저자 키워드 분석을 통한 연구데이터 관련 국외 연구 동향 분석*

An Analysis of Foreign Research Trend on Research Data Using Author Keywords Analysis

한 상 우 (Sangwoo Han)**

목 차

초 록

본 연구는 연구데이터 관련 국외 연구 동향을 분석하기 위하여 WoS에서 2000년부터 2023년까지 발행된 연구데이터 관련 논문을 수집하였으며, 수집된 데이터의 정제 후 총 693건의 연구논문을 대상으로 총 754개의 저자 키워드를 추출하여 키워드 네트워크 분석을 수행하였다. 분석 결과, 첫째, 국외의 연구동향은 국내의 연구동향에 비해 활성화되어 있음을 이해할 수 있었고, 둘째, 연구데이터 관련 연구는 Computer Science와 Information Science 중심으로 진행되었다. 셋째, 연구데이터 관련 연구에서는 Reseach Data Management, Data Repository, Open Data 등의 키워드가 연결정도 중심성이 가장 높은 것으로 나타났다. 본 연구의 결과를 통하여 국외의 연구데이터 관련 연구 동향을 파악할 수 있었고, 연구데이터 관련 국내 연구 동향과 비교를 통해 향후 연구의 방향을 제시할 수 있었다.

ABSTRACT

The goal of this study is to investigate foreign research trend on research data study. To achieve this goal, articles related research data topic were collected from WoS from 2000 to 2023. After data cleansing, author keywords were extracted from a total of 754 author keywords from 593 articles and keyword network analysis was performed. As a result of the analysis, first of all, it was understood that foreign research trends were activated compared to domestic research trends. Second, studies on research data were performed mainly in domain of Computer Science and Information Science. Third, studies related to research showed that keywords such as research data management, data repository, and Open data have the highest degree centrality. Through the results of this study, we were able to identify the foreign research trends on research data, and to compare with the domestic research trends.

키워드: 연구데이터, 과학데이터, 연구동향, 저자 키워드 분석
Research Data, Scientific Data, Research Trend, Author Keywords Analysis

* 이 연구는 2024년도 광주대학교 대학 연구비의 지원을 받아 수행되었음.
** 광주대학교 문헌정보학과 부교수(swhan@gwangju.ac.kr / ISNI 0000 0004 6851 1739)
논문접수일자: 2024년 11월 13일 최초심사일자: 2024년 11월 18일 게재확정일자: 2024년 11월 21일
한국문헌정보학회지, 58(4): 355-375, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.4.355

1. 서 론

연구데이터는 연구를 진행하기 위한 재료에서 연구 전반을 위한 체계 및 플랫폼으로 진화하고 있는 상황으로 보인다. 데이터를 양질의 정보를 생산하기 위해 근거로 사용하기 위해 어렵게 수집하고 가공하여 활용하였으나 보존이나 재사용에 관심을 두지 않았고, 유용한 데이터를 수집하기 어려우니 중복된 생산과 수집을 반복하는 것이 일반적인 상황도 있었다. 그러나 최근 연구데이터에 대한 관심과 중요성에 대한 인식이 전 세계적으로 나타나는 상황에서 연구데이터를 둘러싼 모든 상황이 변화하고 발전하고 있다. 이미 2010년 미국의 NSF는 연구제안서의 제출시 데이터 관리 계획을 필수적으로 첨부하도록 제도화하였고, 국내에서도 국가과제를 수행하는 경우 연구데이터를 필수적으로 관리하는 체계를 만들어가고 있다. 연구데이터 활용에 적극적이며 연구데이터 인프라 및 체계가 발달한 미국, 영국, 독일 등에서도 연구데이터의 접근, 재이용, 활용을 위한 법체계를 정비함으로써 연구데이터를 국가 연구의 기반 플랫폼으로 이용하려고 노력하고 있다.

연구데이터가 연구플랫폼으로 형성되고 활용되기 위해서는 다양한 주제 분야의 연구가 진행될 필요가 있다. 연구데이터 관련 연구는 비교적 최근에 활발하게 진행되고 있으며, 어떠한 방향과 양상으로 진행되고 있는지 연구동향을 살펴보는 것은 의미가 있을 것으로 판단된다. 아울러, 연구데이터와 관련하여 연구 및 운영이 발전한 국외의 동향을 살펴보는 것은 국내 연구의 활성화에도 유용할 것으로 예상된다.

따라서 본 연구에서는 연구데이터 관련 국외 연구논문을 조사하고 연구에 부여된 저자 키워드를 중심으로 분석하여 연구데이터 관련 국외 연구 동향을 분석해보고자 한다. 이를 위해 구체적으로 다음과 같이 연구문제를 설정하였다.

- RQ1. 연구데이터 관련 국외 연구의 주요 키워드는 무엇인가?
- RQ2. 연구데이터 관련 국외 연구의 출판 현황은 어떠한가?
- RQ3. 연구데이터 관련 주요 키워드의 특성 및 군집은 어떤 양상을 나타내는가?

본 연구에서는 위의 연구문제를 해결하기 위하여 국외에서 출판된 연구데이터 관련 연구논문을 수집하고 각 논문에 부여된 저자 키워드를 추출하여 키워드 네트워크 분석 방법으로 관련 연구 동향을 분석하고 그 결과를 바탕으로 국내의 연구데이터 관련 연구의 방향과 시사점을 제시해보고자 한다.

2. 이론적 배경

2.1 연구데이터

연구데이터는 “연구개발과제 수행 과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실자료로서 연구결과의 검증에 필수적인 데이터”로서(국가연구개발정보처리기준, 2020), 여러 연구 및 학문 분야에 따라 다르게 정의되기도 하지만, 핵심적인 개념은 ‘검증(verification)’, ‘재현(feasibility)’, ‘중복

방지(deduplication)', '재사용(reuse)' 등으로 이해할 수 있다. 또한, 연구데이터는 체계적으로 수집되고 관리되어 활용이 가능하도록 구체적인 관리 계획을 수립하는 것을 기본으로 하기 때문에 '데이터 관리 계획(Data Management Plan)' 역시 핵심 개념으로 이해할 수 있다. 또한, 연구데이터의 양적 증가 및 활용이 높아짐에 따라 데이터 리포지터리도 중요하게 인식되고 있다. 최근 20년간 전세계 오픈액세스 논문 저장소는 3배 이상 증가했고, re3data.org에 등록된 리포지토리의 수는 3,000여개를 넘어서고 있다(신은정 외, 2024; 한나은 외, 2024).

2.2 연구동향 분석

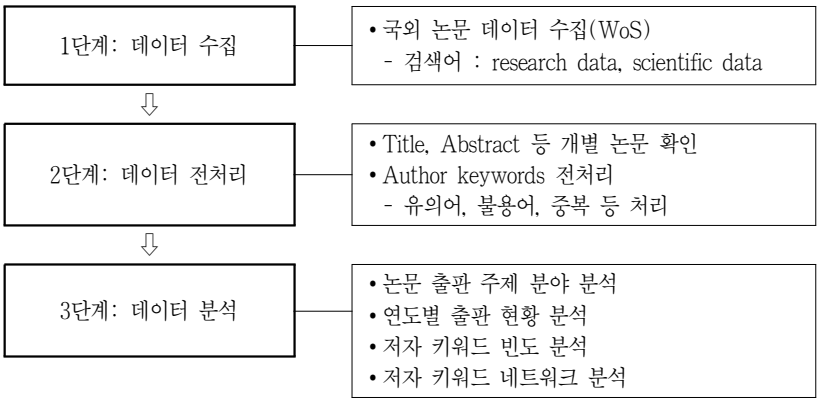
연구동향 분석은 학문의 지속가능한 발전과 미래를 위해 반드시 필요한 작업으로, 누적된 연구결과를 추적하여 학문의 발전 양상 및 특성을 파악할 수 있고, 선행연구의 가치를 이해하고 후속연구의 기반을 마련하는데 중요한 수단으로 이해할 수 있다(배나윤, 오효정, 2024). 국내에서도 여러 학문 분야에서 연구동향 분석을 수행하고 있으며, 문헌정보학 분야에서도 여러 연구가 진행된 바 있다(이세나 외, 2023; 임정훈, 2022; 최재은, 2024; 한상우, 2023). 연구동향 분석에서 주로 이용되는 방법은 키워드 네트워크 분석으로 이는 사회 네트워크 분석 방법을 내용연구 및 문헌연구 분야에 적용한 방법으로 단어와 단어 사이의 관계를 링크로 표시함으로써 구축되는 네트워크를 통해 메시지를 해석하는 분석기법이다. 또한, 구조 분석, 중심성 분석 등을 통해 선행 연구의 주제 및 핵심어, 키워드 사이의 연결 관계를 파악하여 텍스트에 숨어있는 의미, 관계, 연구 동향, 개념구조 등의 특징을 시각적으로 분석할 수 있는 장점이 있다(김준현, 2015; 임정훈, 2022, 한상우, 2023). 키워드 네트워크 분석을 이용한 연구는 연구동향 분석의 목적에 맞게 연구동향을 파악하고 향후의 연구방향을 제시할 수 있으며, 기존의 데이터를 재활용할 수 있는 장점이 있으며 논리적으로 결과를 해석할 수 있다는 점에서 효과적이고 효율적인 연구방법으로 평가받고 있다.

3. 연구방법

연구데이터 관련 국외 연구 동향을 분석하기 위해서 논문의 저자 키워드를 수집하고 빈도분석 및 키워드 네트워크 분석을 수행하였다. 이를 위한 연구 절차는 다음의 절차로 진행되었다(〈그림 1〉 참조).

3.1 데이터 수집

본 연구에서는 연구데이터 관련 국외 연구 동향을 분석하기 위하여 Web of Science(WoS)의 Core Collection을 이용하였다. 상세검색 기능에서 Title, Abstract, Author keywords, Keywords plus를 통합하여 검색이 가능한 Topic 필드를 이용하였으며, 검색어는 'Research data', 'Scientific data'를 사용하여 검색하였다. 또한, 'Data Repository', 'Research Data Management' 등의 연관 검색어로 광범위하게 검색을 수행하였으며, 대상 기간은 2000년부터 2023년까지 총 24년간 생산된 연구논문



〈그림 1〉 연구 단계 개요

을 대상으로 하였다. 이를 위한 검색식은 다음과 같다.

"research data"(Topic) or "scientific data" (Topic) or "research data management"(Topic) and Review Article(Exclude-Document Types)

그 결과 Article 18,213건, Proceedings Paper 6,526건, Review Article 3,047건, Editorial Material 526건, Data Paper 203건, Early Access 457건 등 19개 유형 총 29,463건이 1차로 검색되었으며, 이 중 Article, Data Paper, Proceedings Paper, Early Access 등의 유형으로 필터링하여 25,399건의 연구논문을 데이터 전처리 대상으로 우선 선정하였다.

3.2 저자 키워드 전처리

총 25,399건의 논문을 대상으로 Author Keywords, Title, Abstract 등을 검토하여 연구데이터 관련 논문을 추출하였다. 연구데이터라는 핵심키워드의 특성으로 인해 자연과학 분야의 다양한 실험 결과를 발표한 논문이 다수인 관계로 대상 논문의 내용을 자세히 검토할 필요가 있었다. 해당 작업을 통해 저자 키워드 분석을 위해서 최종적으로 693건의 논문을 선정하였다. 다음으로 논문에 부여된 저자 키워드를 정제하는 작업을 진행하였다. WoS에는 저자가 부여한 Author keywords와 WoS 추천 키워드인 Keywords plus가 구분되어 있으나 본 연구에서는 Author keywords만 대상으로 하였으며, 유사한 키워드가 부여되어 있는 경우 일반적으로 해당 키워드의 의미를 포함할 수 있는 포괄적인 키워드를 선택하여 전처리를 진행하였으며, 전처리 작업을 위해 Excel을 이용하여 반복되는 키워드를 체크하고 용어 변환 과정을 하였으며, 제목과 초록을 참고하여 일일이 확인하는 절차를 거쳤다. 이에 따라 1차로 총 3,502개의 저자 키워드를 추출하였고, 이를 대상으로 〈표 1〉의 내용과 같이 유의어 처리, 용어 분리, 불용어 처리 등의 과정을 거쳐 총 2,646개의 저자 키워드로 정제하였다. 이 중 중복된 키워드를 제외하고 최종적으로 분석에 사용한 저자 키워드는 총 754개이다.

〈표 1〉 저자 키워드 전처리 결과 요약

구분	사례	전처리 결과
유의어 처리	(Open) Research Data(sets), Scientific Data, Science Data,	Research Data
	(Open) Data Repositories (Storage), (Research) Data Repository, Scientific Data Repositories, Digital Repository, Open Access Repository, Institutional Data Repository, Data Archives, Re3Data	Data Repository
	(Research) Data Management (Practices, Policy, Plan), Research Data Services (Governance, Infrastructure)	Research Data Management
	Open (Big) Data	Open Data
	(Research) Data Sharing, Sharing, Information Sharing	Data Sharing
	(Research) Data Curation (Management), Digital (Data) Curation	Data Curation
	Open (Source) Software, (Research) Software (Tools)	Open Software
	(University) Researchers, Scholars, faculty	Researchers
	Open (Scientific) Science(Research)	Open Science
	Copyright, Rights, Property rights, Author's rights	Copyright
	FAIR (meta)(data), FAIR (research) data (management), FAIR (data) principles, FAIR data point, FAIR digital object, FAIR data architecture	FAIR
	(Open, Research) Data (Re)Use, Re-use	Data Reuse
	(Research) Data Preservation, Conservation	Data Preservation
	Publishing data, Data publication	Data Publication
용어 분리	Social Sciences and Humanities 등	Social Science, Humanities 등
불용어 처리	핵심주제어(Research Data)	삭제
	고유명사(국가명, 기관명, 학회명, 서비스명 등)	
	특별한 의미 없는 일반 용어(case study, survey 등)	

저자 키워드를 전처리하는 과정에서 핵심 키워드로 판단할 수 있는 'research data'는 불용어로 처리하였다. 이는 연구데이터와 관련된 연구 동향을 분석함에 있어 해당 키워드는 모든 articles에 포함되어 있을 가능성이 매우 높아 'research data'에 연결정도 중심성이 집중되어 유의미한 분석을 수행하기가 어렵기 때문이다.

정제된 저자 키워드로 키워드 네트워크 분석에 적합하도록 2-mode network 데이터 세트로 가공하여 빈도분석, 연결정도 중심성, 근접중심성, 매개 중심성, 응집구조 분석 등을 수행하였으며, 이를 위해 키워드 네트워크 분석 도구인 NetMiner 4를 이용하였다.

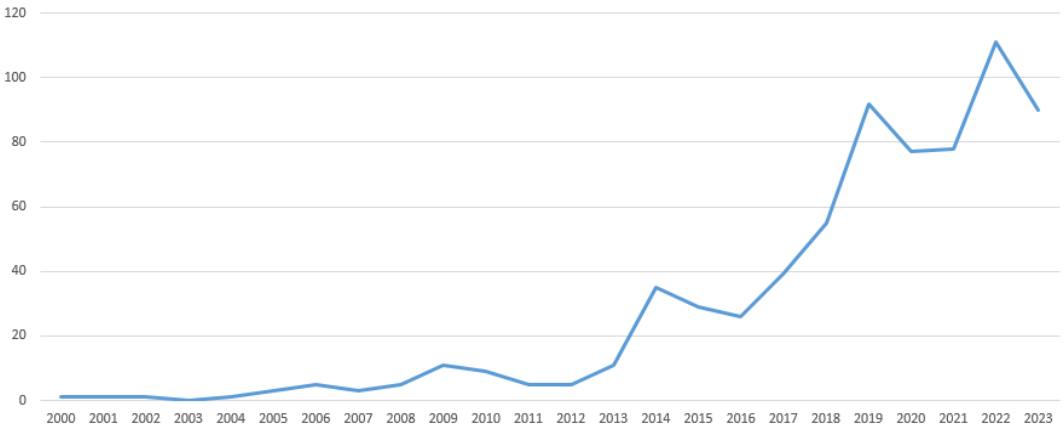
4. 연구결과 분석

4.1 빈도분석

연구데이터 관련 국외 연구논문의 수집 및 전처리 결과 2000년부터 2023년까지 생산된 총 693건의 논문에서 총 754개의 저자 키워드를 수집하여 분석을 수행하였다. 우선, 연구데이터 관련 논문의 유형은 Articles 437건, Data Paper 1건, Early Access 17건, Proceedings Paper 238건 등이며, 발행 건수는 2012년까지 10건 이하로 발행되다가 2014년에 35건으로 크게 증가하였다. 이는 2010년대에 접어들면서 연구데

이터의 중요성 인식, 데이터관리계획의 필수 제출 등 연구데이터와 관련된 상황이 변화하면서 많은 연구가 진행되고 발표된 것으로 분석할 수 있다. 이후 2019년 92건, 2020년 77건, 2021년 78건, 2022년 111건, 2023년 90건 등 최근 5년 이내의 논문수는 대상 기간 중 전체 발행 건수의 절반을 넘어서고 있어 최근의 연구데이터와 관련된 관심과 중요성에 대한 인식이 높아졌음을 이해할 수 있다(〈그림 2〉 참조).

다음으로 출판된 논문의 연구분야는 〈표 2〉에 정리된 바와 같이 가장 높은 빈도를 보인 분야는 Computer Science로 303건의 논문이 발행되었으며, Information Science 분야는 221건으로 나타났다.



〈그림 2〉 연구데이터 관련 국외 논문 연도별 발행 건수

〈표 2〉 주제 분야별 연구데이터 관련 국외 논문 발행 건수(WoS Categories 기준)

주제 분야	빈도수	주제 분야	빈도수
Agriculture	6	Engineering	6
Anesthesiology	2	Environmental Science	5
Anthropology	2	Ethics	5
Archaeology	1	Health	7
Automation & Control Systems	3	Humanities	8
Bio	13	Information Science	221
Business & Management	6	Language & Literature	4
Chemistry	14	Mathematics	6
Communication	16	Medical	3
Computer Science	303	Multidisciplinary	5
Ecology	2	Psychology	6
Economics	5	Social Sciences	6
Education	13	Etc.	23
Energy, Electro chemistry	2	계	693

이어서 연구논문에서 나타난 저자 키워드의 빈도를 분석하였다. 분석 결과 가장 높은 빈도로 출현한 키워드는 'Research Data Management'로 총 390번 나타났으며, 'Data Repository'(143), 'Data Sharing'(117), 'FAIR'(86), 'Open Science'(73), 'Libraries'(67), 'Open Data'(53), 'Metadata'(51), 'Open Access'(50) 등의 순으로 높게 나타났으며, 10번 미만으로 나타난 저자 키워드는 총 718개, 1번만 나타난 저자 키워드는 568개로 나타났다. 저자 키워드의 빈도수가 높게 나타나는 것은 많은 연구에서 해당 키워드가 핵심적인 의미를 가지고 있음을 보여주는 것이며, 낮은 빈도수를 보이는 키워드는 상대적으로 중요성이 떨어지거나 아직 해당 키워드를 중심으로 많은 연구가 진행되지 않았음을 의미하는 것으로 분석할 수 있다(〈표 3〉 참조).

4.2 중심성 분석

연결정도(Degree) 중심성을 분석하는 것은 키워드 네트워크 내에서 어떤 노드에 집중되어 있는지 파악함으로써 연결된 수가 가장 많은 것은 핵심적인 키워드라고 판단할 수 있는 근거가 되며, 근접(Closeness) 중심성은 노드 간의 최단거리를 기반으로 중심성을 분석하므로 노드에 가장 빨리 영향을 주고받는 것을 분석할 수 있다. 또한, 매개(Betweenness) 중심성은 노드 간의 상호의존에 근거하여 중심성을 파악하므로 어떤 루트를 통해 정보가 연결되고 이동하는지를 파악할 수 있다.

저자 키워드의 각 중심성을 분석한 결과, 'Research Data Management', 'Data Repository', 'Data Sharing' 등은 빈도분석에서도 가장 높

〈표 3〉 저자 키워드 빈도분석 결과

저자 키워드	빈도수	비율(%)	저자 키워드	빈도수	비율(%)
Research Data Management	390	14.74	Big Data	17	0.64
Data Repository	145	5.48	Collaboration	16	0.60
Data Sharing	117	4.42	Research Support	16	0.60
FAIR	86	3.25	Data Infrastructure	15	0.57
Open Science	73	2.76	Open Software	14	0.53
Libraries	67	2.53	Ontology	14	0.53
Open Data	53	2.00	Data Governance	14	0.53
Metadata	51	1.93	Research Ethics	13	0.49
Open Access	50	1.89	Data Librarian	12	0.45
Data Curation	46	1.74	Data Visualization	12	0.45
Data Reuse	39	1.47	Scholarly Communication	12	0.45
Data Preservation	32	1.21	Reproducibility	12	0.45
Data Quality	24	0.91	Social Science	11	0.42
Data Literacy	21	0.79	Interoperability	11	0.42
Data Life Cycle	19	0.72	Health Information	11	0.42
Linked Open Data	19	0.72	Data Science	11	0.42
Data Citation	19	0.72	Data Stewardship	11	0.42
Data Publication	19	0.72	Data Librarianship	11	0.42
Researchers	17	0.64	e-Science	10	0.38

* 빈도수 10 이상 출현 순위별

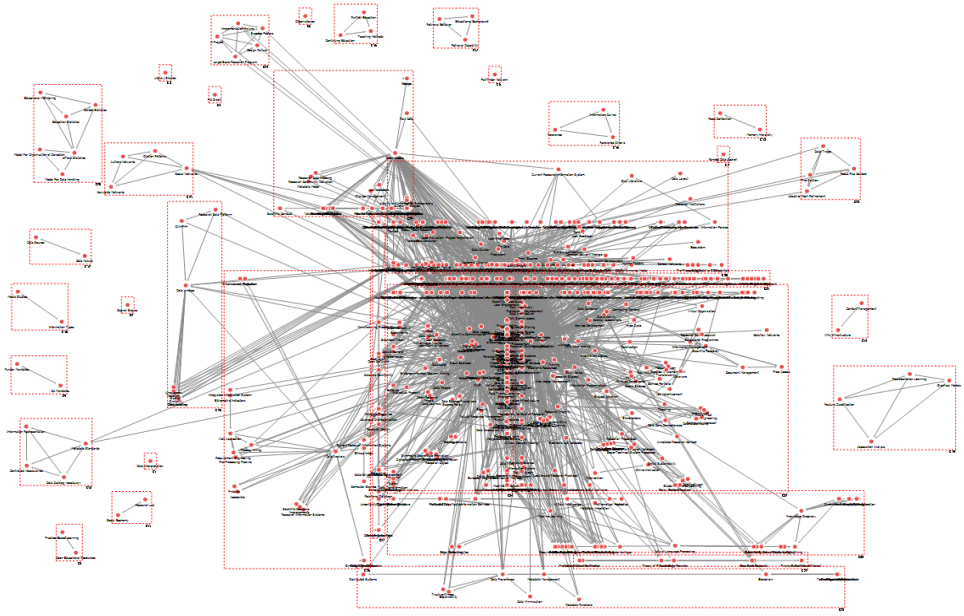
〈표 4〉 저자 키워드의 중심성 지수 분석 결과

키워드	연결정도	근접	매개	출현빈도
Research Data Management	0.549	0.663	0.493	1
Data Repository	0.243	0.533	0.115	2
Data Sharing	0.197	0.510	0.094	3
FAIR	0.173	0.498	0.052	4
Open Science	0.143	0.489	0.037	5
Open Data	0.134	0.492	0.039	7
Metadata	0.127	0.487	0.040	8
Open Access	0.113	0.476	0.029	9
Data Reuse	0.106	0.474	0.039	11
Libraries	0.104	0.467	0.020	6
Data Curation	0.089	0.476	0.011	10
Data Quality	0.080	0.469	0.033	13
Linked Open Data	0.069	0.460	0.009	16
Collaboration	0.069	0.452	0.017	21
Data Preservation	0.068	0.462	0.009	12

은 출현 빈도를 보였으며, 연결정도, 근접, 매개 중심성 모두 가장 높은 결과를 보였다. 이는 연구데이터와 관련된 연구가 3개의 키워드를 중심으로 가장 많이 진행되어 왔고, 동시에 출현하는 빈도도 높으며, 해당 키워드를 통해 다양한 연구들이 서로 연결되어 있음을 나타내는 것으로 이해할 수 있다.

4.3 응집구조 분석

연구데이터 관련 저자 키워드 간의 군집 상태를 확인하기 위해서 community 내의 modularity 분석을 수행하였고, 그 결과 총 34개의 그룹이 나타났다. 그룹의 개수 및 그룹화의 적합 수준을 판단하기 위해서는 modularity 값을 이용하는데, -1~1 사이의 양수값을 가지면 모듈화되었다고 판단하고 있다. 본 연구에서 응집구조 분석을 한 결과값은 0.407로 저자 키워드의 모듈화는 적절한 것으로 나타났다. 응집구조 분석을 수행하는 이유는 키워드 간의 관계뿐만 아니라 전체 네트워크의 하위 그룹이 어떻게 형성되어 있는지를 확인하고 그룹 내에서 핵심적인 키워드는 무엇인지, 하위 그룹 간에는 어떤 관계가 있는지를 파악하여 전체 네트워크의 특징을 파악할 수 있기 때문이다. 총 34개의 저자 키워드 그룹 중 15개는 그룹 간의 연결 관계가 확인되었으며, 그룹 간의 연결 관계가 없이 단독으로 존재하는 그룹은 19개로 나타났다(〈그림 3〉 참조). 연결 관계가 존재하지 않는 19개의 그룹은 대부분 노드수가 5개 이하이며, 빈도분석 결과에서도 낮은 빈도를 보인 키워드로서, 연구논문의 내용적 특성으로 부여되었으며 전처리 과정에서도 불용처리를 하지는 않았으나 연구데이터와 직접적인 관련성이 낮아 내용적 분석 대상에서는 제외하였다(〈표 5〉 참조). 저자 키워드 간의 연결 관계가 확인된 그룹 중 3개의 그룹은 연결된 노드의 수가 5개 이하로 다른 그룹과 유의미한 연결 관계가 나타나지 않아 이를 제외한 12개 그룹(Open Access 그룹, Data Linkage 그룹, Data Analysis 그룹,



〈그림 3〉 저자 키워드 응집구조 분석 결과

〈표 5〉 응집구조 분석 그룹화 결과(노드 수 5개 이하)

그룹	그룹 내 저자 키워드
G1	Data Interpretation
G2	Search Engine
G3	Literary Studies
G4	PID Graph
G5	Observatories
G6	Pahtfinder Network
G7	Ranked Data Search
G8	Practice-based Learning, Open Educational Resources
G9	Funder Mandates, OA Mandates
G10	Media Studies, Information Types
G11	Social Economy, Research Lab
G12	Data Source, Data Nature
G13	Read Contention, Memory Hierarchy
G14	Content Management, XML Infrastructure
G15	Continuing Education, Further Education, Teaching Methods
G16	Relevance, Relevance Criteria, Information Carrier
G17	Retrieval Behavior, Retrieval Capability, Educational Background
G18	Information Representation, Metadata Standard*, Controlled Vocabularies, Data Catalog Vocabulary
G19	Representation Learning, Association Analysis, Graphical Models, Feature Classification
G21	Social Networks*, Keywords Networks, Author Networks, Citation Patterns
G23	Data Triage*, Adaptive Mesh Refinement*, Prioritization*, Focus Plus Context*
G24	IT Project*, Large-scale Research Program*, Design Failure*, Success Factors*, Implementation Failure*

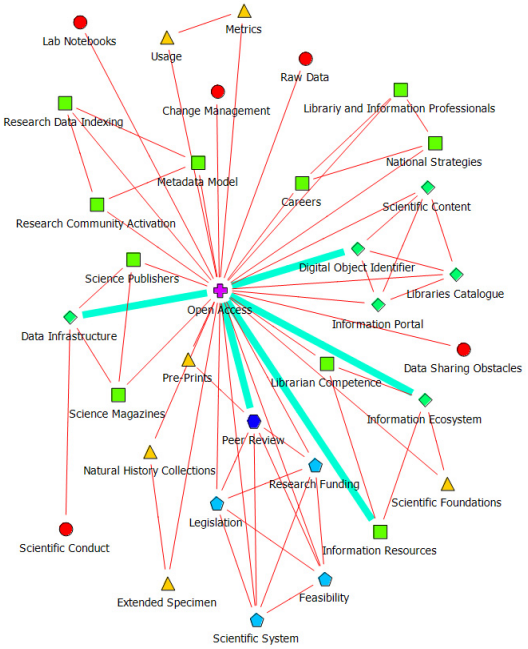
* 다른 저자 키워드 그룹과 관계를 가지고 있는 저자 키워드

Open Software 그룹, Theory 그룹, Socio-Tech 그룹, Computer-AI 그룹, RDM 그룹, Data Utilization 그룹, Data Repository 그룹, Open Data 그룹, Statistics 그룹)을 대상으로 분석을 진행하였다. 이 중 Theory 그룹과 Statistics 그룹은 노드의 특성이 연구데이터의 주제적 특성과 거리가 있는 관련 이론, 기본 통계 기법 등의 내용으로 이를 제외한 나머지 그룹의 특성을 분석하였으며, 내용은 다음과 같다.

Open Access 그룹에는 총 33개의 저자 키워드로 구성되어 있으며, 연결정도 중심성의 평균값은 0.121이다. 해당 그룹에서 가장 높은 연결정도 중심성을 보이는 핵심 키워드는 'Open Access'이며, 이와 관련된 'Peer Review', 'Feasibility', 'Information Ecosystem' 등의 키워드가 높은 연결 관계를 가지고 있는 것으로 나타났다. 또

〈표 6〉 Open Access 그룹의 키워드 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Open Access	0.969	
Peer Review	0.188	
Scientific System	0.156	
Legislation	0.156	
Research Funding	0.156	
Feasibility	0.156	
Data Infrastructure	0.125	
Information Ecosystem	0.125	
Digital Object Identifier	0.125	
Scientific Content	0.125	
Libraries Catalogue	0.125	
Information Portal	0.125	
Information Resources	0.094	
Science Magazines	0.094	
Science Publishers	0.094	
Librarian Competence	0.094	
Metadata Model	0.094	
Research Community Activation	0.094	
Research Data indexing	0.094	
Careers	0.094	
LIS Professionals	0.094	
National Strategies	0.094	
Scientific Foundations	0.063	
Extended Specimen	0.063	
Natural History Collections	0.063	
Pre-Prints	0.063	
Usage	0.063	
Metrics	0.063	
Data Sharing Obstacles	0.031	
Scientific Conduct	0.031	
Change Management	0.031	
Lab Notebooks	0.031	
Raw Data	0.031	



한, 그룹 내의 핵심 키워드인 'Open Access'는 전체 네트워크에서도 높은 빈도로 출현하며, 연결정도, 근접, 매개 중심성 모두 상대적으로 높아 연구데이터 관련 연구에서 핵심 키워드의 위치에 있음을 이해할 수 있다.

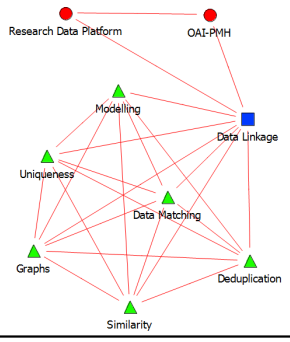
Data Linkage 그룹에는 총 9개의 키워드가 포함되어 있으며, 연결정도 중심성의 평균값은 0.667이다. 동 그룹에서는 'Data Linkage' 키워드가 모든 노드와 연결되어 있으며, 내용상 데이터의 연결, 매칭, 유사성, 모델링 등의 개념들로 연관되어 있음을 이해할 수 있었다. 'Data Linkage' 키워드는 전체 키워드 네트워크 상에서도 비교적 높은 연결정도 중심성(0.017)을 가지고 있어 연구데이터 관련 연구에서 유의미한 위치에 있음을 이해할 수 있다(〈표 7〉 참조).

Data Analysis 그룹은 총 20개의 저자 키워드로 구성되어 있으며, 연결정도 중심성의 평균값은 0.195이다. 그룹 내에서는 'Data Analysis'를 중심으로 'Research Information System', 'Scientific Decisions', 'Pre-processing Pipeline', 'Improvement' 등 데이터의 분석과 연관된 시스템, 분석 전처리 파이프라인, 과학적 결정을 통한 개선 등의 키워드가 서로 밀접하게 연관되어 있음을 나타내고 있다(〈표 8〉 참조).

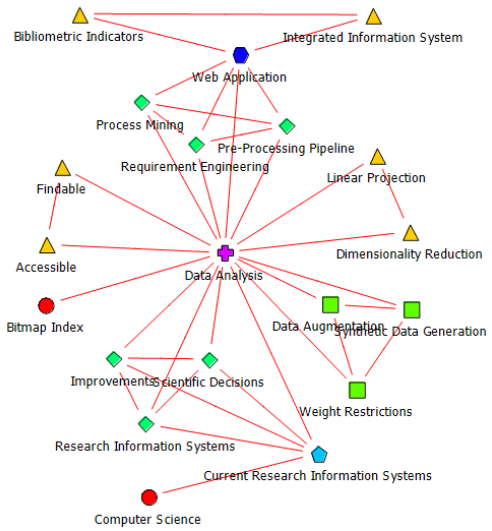
Open Software 그룹에는 총 21개의 저자 키워드가 포함되어 있으며, 그룹의 연결정도 중심성 평균값은 0.167이다. 해당 그룹에서는 'Open Software' 키워드를 중심으로 연결되어 있으며, 'Digitisation', 'Data-driven Process Development', 'Data Filtering', 'Accessibility Literacy' 등의 키워드와 높은 연결정도를 가지고 있는 것으로 나타났다. 'Open Software' 키워드는 전체 네트워크 내의 연결정도 중심성 지수가 0.042로 매우 높은 편으로 많은 키워드와 연결관계를 형성하고 있음을 이해할 수 있다(〈표 9〉 참조).

Socio-Tech 그룹에는 총 12개의 키워드가 포함되어 있으며, 연결정도 중심성 평균값은 0.273이다. 핵심 키워드는 'Data Provenance'로 'Metadata', 'Explainability', 'Trustworhiness' 등과 연관되어 데이터의 출처, 확장, 신뢰 등의 개념과 연관성이 높은 것을 알 수 있었다(〈표 10〉 참조).

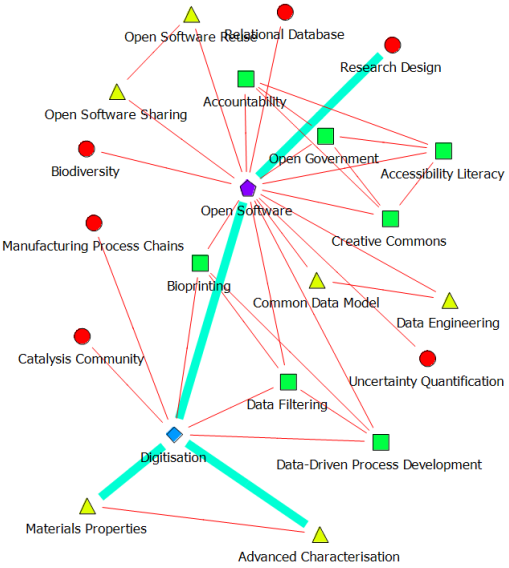
〈표 7〉 Data Linkage 그룹의 키워드 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Data Linkage	1.000	
Data Matching	0.750	
Deduplication	0.750	
Similarity	0.750	
Graphs	0.750	
Modeling	0.750	
Uniqueness	0.750	
Research Data Platform	0.250	
OAI-PMH	0.250	

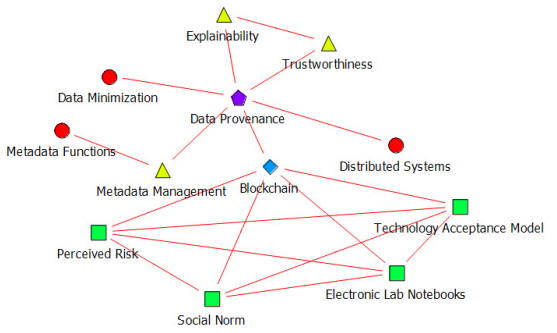
〈표 8〉 Data Analysis 그룹의 키워드 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Data Analysis	0.842	
Web Application	0.316	
CRIS	0.263	
Research Information Systems	0.211	
Improvements	0.211	
Scientific Decisions	0.211	
Pre-Processing Pipeline	0.211	
Process Mining	0.211	
Requirement Engineering	0.211	
Synthetic Data Generation	0.158	
Data Augmentation	0.158	
Weight Restrictions	0.158	
Accessible	0.105	
Findable	0.105	
Bibliometric Indicatiors	0.105	
Integrated Information System	0.105	
Dimensionality Reduction	0.105	
Linear Projection	0.105	
Computer Science	0.053	
Bitmap Index	0.053	

〈표 9〉 Open Software 그룹의 키워드 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Open Software	0.800	
Digitisation	0.400	
Bioprinting	0.200	
Data-Driven Process Development	0.200	
Data Filtering	0.200	
Accountability	0.200	
Open Government	0.200	
Accessibility Literacy	0.200	
Creative Commons	0.200	
Data Engineering	0.100	
Common Data Model	0.100	
Advanced Characterisation	0.100	
Materials Properties	0.100	
Open Software Sharing	0.100	
Open Software Reuse	0.100	
Uncertainty Quantification	0.050	
Relational Database	0.050	
Research Design	0.050	
Biodiversity	0.050	
Catalysis Community	0.050	
Manufacturing Process Chains	0.050	

〈표 10〉 Socio-Tech 그룹의 키워드 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Data Provenance	0.545	
Blockchain	0.455	
Technology Acceptance Model	0.364	
Electronic Lab Notebooks	0.364	
Perceived Risk	0.364	
Social Norm	0.364	
Metadata Management	0.182	
Explainability	0.182	
Trustworthiness	0.182	
Metadata Functions	0.091	
Data Minimization	0.091	
Distributed Systems	0.091	

Computer-AI 그룹에는 총 36개의 저자 키워드가 포함되어 있으며, 연결정도 중심성의 평균값은 0.117이다. 'Machine Learning' 키워드가 가장 높은 연결정도를 보이고 있으며, 'Knowledge Discovery', 'Artificial Intelligence', 'Natural Language Processing', 'Data Mining', 'Deep Learning', 'Artificial Neural Networks', 'Predictive Models' 등 최근의 AI 관련된 키워드가 다수 군집하여 있음을 볼 수 있다(〈표 11〉 참조).

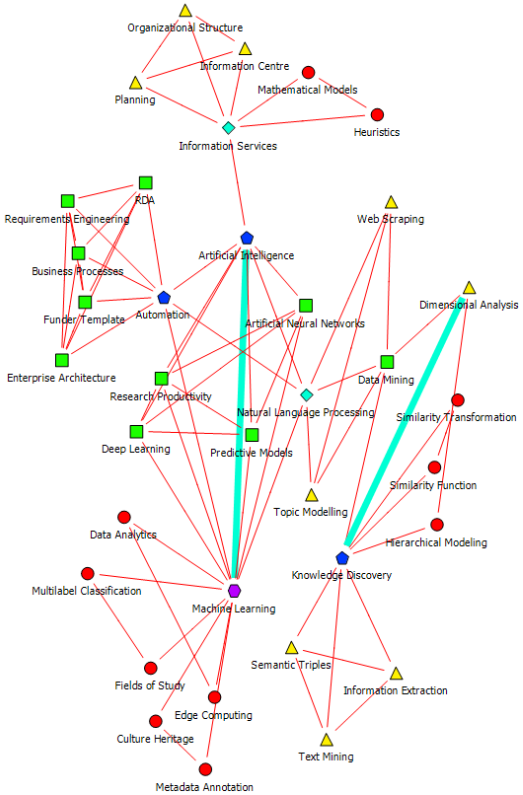
RDM(Research Data Management) 그룹은 총 169개의 키워드가 포함되어 있는 가장 큰 규모의 그룹이며, 연결정도 중심성의 평균값은 0.027이다. 핵심 키워드는 'Research Data Management'로 빈도분석, 연결정도, 근접, 매개 중심성 모두에서 가장 높게 나타나 연구데이터 관련 연구에서 가장 핵심적인 지위를 차지하고 있는 것으로 분석할 수 있다. 물론 본 연구에서 핵심 키워드로 선정한 것은 'Research Data'지만 연구데이터와 관련된 모든 행위를 포괄하는 개념은 연구데이터 '관리'가 될 수밖에 없어 거의 모든 연구에서 해당 키워드가 출현하는 것으로 이해할 수 있다. 아울러, 동 그룹에는 'Libraries', 'Data Literacy', 'Data Librarianship', 'Knowledge Management' 등 문헌정보학 관련 개념들이 높은 연결정도 중심성을 보이고 있어 문헌정보학 분야에서 연구데이터 관련 연구가 다수 진행되었으며, 연구데이터의 관리와 서비스 측면에서 중요한 위치를 차지하고 있다고 분석할 수 있다(〈표 12〉 참조).

Data Utilization 그룹에는 총 102개의 저자 키워드가 포함되어 있으며, 연결정도 중심성의 평균값은 0.035이다. 주요 키워드로는 'Data Sharing', 'Data Visualization', 'Data Integration', 'Evaluation', 'Analytics' 등 데이터의 활용과 관련되어 있으며, 'Data Sharing'은 전체 저자 키워드 네트워크 내에서 높은 빈도로 출현하며 연결정도, 근접, 매개 중심성 모두 높게 나타나 연구데이터 관련 연구에서 핵심적인 키워드로 이해할 수 있다(〈표 13〉 참조).

Data Repository 그룹은 총 151개의 저자 키워

〈표 11〉 Computer-AI 그룹의 키워드 및 네트워크 지도

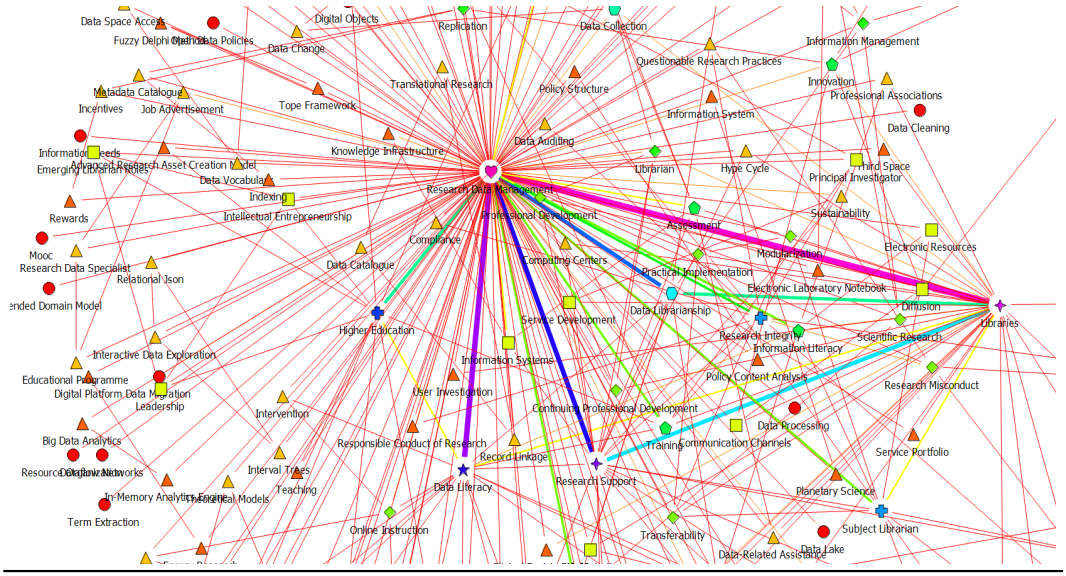
키워드	연결정도 중심성	네트워크 지도
Machine Learning	0.371	
Knowledge Discovery	0.229	
Artificial Intelligence	0.229	
Automation	0.229	
Natural Language Processing	0.171	
Information Services	0.171	
Data Mining	0.143	
Deep Learning	0.143	
Artificial Neural Networks	0.143	
Research Productivity	0.143	
Predictive Models	0.143	
Business Processes	0.143	
Enterprise Architecture	0.143	
Funder Template	0.143	
Requirement Engineering	0.143	
RDA	0.143	
Topic Modeling	0.086	
Web Scraping	0.086	
Information Extraction	0.086	
Semantic Triples	0.086	
Text Mining	0.086	
Planning	0.086	
Organizational Structure	0.086	
Information Centre	0.086	
Dimensional Analysis	0.086	
Metadata Annotation	0.057	
Culture Heritage	0.057	
Data Analysis	0.057	
Edge Computing	0.057	
Fields of Study	0.057	
Multilabel Classification	0.057	
Mathematical Models	0.057	
Heuristics	0.057	
Hierarchical Modeling	0.057	
Similarity Transformation	0.057	
Similarity Function	0.057	



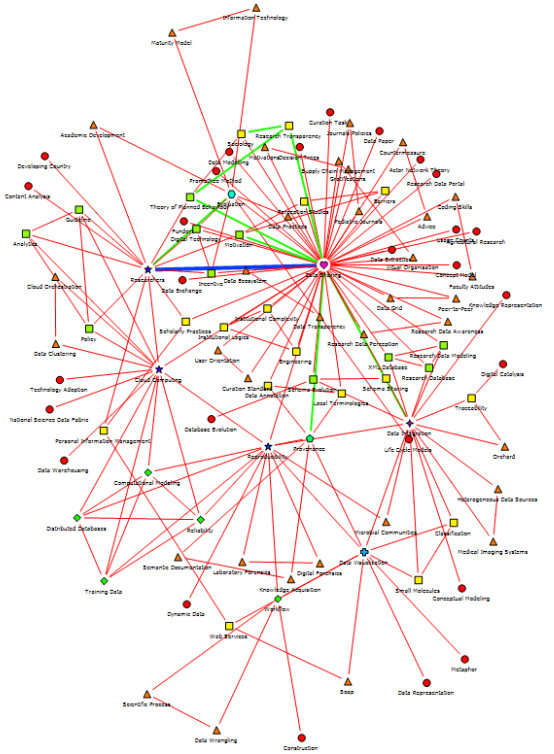
〈표 12〉 RDM 그룹의 키워드(상위 40개) 및 네트워크 지도

키워드	연결정도 중심성	키워드	연결정도 중심성
Research Data Management	0.911	Outreach	0.036
Libraries	0.238	Data Management Skills	0.036
Research Support	0.143	Data Management Training	0.036
Data Literacy	0.120	Metadata Behaviors	0.036
Higher Education	0.083	Tagging Behaviors	0.036
Subject Librarian	0.071	Replication	0.036
Research Integrity	0.071	Librarian	0.036
Data Librarianship	0.054	Modularization	0.030
Data Collection	0.048	Practical Implementation	0.030
Assessment	0.042	Transferability	0.030
Stakeholders	0.042	Scientific Research	0.030
Information Literacy	0.042	Online Instruction	0.030
Knowledge Management	0.042	Electronic Lab Notebook	0.030
Innovation	0.042	Electronic Research Notebook	0.030
Training	0.042	Socio-Technical System Processes	0.030
Information Management	0.036	Research Misconduct	0.030
Datafication	0.036	Continuing Professional Development	0.030
Liaison Librarian	0.036	Professional Development	0.030
Instruction Services	0.036	Document Management	0.024
Collection Development	0.036	RDS Maturity	0.024

네트워크 지도(일부)



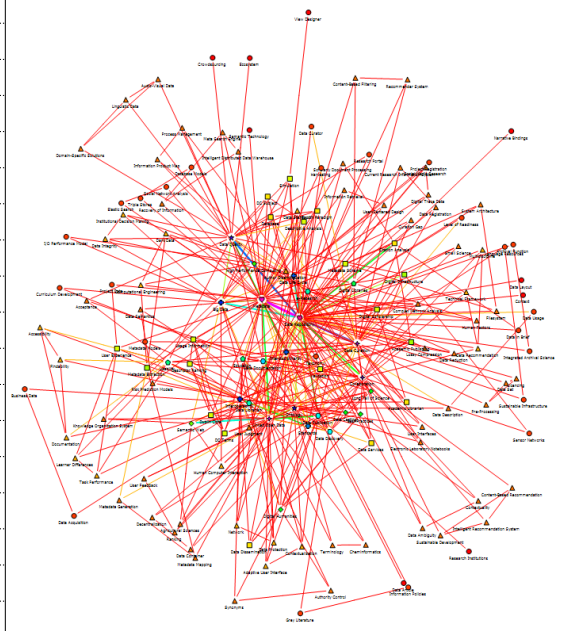
〈표 13〉 Data Utilization 그룹의 키워드(상위 30개) 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Data Sharing	0.564	
Data Integration	0.168	
Researchers	0.149	
Cloud Computing	0.149	
Reproducibility	0.139	
Data Visualization	0.099	
Evaluation	0.089	
Provenance	0.069	
Training Data	0.050	
Computational Modeling	0.050	
Reliability	0.050	
Distributed Databases	0.050	
Workflow	0.050	
Schema Evolution	0.040	
Policy	0.040	
Analytics	0.040	
Guideline	0.040	
Motivation	0.040	
Theory of Planned Behavior	0.040	
Digital Technology	0.040	
Incentive	0.040	
Research Database	0.040	
XML Database	0.040	
Research Data Modeling	0.040	
Traceability	0.030	
Local Terminologies	0.030	
Data Annotation	0.030	
Small Molecules	0.030	
Classification	0.030	
Institutional Logics	0.030	

드로 구성되어 있으며, 연결정도 중심성의 평균값은 0.042이다. 핵심 키워드는 'Data Repository'로 동 그룹에서는 물론 전체 저자 키워드 네트워크 내에서도 0.243의 연결정도 중심성을 보여 중요한 위치를 차지하고 있어 연구데이터 관련 연구에서 핵심적인 키워드로 사용되고 있음을 알 수 있었다. 또한, 'Data Repository'는 'Data Curation', 'Metadata', 'Collaboration', 'Data Quality', 'Data Life Cycle' 등의 키워드 간에 weight가 높은 것으로 나타났다(〈표 14〉 참조).

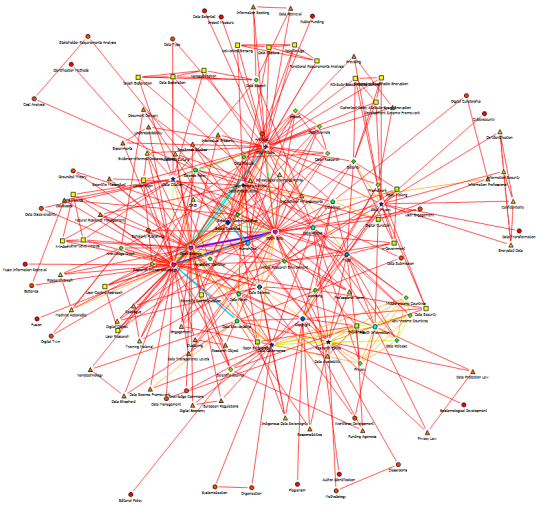
Open Data 그룹은 총 136개의 저자 키워드로 구성되어 있으며, 연결정도 중심성의 평균값

〈표 14〉 Data Repository 그룹의 키워드(상위 30개) 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
Data Repository	0.473	
Metadata	0.387	
Linked Open Data	0.260	
Data Curation	0.227	
Collaboration	0.193	
Data Quality	0.180	
Ontology	0.167	
Interoperability	0.133	
Big Data	0.133	
Data Life Cycle	0.127	
Interdisciplinary	0.127	
Standards	0.113	
Data Documentation	0.107	
Data Publication	0.100	
Data Discovery	0.100	
Data Librarian	0.093	
e-Research	0.093	
Education	0.087	
Usability	0.087	
Digital Libraries	0.080	
Digital Humanities	0.067	
Best Practices	0.067	
Data Casting	0.067	
Long Tail of Science	0.067	
High Performance Computing	0.060	
Semantic Web	0.053	
Digital Infrastructure	0.047	
Academic Publishing	0.047	
Metadata Extraction	0.047	
Validation	0.040	

은 0.049이다. 본 그룹에서는 'FAIR' 키워드가 가장 핵심적인 위치를 차지하고 있으며, 'Open Data', 'Open Science', 'Data Stewardship', 'Data Reuse', 'Data Preservation' 등의 키워드와 높은 weight를 가지고 있어 그룹 내에서 중요한 개념으로 사용되었음을 이해할 수 있다. 또한, 'FAIR'와 'Open Data'는 전체 저자 키워드 네트워크에서도 높은 연결정도 중심성을 가지고 있어 많은 연구에서 중요한 키워드와 함께 출현하고 있으며, 관련 연구도 많이 진행되었다고 판단할 수 있다(〈표 15〉 참조).

〈표 15〉 Open Data 그룹의 키워드(상위 30개) 및 네트워크 지도

키워드	연결정도 중심성	네트워크 지도
FAIR	0.363	
Open Data	0.333	
Open Science	0.333	
Data Reuse	0.304	
Data Preservation	0.163	
Data Governance	0.156	
Data Citation	0.148	
Research Ethics	0.148	
Data Privacy	0.148	
Social Sciences	0.133	
Copyright	0.126	
Data Science	0.119	
Trust	0.119	
Humanities	0.111	
Health Information	0.096	
Data Access	0.089	
Data Policy	0.081	
Data Stewardship	0.081	
e-Science	0.081	
Science Policy	0.081	
Scholarly Communication	0.074	
Virtual Research Environment	0.074	
Security	0.074	
Knowledge Graph	0.059	
Persistent Identifier	0.059	
Data Papers	0.059	
Data Search	0.059	
Licensing	0.059	
Data Release	0.059	
Scientific Journal	0.052	
Framework	0.052	
Research Infrastructure	0.052	

5. 결 론

본 연구에서는 연구데이터 관련 국외 연구의 동향을 분석하고 결과를 바탕으로 국내의 연구 동향과 비교 분석하여 연구데이터 관련 연구의 특성과 발전방향을 이해하고 관련 연구의 지속적 수행을 위한 근거를 제시하고자 하였다. 이를 위해 WoS에서 2000년부터 2923년까지 연

구데이터 관련 논문 총 693건에서 저자 키워드 총 754개를 추출하여 빈도분석과 저자 키워드 네트워크 분석을 수행하였으며, 그 결과를 다음과 같이 정리하였다.

첫째, 국외의 연구데이터 관련 연구는 비교적 활성화되어 있다. 한상우(2023)의 연구에서 국내의 연구데이터 관련 연구 동향을 분석했는데, 동 기간 내 국내의 연구 건수가 58건에 그친 것에 비하면 상대적으로 많은 연구가 진행되었다고 볼 수 있다. 다만, 국내와 국외 전체를 단적으로 비교하여 국내가 연구가 부족하다고 일반화하여 단정지을 수는 없다.

둘째, 국외의 연구데이터 관련 연구도 2014년 이후 비교적 최근에 활발하게 진행되었으며, 연구 분야는 대부분 Computer Science와 Information Science 분야에 집중되는 것으로 나타났다. WoS의 초기 검색 결과에서는 자연과학 분야의 논문이 많이 출현하였으나, 실험 및 관찰 등을 통한 실험적 연구데이터와 관련된 내용으로 본 연구의 분석 대상에서 제외되어 특정 분야로 집중된 것으로 이해할 수 있다.

셋째, 'Research Data Management', 'Open Data', 'Data Repository' 등의 키워드는 높은 빈도로 출현하며 연결정도, 근접, 매개 중심성 등이 높게 나타나 많은 연구에서 핵심적인 키워드로 사용되었음을 알 수 있었고, 아울러 관련된 연구도 많이 진행된 것으로 판단할 수 있었다. 따라서 관련된 키워드 중 상대적으로 낮은 빈도로 출현하는 키워드를 살펴보고 이에 대한 추가 연구를 진행하는 것도 연구데이터와 관련된 전반적 연구 수준 향상에 기여할 것으로 예상된다.

이상의 연구 결과를 바탕으로 연구데이터 관련 연구 동향 분석에 대한 의미를 정리하면 다음과 같다.

첫째, 국외나 국내 모두 연구데이터 관련 연구가 집중적으로 수행된 것은 불과 10여년 정도에 불과하다. 최근 연구데이터 활용과 관련된 체계의 형성, 법제도화 등 연구가 진행되어야 할 분야가 많이 있으므로 지속적인 연구의 진행이 필요하며 다양한 사례 연구도 필요할 것이다.

둘째, 국외 연구데이터 관련 연구에서도 저자 키워드 네트워크 전체에서 연관 관계가 없는 키워드 그룹이 많은 것으로 나타났다. 이는 주요 키워드와 연관성이 없어 관계가 없는 것일 수도 있으나 상대적으로 연구건수가 부족하여 발생하는 현상일 수도 있다. 연구데이터 관련된 다양한 주제의 연구가 진행되어야 할 것이다.

셋째, 연구데이터 관련 국외 연구 동향을 분석함으로써 국내의 연구 동향과 비교 분석이 가능할 수 있을 것으로 판단된다. 국외의 연구데이터 관련 연구는 국내에 비해 활성화되어 있으며, 저자 키워드 역시 다양하고 많은 수로 나타나고 있다. 국내의 연구데이터 관련 연구 동향과 비교했을 때, '연구데이터관리', '데이터리포지터리', '오픈데이터' 등을 중심으로 연구가 진행된 것은 공통적인 현상으로 분석되었다. 다만, 국내와 달리 '오픈액세스', '데이터분석' 등과 관련된 연구가 활성화되어 있음을 확인할 수 있었고, Computer Science 분야의 연구가 다수를 차지하고 있어 Computer-AI 관련 키워드가 높은 것이 국내의 연구와 다른 점으로 이해할 수 있었다. 또한 국내에서 비교적 드물게 나타나고 있는 연구데이터 관련 저자 키워

드를 중심으로 국외의 연구와 비교하여 연구를 진행하기 위한 기초자료로 활용될 수 있다는 점에서 본 연구의 의미를 찾을 수 있을 것이다.

참 고 문 헌

국가연구개발정보처리기준. 과학기술정보통신부고시 제2020-102호.

김준현 (2015). 네트워크 텍스트 분석결과 해석에 관한 소고: 행정학 분야 연구를 중심으로. 인문사회과학연구, 16(4), 247-280. http://doi.org/10.15818/ihss.2015.16.4.247

배나윤, 오효정 (2024). 주요 학문분야 비교를 통한 국내 정보공개 연구동향 분석. 정보관리학회지, 41(2), 295-316. http://dx.doi.org/10.3743/KOSIM.2024.41.2.295

신은정, 최해옥, 김권일 (2024). 국내외 연구데이터 법제도 비교분석 및 개선과제(STEPI Insight 제326호). 과학기술정책연구원.

이세나, 이성신, 백수민 (2023). 저자 키워드와 초록 분석을 통한 법학사서 연구동향 분석. 한국문헌정보학회지, 58(2), 5-31. http://dx.doi.org/10.4275/KSLIS.2024.58.2.005

임정훈 (2022). 키워드 네트워크 분석과 토픽모델링을 활용한 정보활용교육 연구 동향 분석. 정보관리학회지, 39(4), 23-48. http://dx.doi.org/10.3743/KOSIM.2022.39.4.023

최재은 (2024). 텍스트 마이닝을 활용한 국외 데이터 큐레이션 연구 동향 분석. 정보관리학회지, 41(3), 85-107. http://dx.doi.org/10.3743/KOSIM.2024.41.3.085

한나은, 엄정호, 임형준 (2024). 과학기술분야 정부출연연구기관 연구데이터 관리 방안 연구. 한국문헌정보학회지, 58(2), 151-175. http://dx.doi.org/10.4275/KSLIS.2024.58.2.151

한상우 (2023). 키워드 네트워크 분석을 이용한 연구데이터 관련 국내 연구 동향 분석. 한국도서관·정보학회지, 54(4), 393-414. http://dx.doi.org/10.16981/kliss.54.4.202312.393

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

Bae, Nayun & Oh, Hyojung (2024). Analyzing domestic research trends on disclosure of information by comparing major academic disciplines. Journal of Korean Society for Information Management, 41(2), 295-316. http://dx.doi.org/10.3743/KOSIM.2024.41.2.295

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

Han, Naeun, Um, JungHo, & Yim, HyungJun (2024). A study on research data management methods for government-funded research institutes in the field of science and technology. Journal of Korean Society for Library and Information Science, 58(2), 151-175. http://dx.doi.org/10.4275/KSLIS.2024.58.2.151

Han, Sangwoo (2023). An analysis of domestic research trend on research data using keyword network analysis. Journal of Korean Library and Information Science Society, 54(4), 393-414. http://dx.doi.org/10.16981/kliss.54.4.202312.393

Kim, Junhyun (2015). An essay for understanding the meaning of the network text analysis results in study of the public administration. The Journal of Humanities and Social Sciences, 16(4), 247-280. http://doi.org/10.15818/ihss.2015.16.4.247

Lee, Sena, Lee, Seongsin, & Baek, Sumin (2023). An analysis of research trends in law librarians through author keywords and abstract analysis. Journal of Korean Society for Library and Information Science, 58(2), 5-31. http://dx.doi.org/10.4275/KSLIS.2024.58.2.005

Lim, Jeong-Hoon (2022). Analysis of research trends in information literacy education using keyword network analysis and topic modeling. Journal of the Korean Society for Information Management, 39(4), 23-48. http://dx.doi.org/10.3743/KOSIM.2022.39.4.023

National Research and Development Information Processing Standards. Ministry of Science and ICT Notice No. 2020-102.

Shin, Eunjung, Choi, Haeok, & Kim, Kwonil (2024). Recommendations on a national research data regulation, based on a comparative analysis (STEPI Insight v.326). Science and Technology Policy Institute.