야 데이터의 특징은, 테이블 형식으로 표현하기에는 데이터 유형이 너무 다양하고 데이터 구조가 복잡하다는 점이다. 새로운 실험 기법 또는 실험 장비가 지속적으로 개발되기 때문에, 데이터 모델이 지속적으로 확장되어야 할 필요가 있다. 요약하면, 기술 분야의 연구데이터 데이터 공유 대비 전체 인용 횟수가 타 STEM 분야에 비하여 상대적으로 높았으며, 데이터 인용이 주로 일어나는 데이터 유형은 소프트웨어였다.

〈표 18〉은 기술 분야에서 공유 별 데이터 인용을 많이 받은 상위 10위의 데이터 리포지토리를 보여준다. 데이터 리포지토리의 평균적인 데이터 공유별 인용 수는 2.93회였다. 요약하면, 기술 분야는 CRAN이 데이터 공유 및 인용이 높다. 하지만, 데이터 인용 z-score는 0이므로, 데이터 인용이 관례인 리포지토리는 없었다. 기술분야는 데이터 공유별 인용수가 높은 리포지토리는 개별 저자가 대부분이었고, 그룹 저자는 많지 않았다.

4.9 STEM 분야 전체

〈표 19〉는 STEM분야 전체에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형을 보여준다. 총 630,935개의 데이터가 수집되었으며, 총 783,928회의 데이터 인용, 630,807회의 데이터 이용이 있었다. STEM분야 전체에서 상위 10개의 데이터 유형이 전체 데이터 인용의

〈표 17〉 기술 분야의 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형

순위	데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	software(소프트웨어)	201,749	68.8	2,861	40.21
2	(비어 있음)	56,278	19.19	2,716	38.17
3	dataset(데이터 세트)	28,054	9.57	1,221	17.16
4	project(프로젝트)	2,065	0.7	121	1.7
5	figure(이미지)	485	0.17	12	0.17
6	fileset(파일 세트)	464	0.16	33	0.46
7	image(이미지)	330	0.11	11	0.15
8	experimental data(실험 데이터)	293	0.1	0	0
9	audiovisual(시청각)	278	0.09	7	0.1
10	capsule: software(캡슐: 소프트웨어)	224	0.08	5	0.07
총합		290,220	98.97	6,987	98.19

〈표 18〉 기술 분야의 데이터 공유별 인용수가 높은 상위 10위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유 수	데이터 인용 수	데이터 공유별 인용 수	데이터 인용 z-score	그룹 저자 수
1	Comprehensive R Archive Network(종합 R 아카이브 네트워크, CRAN)	6,822	141,070	20.68	0	38
2	Peking University Open Research Data Platform (북경대 오픈 연구데이터 플랫폼)	21	170	8.1	0	10
3	CaltechDATA	162	1,011	6.24	0	3
4	CIRcle	1	5	5	null	0
5	data.bris Research Data Repository(data.bris 연구데이터 리포지토리)	86	270	3.14	0	0
6	ORA - Oxford University Research Archive(옥스포드 대학교 연구 아카이브)	12	36	3	0	1
7	Oak Ridge Leadership Computing Facility(OLCF) Constellation Portal(오크리지 리더십 계산 시설 별자리 포털)	13	36	2.77	0	1
8	DesignSafeCI	15	40	2.67	0	4
9	LINDAT/CLARIN Centre for Language Research Infrastructure Digital Repository(LINDAT/CLARIN 언어 연구 인프라 센터 디지털 리포지토리)	25	64	2.56	0	1
10	University of Reading Research Data Archive(리딩대학교 연구데이터 아카이브)	59	147	2.49	0	2

〈표 19〉 STEM 분야 전체에서 데이터 인용을 많이 받은 상위 10위의 연구데이터 유형

순위	데이터 유형	데이터 인용		데이터 이용	
		횟수	퍼센트	횟수	퍼센트
1	software	344,829	43.99	4,726	0.75
2	(비어 있음)	178,396	22.76	251,159	39.82
3	crystal structure	76,092	9.71	155	0.02
4	dataset	30,395	3.88	1,287	0.20
5	quantitative trait locus map & information	30,348	3.87	2	0.00
6	crystal structure; crystallographic information	23,745	3.03	135	0.02
7	longitudinal	17,930	2.29	1	0.00
8	Image stored as floating point number(4 bytes)	16,883	2.15	56	0.01
9	mass spectral data	6,070	0.77	2	0.00
10	Specialized Mix	5,819	0.74	398	0.06
총합		730,507	93.19	253,195	40.88

93.19%를 차지하고, 데이터 이용은 40.88%를 차지하고 있었다. 이는, STEM분야에서는 연구자들이 선호하는 주요 데이터 유형이 있는 것으로 해석될 수 있다.

〈표 20〉은 STEM분야 전체에서 데이터 공유별 인용수가 높은 상위 10위의 데이터 리포지토리를 분석한 것이다. 데이터 공유별 인용수가 높은 주요 리포지토리가 발견되었으나, 데이터

〈표 20〉 STEM분야 전체에서 데이터 공유별 인용수가 높은 상위 10위의 데이터 리포지토리

순위	데이터 리포지토리 명	데이터 공유 수	데이터 인용 수	데이터 공유별 인용 수	데이터 인용 z-score	그룹 저자 수
1	Database of Genotypes and Phenotypes: dbGaP (유전형 및 표현형 데이터베이스)	626	26,816	42.84	0	3
2	European Centre for Medium-Range Weather Forecasts(유럽 중기 기상 예보센터)	211	5,648	26.77	0	188
3	TCCON Data Archive(TCCON 데이터 아카이브)	17	310	18.24	0	0
4	NCAR Research Applications Laboratory(NCAR 연구 응용 연구실)	14	253	18.07	0	2
5	Computational and Information Systems Laboratory Research Data Archive(계산 및 정보시스템 연구실 연구데이터 아카이브)	167	2,599	15.56	0	126
6	World Glacier Monitoring Service(세계 빙하 모니터링 서비스)	14	201	14.36	0	3
7	International Soil Carbon Network(국제 토양 탄소 네트워크)	1	13	13	null	0
8	UNITE	24	236	9.83	0	9
9	Comprehensive R Archive Network(종합 R 아카이브 네트워크: CRAN)	31,094	281,663	9.06	0	38
10	VLIZ - Flanders Marine Institute(VLIZ-플랜더스 해양연구소)	196	1,594	8.13	0	134

인용 z-score로 표준화 한 값은 모두 0이었다. 이는, 데이터 공유가 적은 데이터 리포지터리가 데이터 인용이 높은 것으로 보일 수 있지만, 표준화 한 데이터 인용값은 0으로 데이터 리포지토리별 차이는 없음을 보여준다. 데이터 인용이 관례인 데이터 리포지토리는 없었다.

5. 논 의

데이터 공유는 학문 분야별로 관례가 다르다. 학문 분야별로 데이터 생산과 이용에 차이가 있으므로(Mongeon et al., 2017), 데이터 공유가 적다는 것이 반드시 데이터의 생산이 적다는 의미는 아닐 수 있다. STEM 분야 간 데이터 공유 빈도가 상이하다는 사실은 데이터 공유 관례를 어떻게 촉진할 것인지에 대한 중요한 질문을 제기한다. 하나의 접근 방법은 데이터 공유자에게 데이터 인용 등의 공식적인 학술 크레딧(credit)을 부여하는 것이다. 데이터 인용 등의 학술 크레딧은 승진 심사나 정년 보장 등에 활용될 수 있다. Andreoli-Versbach와 Mueller-Langer(2014)는 데이터 공유를 필수적으로 요구하는 저널에 게재된 논문의 인용 빈도가 증가함을 확인했다. 이러한 결과는 데이터 공유 촉진을 위한 공식적인 크레딧 부여 방안의 타당성을 뒷받침한다. 또한, 데이터 공유의 표준을 통합된 틀 안에서 정리하는 것은 연구 기관들이 전세계적으로 협력하여 사이버 인프라스트럭처를 구축하고 이를 활용함으로써 데이터 공유의 조화를 이루는 데 중요한 역할을 할 수 있다. 이를 통해 글로벌 연

구센터들은 데이터 공유의 효율성을 높이고, 다양한 국가와 지역 간의 협력을 촉진할 수 있다.

데이터 공유자에게는 데이터를 보존하고 공식적인 학술 크레딧을 받을 수 있는 주요 데이터 리포지토리의 선택이 중요한 문제일 수 있다. 특히, 학문 분야별 리포지토리는 해당 학문 분야의 데이터와 맞춤형 서비스를 제공하여 검색의 효율성을 높이는 장점이 있다. 리포지토리 이용자는 각 리포지토리가 사용하는 통제어와 주제 카테고리에 대한 이해를 통하여, 데이터를 적절히 관리하고 공유할 수 있어야 한다. 또한, 법적, 정책적 측면에서 데이터 공유에는 지적 재산권이나 개인정보보호 문제에 대한 고려가 필수적이다. 예를 들어, 특허, 저작권, 독점 정보 등의 상업 비밀 등은 데이터 공유에 있어 중요한 제약사항 일 수 있다. 기업의 연구에서 생산된 독점 데이터는 미래의 특허와 연관될 수 있으며, 이는 데이터 공유 시에 법적인 위험을 초래할 수 있다. 또한, 법적, 윤리적 관점에서 데이터 공유는 데이터 프라이버시와 보안의 문제를 동반할 수 있다. 이러한 문제는 데이터 리포지토리의 기술적 인프라와 밀접하게 연관되며, 데이터가 잘못 처리될 경우 연구 참가자들의 신뢰를 잃을 위험이 있다. 데이터 리포지토리는 각국의 데이터 규제, 특히 건강 데이터를 다루는 규제를 준수해야 한다. 데이터를 다운로드할 때에는 리포지토리의 이용 약관을 준수하고, 데이터 배포 및 조건에 대한 규정을 명확히 해야 한다. 예를 들어, Creative Commons나 기타 라이선스 조건에 따라 데이터를 사용하고, 추가적인 제한을 두거나, 데이터 가용성에 대한 명시적 설명을 포함하는 등의 절차가 필요하다.

본 연구는 데이터 공유수가 적을 때 인용수가 높게 보이거나, 데이터 공유수가 많은 리포지토리는 인용수가 낮아 보일 가능성이 있음을 확인하였다. 이를 보완하기 위하여 데이터 인용의 Z-Score를 추가로 분석하였다. Springer Nature(n.d.)는 공식 웹사이트에 학문 분야별, 데이터 유형별로 추천하는 데이터 리포지토리 목록을 제공하고 있으며, 이는 학문 분야별 주요 리포지토리(major repositories)로 해석될 수 있다. 데이터 공유자는 연구데이터의 가시성을 위하여 주요 리포지토리에 기탁하는 것을 고려할 수 있다. 하지만, 데이터 공유자가 원하는 것이 연구데이터의 가시성이 아닌, 인용 등의 공식적인 학술 크레딧이라면, 데이터 인용의 Z-score를 기반으로 데이터 인용이 높은 리포지토리에 연구데이터를 공유할 필요가 있다.

본 연구는 데이터 공유 및 인용이 주로 일어나는 데이터 리포지토리는, 학문 특화된 리포지토리(discipline-specific repository)가 대부분이며, Zenodo, Figshare등의 일반 리포지토리(general repositories)가 일부 있고, 기관 리포지토리(institutional repository)는 거의 없음을 확인하였다. 사회과학 분야는 ICPSR(Inter-University Consortium for Political and Social Research) 등의 사회과학 분야 전체를 아우르는 주요 리포지토리가 있다. 하지만, STEM 분야는 다양한 학문 특화된 리포지토리들이 존재하며, 사회과학 분야처럼 분야 전체를 대표하는 리포지토리는 없음을 알 수 있었다. 이는, STEM분야 자체가 세부적인 데이터 유형이 중요하기 때문일 수 있다.

6. 결 론

본 연구의 목적은 데이터 인용이 높은 STEM 분야 연구데이터의 유형 및 데이터 리포지토리를 살펴보는 것이다. STEM 분야는 국립과학재단의 학문분야 코드와 Data Citation Index의 연구분야를 기준으로 총 8개로 분류하였다. 8개의 STEM 분야는 천체물리학, 생명과학, 화학, 컴퓨팅, 지구과학, 공학, 수학, 기술이었다. 각 학문 분야별로 Data Citation Index에서 데이터 인용 횟수가 가장 높은 순서로 레코드를 다운로드 받아서, 총 630만개 이상의 연구데이터를 직접 수집 및 분석하였다.

본 연구는 STEM 분야에서 다양한 종류의 연구데이터 유형이 인용을 받음을 확인하였다. 데이터 인용을 받은 연구데이터는 모두 양적 데이터였다. 데이터 인용을 받은 질적 데이터는 없었다. 데이터 인용을 받은 연구데이터 유형은 STEM 학문 분야별로 다양하였다. 천체물리학 분야는 연구데이터 유형을 설정하지 않고 공유 및 인용되는 경우가 46.16%였으며, 데이터 인용이 주로 일어나는 데이터 유형은 '질량 스펙트럼 데이터'였다. 천체물리학 분야는 데이터 공유가 주로 일어나는 데이터 리포지토리인 MassBank가 있지만, 높은 데이터 공유가 반드시 높은 데이터 인용으로 이어지지는 않았다. 생명과학 분야는 데이터 인용을 많이 받는 데이터 유형이 있었으며, 구체적으로 '양적 형질 유전자좌 지도 및 정보', '부동 소수점 숫자(4바이트)로 저장된 이미지'였다. 생명과학 분야의 데이터 공유 별 인용의 횟수는 STEM 분야 중에서 가장 높았다. 화학 분야에서 데이터 인용이 주로 일어나는 데이터 유형은 '결정구조' 데이터였다. 컴퓨팅 분야는 연구데이터 유형을 특정 짓지 않은 경우의 데이터 인용이 가장 높았으며, 상위 10위의 연구데이터 유형이 전체 데이터 인용의 약 90%를 차지하고 있었다. 지구과학 분야의 연구데이터는 대부분 데이터 유형을 특정 짓지 않고 공유 및 인용되고 있었다. 공학 분야의 연구데이터는 대부분 연구데이터의 유형을 특정 짓지 않고 공유 및 인용되고 있었으며, 상위 10위의 연구데이터 유형이 99.95%의 데이터 인용을 받고 있었다. 공학 분야는 타 STEM 분야에 비하여 데이터 공유 대비 데이터 인용의 수가 가장 낮았다. 공학 분야의 데이터 리포지토리는 인용 횟수가 상대적으로 높은 주요 리포지토리는 없었다. 수학 분야의 연구데이터는 데이터 유형이 2가지 밖에 없었으며, '소프트웨어' 유형이 전체 데이터 인용의 99.85%를 차지하고 있었다. 기술 분야의 연구데이터 데이터 공유 대비 전체 인용수가 타 STEM 분야에 비하여 상대적으로 높았으며, 데이터 인용이 주로 일어나는 데이터 유형은 소프트웨어였다. STEM 분야의 데이터 리포지토리는 '데이터 공유 별 인용'에 있어서는 데이터 인용이 높은 리포지토리들이 다수 확인되었다. 하지만, 데이터 인용을 표준화한 z-score 값은 대부분 0이었으며, 0이 아닌 리포지토리는 GroupLens Datasets(GroupLens 데이터세트)가 0.35로 유일했다. 즉 공식적인 데이터 인용(formal data citation)은 STEM 분야의 학술 커뮤니티의 관례가 아니다. 데이터 공유 및 인용이 주로 일어나는 리포지토리는 학문특화 리포지토리(discipline-specific repositories)이며, 일반 리포지토리(general repositories)는 일부 발견되고, 기관 리포지토리(institutional repositories)

는 거의 없음을 확인하였다. STEM분야의 주요 리포지토리에서는 그룹저자가 많지 않지만, 지구과학, 공학 분야는 그룹저자를 통한 연구데이터의 공유 및 인용이 활발함을 확인하였다.

본 연구의 한계는, 비록 전세계 450개 이상의 데이터 리포지토리에서 추적 및 색인되는 연구데이터를 살펴보고, 공식적인 데이터 인용(formal data citation)을 분석할 수 있었지만, Data Citation Index에서 색인하지 않는 연구데이터는 살펴볼 수 없었다는 점이다. 데이터 이용의 경우 2013년 이후의 이용수만 분석했는데, 이는 수집된 데이터의 한계이다. 향후 연구는 첫째, 더 많은 데이터 리포지토리에 공유된 연구데이터를 분석하고, 둘째, 사회과학, 의학 분야 등 보다 다양한 학문 분야의 연구데이터로 분석을 확장하는 것이다. 본 연구의 공헌은 630만개 이상의 연구데이터를 직접 수집하여, STEM 분야의 연구데이터의 관례를 실질적으로 살펴보았다는 것이다. 특히, 본 연구의 결과가 학술 커뮤니티에 제시하는 시사점은, 높은 데이터 공유 및 데이터 이용이 높은 데이터 인용 등의 학술 크레딧을 의미하지는 않는다는 점을 밝혔다는 점이다. 즉, 연구자가 자신의 연구데이터를 공유할 때, 연구데이터의 가시성을 염두에 두고 데이터 공유와 이용이 활발한 주요 리포지토리(major repositories)에 연구데이터를 기탁해왔을 수 있다. 하지만, 연구자가 인용 등의 공식적인 학술 크레딧을 염두에 둔다면, 공유와 이용이 활발한 주요 리포지토리에 기탁하는 것보다는, 데이터 인용이 높은 리포지토리를 선택할 수 있다는 점을 밝혔다는 점에서 본 연구가 학술 커뮤니티에 제시하는 시사점이 있다.

참 고 문 헌

강준희, 한병찬 (2019). 멀티스케일 계산을 위한 제일원리 전산 데이터 기반 머신 러닝 포텐셜 개발. 한국공업화학회, 22(4), 13-19.

김용민 (2016). 유전체 빅데이터 연구 동향. 분자세포생물학 뉴스레터, 1-8.

김주섭, 강효숙, 김선태 (2023). 생태 분야 데이터 리포지터리 운영 현황 분석 및 EcoBank 서비스 제안. 한국문헌정보학회지, 57(4), 289-310. https://doi.org/10.4275/KSLIS.2023.57.4.289

박형주 (2024). 생명과학분야 학술논문의 전문에 나타난 연구데이터의 공유, 재이용, 인용 분석. 한국문헌정보학회지, 58(4), 335-353. https://doi.org/10.4275/KSLIS.2024.58.4.335

송영옥, 김성영, 장덕진 (2010). 바이오 데이터 패턴 분석을 위한 시스템 및 알고리즘 설계. 한국콘텐츠학회 논문지, 10(8), 104-110.

이혜림 (2023). 데이터 리포지토리의 보존 정책 프레임워크에 관한 연구: CoreTrustSeal 인증을 중심으로. 한국문헌정보학회지, 57(4), 119-139. https://doi.org/10.4275/KSLIS.2023.57.4.119

이혜림, 엄정호, 신영호, 임형준, 한나은 (2024). CoreTrustSeal 인증 획득을 통한 데이터 리포지토리의 신뢰성 향상을 위한 연구. 정보관리학회지, 41(2), 245-268.

https://doi.org/10.3743/KOSIM.2024.41.2.245

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).

Andreoli-Versbach, P. & Mueller-Langer, F. (2014). Open access to data: an ideal professed but not practicised. Research Policy, 43(9), 1621-1633. https://doi.org/10.1016/j.respol.2014.04.008

Ball, A. (2009, June 3). Scientific data application profile scoping study report. Available: http://www.ukoln.ac.uk/projects/sdapss/papers/ball2009sda-v11.pdf

Borgerud, C. & Borglund, E. (2020). Open research data, an archival challenge? Archival Science. https://doi.org/10.1007/s10502-020-09330-3

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

Bulter, D. (2016). Zika researchers release real-time data on viral infection study in monkeys. Nature, 530(5). https://doi.org/10.1038/nature.2016.19438

Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional repositories: exploration of costs and value. D-Lib Magazine, 19(1/2), 1-14. https://doi.org/10.1045/january2013-burns

Cambridge Crystallographic Data Centre (n.d.). The Cambridge Structural Database. Available: https://www.ccdc.cam.ac.uk/structures/?

Clarivate Analytics (2024). Data Citation Index. Available: https://clarivate.com/academia-government/ko/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/data-citation-index/

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

Corti, L., Eynden, V., Bishop, L., & Woollard, M. (2014). Managing and sharing research data: A guide to good practice. Los Angeles, CA: SAGE.

Costa, M. R., Qin, J., & Bratt, S. (2016). Emergence of collaboration networks around large scale data repositories: a study of the genomics community using GenBank. Scientometrics, 108, 21-40. https://doi.org/0.1007/s11192-016-1954-x

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

Diamantopoulos, N., Sgouropoulos, C., Kastrantas, K., & Manouselis, N. (2011). Developing a metadata application profile for sharing agricultural scientific and scholarly research resources. In E. García-Barriocanal, Z. Cebeci, M. C. Okur, & A. Öztürk eds. Research Conference on Metadata and Semantic Research. Berlin, Heidelberg: Springer, 240, 453-466.

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

Downs, R. & Chen, R. (2006). Organizational needs for managing and preserving geospatial data and related electronic records. Data Science Journal, 4, 255-271.

Elsevier (2020, January 27). Novel coronavirus information center. Available: https://www.elsevier.com/connect/coronavirus-information-center

EMBL-EBI (2025). Electron Microscopy Data Bank. Available: https://www.ebi.ac.uk/emdb/

Funk, C., Hefferon, B., & Johnson, C. (2019, August 2). Trust and mistrust in Amerians' 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

Gries, C., Budden, A., Laney, C., O'Brien, M., Servilla, M., Sheldon, W., Vanderbilt K., & Vieglais, D. (2018). Facilitating and improving environmental research data repository interoperability. Data Science Journal, 17(22), 1-8. https://doi.org/10.5334/dsj-2018-022

Helbig, K., Hausstein, B., & Toepfer, R. (2015). Supporting data citation: experiences and best practices of a DOI allocation agency for social sciences. Journal of Librarianship and Scholarly Communication, 3(2), eP1220. https://doi.org/10.7710/2162-3309.1220

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

Mongeon, P., Robinson-Garcia, N., Jeng, W., & Costas, R. (2017). Incorporating data sharing to the reward system of science: linking DataCite records to authors in the Web of Science. Aslib Journal of Information Management, 69(5), 545-556. https://doi.org/10.1108/AJIM-01-2017-0024

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

Oblasser, S., Mika, T., & Kitamoto, A. (2020). Finding a repository with the help of machine-actionable DMPs: opportunities and challenges. International Journal of Digital Curation, 15(1), 1-12. https://doi.org/10.2218/ijdc.v15i1.704

OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Available: https://www.oecd-ilibrary.org/science-and-technology/oecd-principles-and-guidelines-for-access-to-research-data-from-public-funding_9789264034020-en-fr

Open Science Framework (2019). FAQs. Available: https://help.osf.io/hc/en-us/articles/360019737894-FAQs#what-is-the-individual-file-size-limit

Park, H. & Wolfram, D. (2018). Research software citation in the Data Citation Index: current practices and implications for research software sharing and reuse. Journal of Informetircs, 13(2), 574-582. https://doi.org/10.1016/j.joi.2019.03.005

Perrier, L., Blondal, E., & MacDonald, H. (2020). The views, perspectives, and experiences of academic researchers with data sharing and reuse: a meta-synthesis. PLoS ONE, 15(2), e0229182. https://doi.org/10.1371/journal.pone.0229182

Piwowar, H. A. (2011). Who shares? Who doesn't? Factors associated with openly archiving raw research data. PLoS ONE, 6(7), e18657. https://doi:10.1371/journal.pone.0018657.g003

Smith, R. D. & Kelleher, N. L. (2018). Mass spectrometry and its evolving role in life sciences research. Nature Methods, 15(7), 491-492. https://doi.org/10.1071/CH13284

Springer Nature (2019). Research data policies - Recommended repositories. Available: https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories

Springer Nature (n.d.). Available: https://www.nature.com/sdata/policies/repositories

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

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, U. A., 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

Vilnius University (n.d.). Crystallography Open Database. Available: https://www.crystallography.net/

Wang, J., Li, P., Zhuang, X., Li, X., Jiang, X., & Wu, J. (2024). Multi-source data integration and multi-scale modeling framework for progressive prediction of complex geological interfaces in tunneling. Underground Space, 15, 1-25. https://doi.org/10.1016/j.undsp.2023.08.006

Yakel, E., Faniel, I. M., Kriesberg, A., & Yoon, A. (2013). Trust in digital repositories. International

Journal on Digital Curation, 8, 143-156. https://doi.org/10.2218/ijdc.v8i1.251

Yoon, A. (2014). End users' trust in data repositories: Definition and influences on trust development. Archival Science, 14, 17-34. https://doi.org/10.1007/s10502-013-9207-8

• 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

Kang, Joonhee & Han, Byungchan (2019). Development of first-principles database driven machine learning potential for multi-scale simulations. KIC NEWS, 22(4), 13-19.

Kim, Juseop, Kang, Hyosuk, & Kim, Suntae (2023). Analysis of ecological data repository operation status and EcoBank service proposal. Journal of the Korean Society for Library and Information Science, 57(4), 289-310. https://doi.org/10.4275/KSLIS.2023.57.4.289

Kim, Yongmin (2016). Trends in genomic big data research. Molecular Cell Biology Newsletter, 1-8.

Park, Hyoungjoo (2024). Data sharing, reuse, and citation in biological research: an analysis of full-text literature. Journal of the Korean Society for Library and Information Science, 58(4), 335-353. https://doi.org/10.4275/KSLIS.2024.58.4.335

Rhee, Hea Lim (2023). A study on the preservation policy framework of data repository: focusing on CoreTrustSeal certification. Journal of the Korean Society for Library and Information Science, 57(4), 119-138. https://doi.org/10.4275/KSLIS.2023.57.4.119

Rhee, Hea Lim, Um, Jeong-Ho, Shin, Youngho, Yim, Hyung-jun, & Han, Na-eun (2024). A study to improve the trustworthiness of data repositories by obtaining CoreTrustSeal certification. Journal of the Korean Society for Information Management, 41(2), 245-268. https://doi.org/10.3743/KOSIM.2024.41.2.245

Song, Young-ok, Kim, Sung-Young, & Chang, Duk-Jin (2010). Design of the system and algorithm for the pattern analysis of the bio-data. The Journal of the Korea Contents Association, 10(8), 104-110.