The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 1, pp.443-451, January 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.1.443
JCCT 2024-1-52

데이터 스칼라십: 데이터 저널과 데이터 리포지토리를 중심으로

Data Scholarship: Data Journals and Data Repositories

박형주*

Hyoungjoo Park*

요 약 본 연구는 데이터 스칼라십을 이해하기 위하여 데이터 논문으로 색인되는 저널의 지적 구조를 분석 및 시각화하고 데이터 리포지토리의 운영을 비교하였다. 동료 평가(peer review) 유형을 살펴보고, 공동 출현 분석(co-occurence analysis) 및 네트워크 분석(network analysis)을 실시하였다. WoS에 데이터 논문으로 색인되는 상위 10위 저널은 전통적인 유형과 데이터 논문 유형을 혼재해서 발간하고 있었다. DCI에 색인되는 데이터 리포지토리는 대부분 북미 및 유럽 국가에서 운영하고 있다. 국내의 데이터 리포지토리는 대부분 연구원에서 운영하고 있다. 본 연구의 결과는 데이터 저널, 데이터 리포지토리 등 데이터 스칼라십의 관행을 이해하는 데 도움이 되기를 바란다.

주요어 : 데이터 저널, 공동 출현 분석, 네트워크 분석

Abstract The purpose of this study is to comprehend the knowledge structure of data scholarship within data journals and repositories. The study explored various aspects, including types of peer review, co-occurrence analysis through author keywords, and network analysis via article titles. The majority of data repositories in the DCI are maintained by countries in North America and the European Union. In Korea, data repositories are predominantly managed by research institutions. This study contributes to enhancing our understanding of the practices in data scholarship.

Keywords : Data journal, Co-occurrence analysis, Network analysis

Ⅰ. 서 론

데이터 스칼라십은 연구 과정에서 발생하는 데이터의 복잡한 관계와 스칼라십 간의 관계를 설명하며, 데이터의 수집, 관리, 처리, 분석, 시각화에 중점을 둔다[1]. 데이터 스칼라십은 연구 투명성에 대한 요구의 증가와 함께 그 중요성이 증대되고 있다. 오픈 사이언스 패러다임에서 데이터 저널과 데이터 리포지토리는 연구데이터가 공유되고 재이용되도록 지원하는 새로운 매개물이며, 새로운 과학적 발견에 기여할 수 있다는 점에서 데이터 스칼라십은 주목을 받고 있다. 데이터 저널은 연구데이터가 개별적으로 잘 문서화되고 출판이 가능하며 인용이 가능하도록 하는데 상당히 중요한 역할을 수행한다[2]. 데이터 저널의 논문 심사와 평가 과정을 통하여 연구데이터는 동료 평가(peer review)를 받을 수 있다. 데이터 저널은 데이터 출판에 따른 학술

*정회원, 충남대학교 문헌정보학과 조교수 (제1저자, 교신저자)
접수일: 2023년 10월 16일, 수정완료일: 2023년 11월 1일
게재확정일: 2023년 11월 10일

Received: October 16, 2023 / Revised: November 1, 2023
Accepted: November 10, 2023
*Corresponding Author: hyoungjoo.park@cnu.ac.kr
Dept. of Library and Information Science, Chungnam National Univeristy, Korea

크레딧을 부여할 수 있는 매개체이자, 데이터의 품질을 향상시키는 데 도움을 줄 수 있으며, 예로는 Scientific Data, Data in Brief, GigaScience 등이 있다. 전통적인 학술 논문과 데이터 논문의 공통점은 두 유형의 논문이 동료 심사 과정을 대부분 거쳐야 한다는 점이고, 차이점은 데이터 논문은 연구데이터 객체를 기술하기 위하여 디자인되었다는 점이다[3]. 데이터 논문은 궁극적으로 연구데이터의 공유와 재이용을 촉진하는 역할을 수행할 것이다[4]. 논문이 연구데이터와 함께 제공되는 경우 논문의 인용이 25% 증가한다[5]. 연구데이터의 동료 심사가 일반적이지 않은 현재의 학술 관행은, 공유되는 연구데이터의 품질 저하로 이어질 수 있다. 해결책 중의 하나는 학술 논문 내에 연구데이터를 함께 출판하는 것이다[6, 7]. 연구데이터가 저널 내부 혹은 외부의 데이터 리포지토리에 저장된 경우, 안정된 상태로 장기간 보존될 수 있다[8]. 요약하면, 데이터 저널은 연구데이터의 동료심사를 통하여 공유된 연구데이터의 품질을 향상시킬 수 있고, 지속적으로 장기간 저장을 가능하게 하고, 연구데이터를 추적하고 색인되도록 도와서 연구데이터가 학문적 가치를 발휘할 수 있도록 한다.

본 연구의 목적은 국내외 데이터 저널의 지적 구조를 분석 및 시각화하고 데이터 리포지토리의 비교를 통하여 데이터 스칼라십을 이해하는 것이다. 구체적인 연구 문제는 다음과 같다.

*연구 문제 1: 데이터 저널의 출판 현황은 어떠한가?

*연구 문제 2: 데이터 저널의 지적 구조는 어떠한가?

*연구 문제 3: 데이터 리포지토리의 운영 현황은 어떠한가?

II. 선 행 연 구

해외 주요 연구재단의 연구데이터 관리 정책, 영향력 있는 출판사의 연구데이터 공유 정책, 학술 데이터베이스의 연구데이터 색인은 연구데이터의 공유, 재이용, 인용에 효과적이다. 연구데이터는 해외 주요 연구 재단인 미국국립과학재단, 미국국립보건원과 영향력 있는 출판사인 Springer, Taylor & Francis Group, Wiley 등의 연구데이터 공유 정책과 더불어 그 중요성이 증가하고 있다. 예를 들어, Springer Nature는 자사 저널의 연구데이터 공유 정책과 함께 데이터 리포지토리 가이드라인을 저자 및 에디터에게 제공하여 연구데이터의 공유를 위한 데이터 리포지토리의 선택을 돕고 있으며, 논문과 연구데이터의 제출에서 출판까지의 전 과정에서 연구데이터 공유 정책을 제공하고 있다[9]. Scientific Data는 데이터 리포지토리 가이드라인에 학문 분야 별 데이터 리포지토리(discipline-specific data repositories) 목록을 제공하는데, 구체적으로 사회과학 분야의 Inter-university Consortium for Political and Social Research(ICPSR)와 물리학 분야의 High-Energy Physics Data(HEPData) 등이 있다[10]. 학술데이터베이스도 전 세계에 공유된 연구데이터의 추적과 색인을 하고 있다. 예를 들어, Clarivate Analytics사의 Data Citation Index(DCI)는 전 세계 450여개의 데이터 리포지토리에서 1,400만개 이상의 데이터세트, 160만개 이상의 데이터연구, 40만개 이상의 소프트웨어를 추적 및 색인하고 있다[11].

데이터 저널은 연구데이터의 수집 방법에 대한 상세 기술(descriptive) 정보를 제공하는 학술 저널이며, 연구데이터의 공유 및 인용 등을 통하여 연구의 투명성을 향상시킬 수 있다. 데이터 저널은 학술지의 새로운 유형의 하나로 데이터 논문을 출판하는 학술지를 일컬으며, Scientific Data, Data in Brief, Biodiversity Data Journal 등이 있다. 데이터 저널은 데이터의 수집, 방법론, 변수 등에 대한 정보를 기존의 전통적인 학술지보다 상세하게 기술하므로 데이터 저널은 연구 및 출판 과정에 대한 투명성을 증대시킬 수 있다. 생명과학 분야와 로봇 분야는 데이터 논문이 이미 정립된 학문이며[4, 12], 심리학, 역사학 등은 데이터 논문의 장점을 발견하고 활용하는 사례가 증가하고 있다[13, 14]. 데이터 저널은 연구데이터의 공유 및 인용 등을 통하여 연구의 투명성을 향상시킬 수 있다. 데이터 저널이 급격히 증가하고 있으며, WoS는 약 70%의 데이터 저널을 색인하고 있다[15]. 데이터 저널이 갖추어야 할 최소 요소는 연구의 전체 개요, 연구 참여자 정보, 방법론, 표현형 평가 프로토콜, 스캔 상세 정보, 연구데이터의 배포에 대한 정보가 있다. 데이터 저널은 논문의 동료심사 과정 중에 연구데이터의 품질을 심사하기 때문에 양질의 연구데이터 공유와 재이용에 순기능을 한다. 효과적인 연구데이터의 공유를 위해서는 연구데이터 재이용의 근거를 제시하는 충분한 메타데이터 및 연구데이터의 유효성이 필요하다[16]. 연구데이터의 수집, 분석, 공유에 대한 구체적인 기술(description)은 다른 연구자가 데이터 저널을 발견하고 평가할 수 있도록 돕는다[17,

18]. 대부분의 연구자들은 공식적인(formal) 학술 크레딧을 받을 수 있다면 연구데이터를 공유하고자 하는 의지를 보인다[19]. 논문과 연구데이터가 함께 공유될 경우 논문의 인용이 25% 증가한다[6].

데이터 저널, 논문의 부록(appendix), 학술 데이터베이스는 연구데이터의 공유, 재이용, 인용을 촉진하도록 돕는 매개체이다. 첫째, 데이터 저널은 학술 논문의 새로운 형태로, 연구데이터에 대한 자세한 기술을 통하여 연구자가 연구데이터를 더욱 잘 이해하도록 한다. 데이터 논문은, 연구데이터에 대한 상세한 기술을 포함하기에 논문의 독자는 연구데이터의 방법론에 대하여 이해할 수 있다. 데이터 논문은 참고문헌과 재이용에 편리한 연구데이터 공유 선언(declaration)과, 디지털객체식별자(digital object identifier; DOI)를 제공하므로 연구데이터의 공유가 쉬우며 독자가 공유된 연구데이터에 접근하기에 용이하다[20]. Cao는 전통적인 논문은 연구의 가정(hypothesis), 데이터 분석, 결과, 논의, 결론 등으로 구성되는 반면, 데이터 논문은 연구데이터가 데이터 논문의 주요 내용이라는 점에서 차이가 있다고 하였다. 둘째, 논문의 부록(appendix)에 연구데이터를 공유하는 것은, 논문의 필수 사항은 아니지만 논문의 신뢰도를 향상시킬 수 있다. 논문의 부록에는 일반적으로 디지털객체식별자가 부여되지 않으므로, 해당 논문을 통해서만 인용이 될 수 있다는 단점이 있다. 셋째, 데이터베이스를 통한 연구데이터의 공유는 학술 데이터베이스, 전자 아카이브, 데이터 센터 등이 있다. 데이터베이스는 규칙에 따라서 데이터 자원을 저장하므로, 데이터의 검색이 용이하고, 데이터의 접근과 공유를 안정적으로 제공한다. 연구데이터가 생성된 방법론을 데이터베이스에 기록하지 않으므로 데이터의 신뢰성이 부족할 수 있다.

데이터 저널은 연구데이터의 품질을 점검하고 향상하는 데 도움을 줄 수 있다. 연구데이터의 공유는 신경과학, 유전체학, 지구과학, 천문학 등의 학문에서는 일반적인 관행이다[19, 21]. 연구데이터의 출판은 연구데이터의 품질을 점검하고, 동료 심사와 편집 위원회의 의사 결정이 포함되므로, 연구데이터의 품질을 향상시킬 수 있다[22]. 연구데이터의 동료심사는 아직 학계의 관행이 아니다. 연구데이터의 품질을 향상시키기 위하여 학술 논문 내에 연구데이터를 동시에 출판하는 방법이 제안되기도 하였다[6, 7]. 전통적인 논문과 데이터 논문의 동료 심사의 차이점은, 전통적인 논문의 동료 심사 절차는 대부분 표준화되어 있는 반면, 데이터 논문의 연구데이터 동료 심사는 학문 분야에 따라 관행이 다르고 일관된 표준이 부족한 점이다[2, 23]. 데이터 저널은 연구데이터 소유권 등의 지적재산권 문제를 명확히 하지만, 데이터 논문의 동료심사는 심사위원의 수가 한정적이므로 학술커뮤니케이션의 또 다른 부담이 될 수 있다[24].

선행 연구를 통하여 데이터 저널은 공유된 연구데이터의 발견을 촉진하고, 데이터의 품질을 향상시키고, 데이터 인용을 색인하기에 효과적임을 확인하였다. 기존의 연구는 지적 구조를 이해하기 위하여 키워드를 통한 자료의 내용 분석을 하거나[25] 학술지 논문을 중심으로 분석되어 왔다[26].

III. 방 법 론

국내외 데이터 저널을 비교 분석하였다. 첫째, 국외 데이터 저널은 WoS(Web of Science)에 색인된 데이터 논문을 기준으로 수집 및 분석하였다. WoS의 핵심 컬렉션(Core Collection)에서 고급 검색을 수행하였으며, 문서 유형(document type)은 데이터 논문(data paper)으로 한정하였다. 2022년 8월 22일에 데이터를 수집하였으며, 171개 데이터 저널의 11,370개 데이터 논문에 대한 서지 정보와 인용 정보를 수집하였다. 수집된 데이터 중에서 저자 명에 논문 명이 있는 등 데이터가 잘못 정렬된 경우는 제외하였다. 이는 WoS의 한계로 보인다. 수집된 데이터에는 사설(editorial)이나 정오표(errata)로 색인된 논문은 없었다. 최종 분석의 대상은 WoS에 데이터 논문으로 색인된 11,362개의 데이터 논문의 서지 정보와 인용 정보였다. 상위 10위의 논문을 집중 분석하였다. 둘째, 국내에서 발간되는 모든 데이터 저널을 발견하고 식별하기는 쉽지 않았지만, 저널 유형을 '데이터 저널'로 분류해 놓아 식별이 가능한 경우만을 분석의 대상으로 삼았다. WoS에서 데이터 논문으로 색인되는 국내의 저널은 발견되지 않았다. 국내에서 발간되는 데이터 저널을 확인하기 위하여 구글 등의 검색엔진에서 '데이터 저널'과 'data journal'을 검색어로 활용하였다. 이 후 사이트에 개별 방문하여 데이터 저널 여부를 재확인하였다.

데이터 리포지토리의 식별과 분석은 Clarivate

Analytics의 DCI에 색인된 모든 데이터 리포지토리 목록을 확보하여[27] 분석하였다. DCI는 전 세계 450여개 데이터 리포지토리에 저장된 1,400만개 이상의 데이터 세트, 160만개 이상의 데이터 연구, 40만개 이상의 소프트웨어를 추적 및 색인한다.

IV. 실험 및 결과

<표 1>은 WoS에 색인된 데이터 논문의 수가 많은 상위 10위 저널을 보여준다. 창간 년도는 Clarivate Analytics의 Master Journal List[28]를 참고하였다. 저널 영향도는 2021년 Journal Citation Report(JCR)를 활용하였다. 각 저널이 채택한 동료 심사의 유형 정보를 수집하기 위하여 각 저널의 웹사이트를 개별 방문하였다. 대규모의 연구데이터가 활용되어야 하는 과학 분야와 다학제 분야의 데이터 논문 발간이 활발했다.

표 1. WoS에 색인된 데이터 논문을 수록한 상위 10위 저널
Table 1. Top 10 journals indexed as data papers in WoS

순위	저널 명	학문 분류	출판사	저널 영향도	저널 랭크
1	Data in Brief	다학제	Elsevier	0.28	Q3
2	Scientific Data	다학제	Springer Portfolio	8.5	Q1
3	Earth System Science Data	지구과학, 다학제	Copernicus Publications	11.82	Q1
4	Data	컴퓨터 과학, 다학제	MDPI	0.66	Q2
5	Biodiversity Data Journal	생물 다양성 보전	Pensoft	1.54	Q3
6	GigaScience	다학제	Oxford University Press	7.66	Q1
7	Human Genome Variation	유전학	Springer Nature	0.39	Q4
8	Ecology	생태학	Wiley	6.43	Q1
9	Microbiology Resource Announcements	미생물학	American Sociology for Microbiology	0.19	Q4
10	BMC Research Notes	다학제	Springer Nature	0.5	Q2

표 2. WoS에 색인된 상위 10위 저널의 데이터 논문 비교: 인용
Table 2. Top 10 journals indexed as data papers in WoS: citations

순위	저널 명	전체 논문	데이터 논문	
		출판 건수	출판 건수(%)	총 인용수
1	Data in Brief	7,127	6,705(94.08%)	40,128
2	Scientific Data	2,395	1,866(77.91%)	18,146
3	Earth System Science Data	1,090	560(51.38%)	12,729
4	Data	605	286(47.27%)	4,156
5	Biodiversity Data Journal	1,095	286(26.12%)	1,439
6	GigaScience	1,032	201(19.48%)	1,378
7	Human Genome Variation	270	182(67.41%)	1,020
8	Ecology	10,377	136(1.31%)	1,007
9	Microbiology Resource Announcements	3,830	80(2.09%)	757
10	BMC Research Notes	2,092	66(3.15%)	719

<표 2>는 저널의 전체 논문 건수와 데이터 논문 건수를 비교한 표이다. 순위는 데이터 논문의 수가 많은 저널을 내림 차순한 것이다. 전체 논문 중 데이터 논문의 수가 50%를 상회하는 논문은 Data in Brief, Scientific Data, Human Genome Variation, Earth System Science Data였다. 전체 논문 수 당 데이터 논문의 비율은 Ecology 저널이 1.31%로 가장 낮았고, Data in Brief가 94.08%로 가장 높았다. Ecology가 1.31%로 낮은 이유는 창간 년도가 1920년도로 100년 이상의 역사를 가진 저널이므로 전체 출판 논문에 비하여 데이터 논문의 수가 적기 때문으로 해석될 수 있다. 요약하면, 순수하게 데이터 논문만을 수록하는 저널은 찾기 힘들며, 데이터 논문과 전통적인 논문을 혼재해서 발간하고 있다. 학술 논문 출판의 새로운 유형으로 데이터 논문을 게재하는 저널이 증가 추세에 있으며, 데이터 논문은 전통적인 학술 논문과 동등한 지위를 유지하고 있다고 해석될 수 있다.

WoS에 데이터 논문을 수록한 상위 10위 저널의 동료 심사(peer review) 유형을 분석하였다. 상위 10위 저널은 역순으로 Data in Brief, Scientific Data, Earth System Science Data, Data, Biodiversity Data Journal,

GigaScience, Human Genome Variation, Ecology, Microbiology Resource Announcements, BMC Research Notes였다. 동료 심사의 유형은 익명 심사(blind review)를 채택한 저널, 익명 심사와 공개 심사를 절충한 저널, 공개 심사(open peer review)를 채택한 저널이 혼재되어 있었다. Biodiversity Data Journal은 일반인이 공개적으로 논문의 심사평을 올릴 수 있다. BioDiversity Data Journal은 에디터(editor)의 이름은 공개하지만 논문 게재 여부에 대한 에디터의 결정문(decision letter)은 공개하지 않는다. Microbiology Resource Announcements는 2013년 Genome Announcement로 창간된 후 2018년 현재의 명칭으로 변경되었다[29]. Ecology가 오픈 액세스와 closed access를 혼재해서 출판할 수 있는 유일한 저널이었고, 그 외의 저널은 모두 오픈 액세스를 채택하고 있었다. 즉, 데이터 논문을 수록한 저널은 상대적으로 공개된 오픈액세스 형태를 채택하고 있다. 저널의 WoS의 주제 범주(subject category)는 대부분 다학제적 분야 혹은 STEM(Science, Technology, Engineering, Mathematics) 분야였으며, 구체적으로 기상학, 지리 과학, 컴퓨터 과학, 정보 시스템, 생물다양성 보전, 유전학, 생태학, 미생물학이었다. 상위 10위 저널의 학문 분야에 인문학 및 사회과학 분야는 없었다. 요약하면, 데이터 논문을 발간하는 주체에 따라서 데이터 논문을 운영하는 방식이 상이했으며, 데이터 논문을 발간하는 주체에 따라서 출판 모델이 다양했다. 데이터 저널이 가장 많이 출판된 학문 분야(subject category)의 상위 5위는 내림차순으로 과학 기술(8,568회, 75.48%), 지질학 및 기상학(623회, 5.49%), 생물 다양성 및 보존(297회, 2.62%), 전산학(288회, 2.54%), 유전학(270회, 2.38%)이었다.

국내에서 출판되는 모든 데이터 저널을 발견하고 식별하기는 쉽지 않으나, 저널 유형을 데이터 저널로 구분한 저널은 2023년 5월을 기준으로 국내에서는 Geo Data가 유일한 것으로 보인다. Geo Data는 2019년 국립생태원, 한국지질자원연구원, 한국해양과학기술원 및 부설 극지연구소, 한국항공우주연구원의 5개의 공공연구기관이 공동으로 발간한 데이터 저널이다. 연 4회 발간되는 저널이며, 논문의 종류는 '데이터 설명 논문', '데이터 리뷰 논문', '데이터 단보'의 세가지 유형 중 한 종류로 투고할 수 있다. GeoData가 요구하는 논문의 분량은 2,000 단어 내외로 비교적 짧다. 연구 분야는 지구과학 분야이며, 구체적으로 극지학, 생태학, 우주 항공, 지질학, 환경학, 해양학이다. 데이터의 필수 기술 항목은 데이터의 제목, 디지털객체식별자(digital object identifier; DOI), 카테고리, 초록, 수집 기간, 등록자 정보, 크리에이티브 커먼즈 라이선스(Creative Commons license)이다. 데이터의 선택 기술 항목은 수행 과제 및 수집 장비이다. 연구데이터의 저장소는 외부의 데이터 리포지토리가 아닌, Geo Data 데이터 저널의 자체 웹사이트에 저장하도록 하고 있다. 요약하면, 국내의 데이터 저널 발간은 초기 단계에 있다고 해석될 수 있다.

<그림 1>은 데이터 저널의 저자 키워드에 기반한 공동 출현 단어 분석(co-occurrence analysis)이며, <표 3>은 클러스터별 주요 키워드를 보여준다. 26,491개의 키워드 중에서 키워드 출현이 최소 10회인 253개의 키워드를 대상으로 분석하였다. 총 9개의 클러스터로 구성되어 있었다. 클러스터1은 시민 과학(citizen science)에 기반한 조류 연구, 클러스터 2는 암, RNA(Ribonucleic acid) 등 생체 정보, 클러스터3은 유전자, 클러스터4는 가뭄 등의 기후 변화로 인한 아프리카 지역 곡물 등의 식량 불안정에 대한 내용이었다. 클러스터5는 남극 해양 지역의 탄소 및 질소와 관련된 환경 이슈였다. 클러스터6은 폐수, 중금속 등의 지하수 식수 오염과 관련된 내용이었다. 클러스터7은 빅데이터 분석, 사물 인터넷, 에너지 소비 및 효율 등의 4차 산업과 관련된 내용이었다. 클러스터8은 인공지능, 머신 러닝, 데이터 마이닝, 자연어 처리 등의 빅데이터 분석에 활용되는 방법론에 대한 내용이었다. 클러스터9는 코로나19 전염병에 따른 불안, 우울 등의 정신 건강 이슈로 인한 삶의 질에 대한 내용이었다.

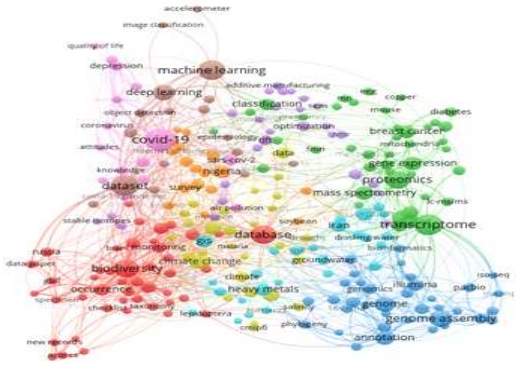

그림 1. 저자 키워드에 기반한 공동 출현 분석
Figure 1. Co-occurence analysis through author keywords

표 3. 저자 키워드에 기반한 공동 출현 분석 클러스터의 주요 키워드
Table 3. Clusters: co-occurence analysis through author keywords

클러스터	주요 키워드
1	biogeography, birds, citizen science, monitoring
2	breast cancer, cancer, DNA, methylation, gene expression, metabolism, RNA sequencing
3	antibiotic resistance, bioinformatics, genome, evolution
4	Africa, agriculture, climate change, drought, food security, rice
5	Antarctica, biocontrol, carbon, metabolomics, nitrogen, oxidation
6	drinking water, groundwater, heavy metals, wastewater, water, water quality
7	Big data, data analysis, energy consumption, energy efficiency, internet of things
8	artificial intelligence, data mining, deep learning, machine learning, natural language processing
9	anxiety, covid-19, depression, mental health, pandemic, public health, quality of life

<그림 2>는 논문의 제목(title)을 기준으로 네트워크 분석을 한 결과이며 <표 4>는 클러스터 별 주요 키워드를 보여준다. Binary counting으로 분석한 33,401개의 용어 중에서 최소 10번 이상 출연한 용어는 386개였다. 클러스터 1은 서식지, 숲, 강, 식물 등에서 일어나는 환경 변화에 관한 내용이었다. 클러스터 2는 세포에 있는 RNA 전사체(transcriptome), 클러스터 3은 코로나 전염병 발발, 클러스터 4는 화학 물질 실험, 클러스터 5는 게놈(genome), 클러스터 6은 돌연변이 게놈, 클러스터 7은 지하수 수질 및 강수량, 클러스터 8은 메타유전체학, 클러스터 9는 미국 캘리포니아에 대한 내용이었다.

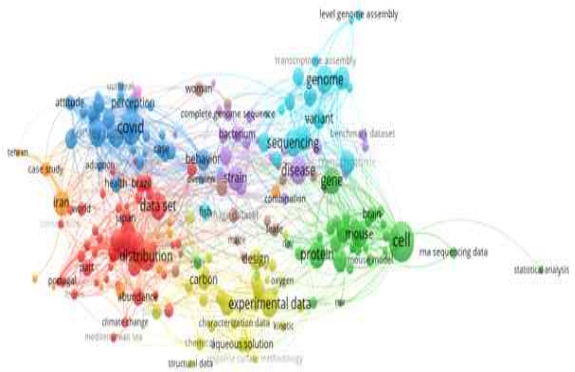

그림 2. 논문 제목에 기반한 네트워크 분석
Figure 2. Network analysis through article titles

표 4. 제목에 기반한 네트워크 분석 클러스터의 주요 키워드
Table 4. Clusters: network analysis through article titles

클러스터	주요 키워드
1	climate change, forest, habitat, river, vegetation
2	Alzheimer, brain, cell, gene protein, proteomic, RNA sequencing data, transcriptomic data
3	Covid, outbreak, pandemic, perception
4	chemical, correlation, design, experimental data, mechanical property, metal, nitrogen, oxygen, structural data
5	complete genome sequence, draft genome sequence
6	genome, mutation, sequence, variant
7	global dataset, groundwater quality, precipitation, sediment, water quality
8	metagenomics data
9	California, USA

<표 5>와 <표 6>은 DCI에서 색인하는 데이터 리포지토리 목록[27]을 활용하여 분석한 결과이다. DCI에서 색인하는 전 세계 450여개의 데이터 리포지토리가 분석되었다. 구체적으로, 생명과학(38%), 물리과학(22%), 사회과학(19%), 다학제(17%), 기술(2%), 예술•인문학(2%)이었다. 요약하면, DCI의 데이터 리포지토리는 생명과학, 물리과학, 사회과학, 다학제 분야가 주로 색인되며, 기술 및 예술•인문학 분야는 색인되는 데이터 리포지토리의 수가 많지 않다.

표 5. 논문의 제목에 기반한 네트워크 분석 별 클러스터
Table 5. Clusters: network analysis through article titles

학문 분야	총수	퍼센트	데이터 리포지토리 예시
생명과학	170	38%	GenBank, ModelDB
물리과학	97	22%	PANGAEA, HEPData,
사회과학	84	19%	ICPSR
다학제	75	17%	Figshare, Zenodo, Mendeley Data
기술	11	2%	NanoHUB, Mantid Project, MatDB
예술•인문학	9	1.99%	Archaeology Data Service

<표 6>은 DCI에서 색인하는 데이터 리포지토리의 국가별 상위 10위를 보여준다. 중국과 일본을 제외하고는 북미와 유럽에 위치한 국가가 상위에 랭크되어 있었다. 북미 국가가 약 50%를 차지하고 있으며, 이 중 미국이 45.7%를 차지하고 있었다. 상위 10위에 랭크된 데이터 리포지토리의 유형은 다양하였는데, 일반 리포지토리(generic repository)인 Dryad, 기관 리포지토리인 하버드 대학교의 Harvard Dataverse, 학문 특화 리포지

토리(discipline-specific repository)인 PANGAEA 등이 있었다. 요약하면, DCI에서 색인하는 데이터 리포지토리는 주로 북미 국가 및 유럽에서 운영되고 있다. 국내의 데이터 레포지토리는 대부분 연구소에서 운영하고 있었다.

표 6. DCI에서 색인하는 데이터 리포지토리의 국가 별 비교
Table 6. Data repositories organized by countries in the DCI

순위	국가명	총 수(%)	데이터 리포지토리 예시
1	미국	204(45.7%)	Dryad, Harvard Dataverse, IEEE Data Portal
2	영국	60(13.5%)	1000 Genomes, ArrayExpress Archive, HEPData
3	독일	30(6.7%)	PANGAEA, World Data Centre for Climate
4	호주	17(3.8%)	Australian Antarctic Data Centre
5	캐나다	17(3.8%)	cIRcle, Scholars Portal Dataverse
6	네덜란드	15(3.4%)	Longitudinal Aging Study Amsterdam
7	중국	11(2.5%)	GigaDB, Science Data Bank
8	일본	11(2.5%)	Materials Data Repository
9	스위스	9(2%)	ETH Data Archive, Zenodo
10	프랑스	7(1.6%)	Gazel Cohort

Ⅴ. 논의 및 결 론

학술커뮤니케이션에서 전통적인 논문은 보편적인 규범이지만, 연구데이터의 인용은 아직 보편적인 규범이 아니다. 연구데이터는 인용이 가능해야 한다. 공유된 연구데이터의 검증은 데이터 리포지토리에 기탁 시에 이루어져야 하지만, 기탁되는 수많은 연구데이터의 품질 검증에는 현재의 학술커뮤니케이션의 관행에서는 많은 어려움이 따를 수 있다. 데이터 저널의 데이터 논문은 연구데이터가 데이터 논문을 출판하는 과정에서 동료 심사가 이루어지므로, 데이터 인용을 현실적이고 효과적으로 장려할 수 있다. 연구데이터를 공유할 때 데이터 인용이 연구자에게 중요한 인센티브를 제공함에도 불구하고, 데이터 공유자는 감사의글(acknowledgment) 등에 자신의 이름이 언급될 때 받을 수 있는 크레딧(credit)을 받지 못하고 있다[30]. 연구데이터는 논문처럼 출판, 인용, 색인되어야 한다[31]. 연구데이터 관리의 장애 요소는 연구데이터에 대한 공통적인 인프라의 부족, 연구 환경의 차이, 데이터의 규모 및 데이터 포맷의 부족이다.

데이터 저널은 연구데이터의 공유와 인용을 권고하며, 공유된 연구데이터는 학술 논문과 동등한 지위를 유지하는 방안이 될 수 있다. 데이터 저널은 연구데이터의 접근과 활용을 촉진하는 수단이 될 수 있다[15]. 국외는 다양한 주제 분야에서 다양한 출판사가 데이터 저널을 발간하고 있으나, 국내는 데이터 저널을 발간할 수 있는 출판 시스템 인프라가 미비한 실정이다. 본 연구는 2023년을 기준으로 국내의 데이터 저널은 Geo Data 하나만 있음을 확인하였다. 국외에는 Springer Nature 등의 대형 출판사의 데이터 공유 정책에 기인한 데이터 저널의 출판이 활발하다. 국내는 연구데이터와 관련된 전문 출판사가 부재하고 데이터 저널에 대한 인식의 부족으로 데이터 저널이 초기 단계에 있다. 데이터 저널은 연구데이터의 수집 방법에 대한 상세 기술은 포함하지만 데이터세트에 대한 분석 또는 결론은 포함하지 않을 수 있다. 데이터 저널에 논문을 출판할 경우, 데이터 저널이 논문의 원시 데이터(raw data) 공유를 요구한다는 점에서 윤리적인 문제와 법적 문제가 발생할 수 있다. 연구 참여자 중에서 미성년자가 포함되어 있거나, 의료 정보 등의 민감한 정보가 포함되어 있는 경우에는 더욱 주의를 기울여야 한다. 오픈사이언스의 패러다임에서는, 연구자는 연구의 시작 시점부터 연구데이터의 공유 가능성을 고려하면서 가능한 개방적이면서도 필요하다면 폐쇄성을 유지해야 한다. 오픈사이언스의 개방 패러다임과 실제 연구 프로젝트의 실정은 도전적인 과제일 수 있다.

본 연구는 데이터 스칼라십을 이해하기 위하여 국내외 데이터 저널의 지적 구조를 분석 및 시각화하고, 국내외 데이터 리포지토리의 운영 현황을 비교 분석하였다. 지적 구조 분석을 위하여 공동 출현 분석과 네트워크 분석을 실시하였다. WoS에 데이터 논문으로 색인되는 상위 10위 저널은 오픈 액세스를 채택하는 경우가 많았고, 학문 분야는 다학제 및 STEM 분야가 많았으며, 혼재된 동료평가 유형을 운영하고 있었다. DCI의 데이터 리포지토리는 생명과학, 물리과학, 사회과학, 다학제 리포지토리가 학술 데이터베이스에 의해서 주로 색인되고 있다. DCI에 색인된 리포지토리는 북미 및 유럽 연합에서 주로 운영하고 있으며, 국내의 데이터 리포지토리는 연구원에서 주로 운영하고 있다.

References

[1] C. L. Borgman, Big Data, Little Data, No Data, Cambridge: MIT Press, 2015.

[2] M. A. Parsons and P. A. Fox, “Is Data Publication the Right Metaphor?,” Data Science Journal, Vol. 12, pp. WDS32-WDS46, 2013. DOI: 10.2481/dsj.WDS-042

[3] D. Carlson and T. Oda, “Editorial: Data Publication-ESSD Goals, Practices and Recommendations,” Earth System Science Data, Vol. 10, pp. 2275-2278, 2018. DOI: 10.5194/essd-10-2275-2018

[4] V. Chavan and L. Penev, “The Data Paper: a Mechanism to Incentivize Data Publishing in Biodiversity Science,” BMC Bioinformatics, Vol. 12, No. 15, p. S2, 2011. DOI: 10.1186/1471-2105-12-S15-S2

[5] G. Colavizza, I. Hrynaszkiewicz, I. Staden, K. Whitaker and B. McGillivray, “The Citation Advantage of Linking Publications to Research Data,” PLoS ONE, Vol. 15, No. 4, p. e0230416, 2020. DOI: 10.1371/journal.pone.0230416

[6] X. Huang, B. A. Hawkins and G. Qiao, “Biodiversity Data Sharing: Will Peer-reviewed Data Papers Work?,” BioScience, Vol. 63, No. 1, pp. 5-6, 2013. DOI: 10.1525/bio.2013.63.1.2

[7] M. S. Mayernik, S. Callaghan, R. Leigh, J. Tedds and S. Worley, “Peer Review of Datasets: When, Why, and How,” Bulletin of the American Meteorological Society, Vol. 96, No. 2, pp. 191-201, 2015. DOI: 10.1175/BAMS-D-13-0083.1

[8] M. Assante, L. Candela, D. Castelli and A. Tani, “Are Scientific Data Repositories Coping with Research Data Publishing?,” Data Science Journal, Vol. 15, No. 6, 2016. DOI: 10.5334/dsj-2016-006

[9] Springer Nature, “Data Policies,”. https://www.nature.com/sdata/policies/data-policies.

[10] Springer Nature, “Data Repository Guidance,” https://www.nature.com/sdata/policies/repositories

[11] Clarivate Analytics, “Data Citation Index,” 2022. https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/.

[12] P. Newman and P. Corke, “Editorial: Data Papers-peer Reviewed Publication of High Quality Data Sets,” The International Journal of Robotics Research, Vol. 28, No. 5, p. 587, 2009.

[13] J. J. Kossakowski, P. C. Groot, J. M. Haslbeck, D. Borsboom and M. Wichers, “Data from ‘Critical Slowing Down as a Personalized Early Warning Signal for Description’,” Journal of Open Psychology Data, Vol. 5, No. 1, 2017. DOI: 10.5334/jopd.29

[14] K. Xu, B. Nosek and A. G. Greenwald, “Psychology Data from the Race Implicit Association Test on the Project Implicit Demo Website,” Journal of Open Psychology Data, Vol. 2, No. 1, pp. 1-3, 2014.

[15] L. Candela, D. Castelli, P. Manghi and A. Tani, “Data Journals: a Survey,” Journal of the Association for Information Science and Technology, Vol. 66, No. 9, pp. 1747-1762, 2015. DOI: 10.1002/asi.23358

[16] A. Zimmerman, “Not by Metadata Alone: the Use of Diverse Forms of Knowledge to Locate Data for Reuse,” International Journal on Digital Libraries, Vol. 7, pp. 5-16, 2007. DOI: 10.1007/s00799-007-0015-8

[17] I. X. Faniel, R. D. Frank and E. Yakel, “Context from the Data Reuser's Point of View,” Journal of Documentation, Vol. 75, No. 6, pp. 1274-1297, 2019. DOI:10.1108/JD-08-2018-0133

[18] X. Wang, Q. Duan and M. Liang, “Understanding the Process of Data Reuse: an Extensive View,” Journal of the Association for Information Science and Technology, Vol. 72, No. 9, pp. 1161-1182, 2021. DOI: 10.1002/asi.24483

[19] C. Tenopir, N. M. Rice, S. Allard, L. Baird, J. Borycz, L. Christian, B. Grant, R. Olendorf and R. J. Sandusky, “Data Sharing, Management, Use, and Reuse: Practices and Perceptions of Scientists Worldwide,” PLoS ONE, Vol. 15, No. 3, p. e0229003, 2020. DOI: 10.1371/journal.pone.0229003

[20] X. Cao, “Data Paper: an Important Type of Academic Articles,” Resources Data Journal, Vol. 1, pp. 10-16, 2022.

[21] H. H. Pierce, A. Dev, E. Statham and B. E. Bierer, “Credit Data Generators for Data Reuse,” Nature, Vol. 570, pp. 30-32, 2019. DOI: 10.1038/d41586-019-01715-4

[22] M. J. Costello, W. K. Michener, M. Gahegan, Z. Q. Zhang and P. E. Bourne, “Biodiversity Data Should be Published, Cited, and Peer Reviewed,” Trends in Ecology & Evoluation, Vol. 28, No. 8, pp. 454-461, 2013. DOI: 10.1016/j.tree.2013.05.002

[23] F. Murphy, “An Update on Peer Review and Research Data,” Learned Publishing, Vol. 29, No. 1, pp. 51-53, 2016. DOI: 10.1002/leap.1005
[24] M. E. Hochberg, J. M. Chase, N. J. Gotelli, A. Hastings and S. Naeem, “The Tragedy of the Reviewer Commons,” Ecology Letters, Vol. 12, pp. 2-4, 2009. DOI: 6.x
[25] S. Jung, “A Study on Changes in the Food Service Industry about Keyword before and after COVID-19 using Bid Data,” International Journal of Internet, Broadcasting and Communication, Vol. 14, No. 3, pp. 85-90, 2022. DOI: 10.7236/IJIBC.2022.14.3.85
[26] M. K. Lee and S. L. Kim, “Analysis on Research Trends of Early Childhood Software Education: Korean Articles Published in 2017 through 2022,” The Journal of the Convergence on Culture Technology, Vol. 9, No. 6, pp. 281-289, 2023. DOI: 10.17703/JCCT.2023.9.6.281
[27] Clarivate Analytics, “Master Data Repository list," 2022. https://clarivate.com/webofsciencegroup/master-data-repository-list/.
[28] Clarivate Analytics, “Master Journal List,” 2022. https://mjl.clarivate.com/journal-profile.
[29] International Standard Serial Number International Centre, “Key-title Gigascience,” https://portal.issn.org/resource/ISSN/2047-217X.
[30] H. Park and D. Wolfram, “An Examination of Research Data Sharing and Re-use: Implications for Data Citation Practice,” Scientometrics, Vol. 111, No. 1, pp. 443-461, 2017. DOI: 10.1007/s11192-017-2240-2
[31] M. J. Costello, “Motivating Online Publication of Data,” BioScience, Vol. 59, No. 5, pp. 418-427, 2009. DOI: 10.1525/bio.2009.59.5.9

※ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임.(NRF-2022M3J6A1084843)