텍스트 마이닝을 활용한 국외 데이터 큐레이션 연구 동향 분석

Analysis of Research Trends in Data Curation Using Text Mining Techniques

최재은(Jaeeun Choi)*

초 록

본 연구의 목적은 국외 데이터 큐레이션 연구 동향을 분석하는 것이다. 이를 위해 Scopus와 WoS에서 1,849건의 학술 정보를 추출하였으며 중복 제거 등을 통해 최종 1,797건의 논문, 학술대회 발표자료 등의 표제, 키워드, 초록을 분석 대상으로 하였다. 전처리를 거친 키워드를 빈도분석 하였으며, LDA 토픽 모델링 분석을 통해 주요 주제를 도출하고 토픽의 키워드를 대상으로 네트워크 분석을 통해 중심성을 도출하였다. 키워드 빈도 분석 결과, 'research', 'information' 등이 자주 등장했으며, 이는 데이터 큐레이션이 의학 연구, 생의학 연구 및 연구데이터 관리, 연구 인프라 등 다양한 측면에서 이루어지고 있음을 보여준다. LDA 토픽 모델링을 통해서는 '임상 의료 데이터의 품질 제고와 분석', '빅데이터 관리와 처리 시스템의 효율성 향상', '과학 데이터의 관리와 디지털 리포지터리', '의료 및 생물학적 데이터의 주석과 모델링', '유전자 및 단백질 데이터베이스 연구' 5가지 토픽을 도출하였다. 키워드 네트워크 분석 결과, 'analysis'는 전역 중심성에서 높은 수치를 나타내 데이터 활용 측면에서 분석 방법이나 분석 시스템 등으로 폭넓게 논의되고 있음을 알 수 있었고, 지역 중심성에서는 'research', 'gene', 'system' 등이 상위에 위치한 것으로 나타났다.

ABSTRACT

This study analyzes trends in data curation research. A total of 1,849 scholarly records were extracted from Scopus and WoS, with 1,797 papers selected after removing duplicates. Titles, keywords, and abstracts were analyzed through keyword frequency analysis, LDA topic modeling, and network analysis. Frequent keywords like 'research' and 'information' suggest that data curation is widely applied in medical research, biomedical research, data management, and infrastructure. LDA modeling identified five main topics: improving medical data quality, enhancing big data management, managing scientific data and repositories, annotating and modeling medical data, and gene/protein database research. Network analysis showed that 'analysis' was central in global discussions, while 'gene' and 'system' were locally central. These findings highlight the importance of data curation in various research areas.

키워드: 데이터 큐레이션, 연구동향, 토픽 모델링, LDA, 네트워크분석
data curation, research trends, topic modeling, LDA, network analysis

* 이화여자대학교 문헌정보학과 박사과정(jaeeunchoi2@ewhain.net)

■ 논문접수일자 : 2024년 8월 14일 ■ 최초심사일자 : 2024년 8월 29일 ■ 게재확정일자 : 2024년 9월 3일
■ 정보관리학회지, 41(3), 85-107, 2024. http://dx.doi.org/10.3743/KOSIM.2024.41.3.085

1. 서 론

데이터는 연구의 재현성과 신뢰성을 담보하며 재사용을 통해 새로운 연구를 가능하게 하기 때문에 매우 중요한 요소라고 할 수 있다. 미국 과학기술정책실은 개방적이고 공정한 연구를 확산시키기 위해 2023년을 오픈사이언스의 해로 선포하였다. 이에 NIH(National Institute of Health)와 같은 미국 주요 연방기관은 공적자금이 투입된 연구데이터에 대한 접근과 공유 계획을 수립하였다(이민정, 2023).

국내에서도 과학기술정보통신부를 중심으로 관련 정책과 규정이 신설되었다. 연구데이터 관리 및 활용에 대한 세부 내용을 연구관리 매뉴얼에 반영하였으며(과학기술정보통신부, 한국과학기술기획평가원, 2020), 「국가연구개발혁신법」의 하위 행정규칙인 「국가연구개발정보처리기준」에서 연구데이터 및 DMP를 정의하고 DMP 제출 사항 등을 규정하였다. 또 연구데이터의 체계적인 관리와 활용을 위한 근거법 제정을 위해 「국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안」을 입법 예고하였다. 해당 법에는 연구데이터의 공개와 이용, 공동 보유, 국제적 연계 등에 관한 사항이 포함되어 있다(과학기술정보통신부, 2023).

이렇듯 연구데이터의 관리와 공유, 활용 등이 점차 중요해지는 환경 속에서 데이터 큐레이션의 개념은 좀 더 주목을 받을 필요가 있다. 데이터 큐레이션은 데이터의 발견 및 검색을 가능하게 하고, 품질을 유지하며, 가치를 추가하고, 인증, 보관, 관리, 보존 및 표현을 포함한 활동을 통해 시간이 지남에 따라 재사용을 제공하는 활동(Johnston, 2017)으로서 모든 유형의 디지털 자료를 대상으로 하는 디지털 큐레이션의 하위 범주로 여겨진다(김판준, 2015; 진보라, 윤유라, 2017; Johnston, 2017).

디지털 큐레이션에 관한 연구 동향은 국내외 연구를 대상으로 몇 차례 분석이 수행되었다(김판준, 2015; 박민석, 이지수, 2024; Kim, 2014). 또한 유사 개념인 연구데이터(한상우, 2023)와 연구데이터 관리(Lee et al., 2024), 오픈데이터(이혜경, 이용구, 2023)의 연구 동향 분석도 있으나 디지털 큐레이션의 하위범주로서 연구지원에 중점을 둔 개념인 데이터 큐레이션의 연구 동향 분석은 아직까지 수행되지 않은 것으로 조사되었다.

국내 데이터 큐레이션 연구는 주로 데이터 큐레이션 수행을 위한 개념모델, 프레임워크, 가이드라인 제시에 중점을 두고 주로 문헌정보학 분야에서 수행되었고, 연구의 수나 다양성 면에서 아직까지 많지 않은 편이었다. 따라서 본 연구는 국외 데이터 큐레이션 연구 동향을 분석하여 연구자들이 향후 참고할 수 있는 시사점을 주는 것을 목적으로 한다.

연구 대상으로는 심사를 통해 신뢰할 수 있는 저널과 논문을 포괄적으로 수집하는 Scopus와 WoS의 학술 정보 1,849건이다. 중복 및 관련 없는 논문 제거 등을 통해 총 1,797건의 학술논문, 학술대회 발표자료 등을 선정했으며 이 자료들의 표제, 키워드, 초록의 텍스트가 최종 연구 대상이 되었다.

연구 방법으로는 대량의 텍스트에서 의미 있는 정보를 추출하는 텍스트 마이닝을 채택하였다. 텍스트 전처리 후 주제를 추출할 수 있는 방법인 토픽 모델링을 통해 국외 데이터 큐레이션 연구의 토픽 5개를 도출하였으며,

해당 토픽의 키워드를 바탕으로 네트워크 분석을 수행하였다. 키워드 네트워크에서 키워드 간의 관계를 통해 중심성을 계산했으며, 이를 통해 중요한 키워드를 선정하였다. 이러한 분석 결과를 통해 데이터 큐레이션의 개념적인 맥락을 확인할 수 있다. 또 데이터 큐레이션 하위 범주 내 연구가 활발한 분야와 그렇지 않은 분야를 확인하는 등 향후 연구자들의 연구 방향 설정에 도움이 되는 시사점을 얻을 수 있다.

2. 관련 연구

2.1 데이터 큐레이션의 중요성과 정의

데이터 관리는 지식 발견과 혁신으로 이어지는 주요 통로이며, 데이터 발행 이후 지식 통합과 재사용을 이끌어내기 위한 핵심이다. 2014년 네덜란드 라이덴에서 열린 '데이터 페어포트 공동 설계' 워크숍에서는 데이터 발견과 재사용 장벽을 극복하고자 다양한 학계와 이해관계자들이 모였다. 여기서 데이터는 찾을 수 있어야 하며(Findability), 접근 가능해야 하며(Accessibility), 상호 운용 가능하며(Interoperability), 재사용이 가능해야 한다(Reusability)는 FAIR 가이드 원칙이 발표되었다(Wilkins et al., 2016).

그러나 FAIR 원칙은 인간 연구 참여자와의 상호작용을 통해 생성된 데이터를 사용할 때 발생하는 사전 동의, 프라이버시 동의, 지적재산권 등과 같은 인식론적, 법적, 윤리적 문제를 다루지 않는다. 이러한 문제를 해결하기 위해 데이터 관리 계획 수립, 데이터 공유를 용이하게 하기 위한 연구 설계, 맥락 정보를 포착하기 위한 메타데이터 작성 등이 포함되는 데이터 큐레이션이 중요하다(Mannheimer, 2024, 5-6).

최근에는 이러한 데이터 큐레이션이 재사용에 미치는 영향과 연구자들이 경험한 데이터 큐레이션의 가치를 규명하기 위한 실증적인 연구들이 수행되고 있다.

Hemphill et al.(2022)은 데이터 큐레이션과 재사용 간의 관계를 규명하기 위해 ICPSR에서 Level 1부터 Level 3까지 데이터 큐레이션 수준이 표기된 데이터셋과 그 데이터셋의 다운로드 간의 상관관계를 규명하였다. Level 1은 각 변수를 설명하는 코드북 등의 작업을 의미하며, Level 2는 Level 1에 더해 데이터 재형식화, 누락된 값 표준화, 철자 수정 등이, 가장 높은 Level 3은 맞춤형 문서화와 텍스트 인덱싱 등이 포함된다. 분석 결과, 높은 수준의 큐레이션이 다운로드 수와 강하게 연결되어 있음을 확인하였다.

Marsolek et al.(2023)은 미국 6개 데이터 저장소에서 데이터 큐레이션 서비스를 받은 연구자 238명을 대상으로 데이터 큐레이션에 대한 인식을 조사하였다. 98%의 연구자가 큐레이션 서비스에 강하게 만족했으며, 90%의 연구자는 큐레이션 과정을 거치고 데이터 공유에 자신감을 느꼈다고 밝혔다. 그 밖에 질적 의견으로 데이터 큐레이션 덕분에 자신의 데이터가 이해하기 쉬워졌다는 의견이 다수 있었다.

데이터 큐레이션에 대한 정의는 여러 연구자마다 조금씩 다르지만(〈표 1〉 참조), 공통적으로는 데이터를 관리하여 재사용이 가능하도록 하게 하는 활동으로 정리할 수 있다.

〈표 1〉 데이터 큐레이션의 정의

연구자	정의
최동훈 외 (2017)	데이터의 발견, 접근, 이해, 통합, 재사용을 위해 메타데이터를 정의 및 생성하고 데이터 부가가치 제고를 위한 노력. 메타데이터와 문맥정보가 데이터 재사용을 촉진함. 연구데이터 관점에서 정의
진보라와 윤유라(2017)	학술 및 연구활동을 통해 생산된 데이터를 생애주기에 따라 지속적으로 관리하는 활동을 지칭. 다양한 학문분야에서 중요한 요소인 데이터를 생애주기에 맞춰 저장하고 관리하며 보존하여 재사용을 도모하는 활동으로 정의함. 디지털 큐레이션의 하위범주
이상헌(2020)	기록관리 현장에서 데이터 기록을 수집부터 서비스까지 관리하고, 특히 데이터 기록의 융합을 통하여 새로운 평가가치를 발견하고 창출하는 모든 활동
이정미(2020)	디지털 정보자원의 수집과 보존을 전제로 데이터의 생애주기에 따라 발견, 정리, 지속적 분석, 재평가를 통해 재사용이 가능하게 하는 활동
이유경과 정은경(2015)	데이터를 체계적으로 보존해 재사용을 촉진하고 그것에 가치를 부여하는 활동
Marsolek et al. (2023)	연구데이터가 목적에 맞고 발견 및 재사용이 가능하며, FAIR 원칙을 더 잘 충족할 수 있도록 하는 다양한 작업을 의미
Johnston et al. (2024)	데이터가 목적에 맞고 발견과 재사용이 가능하도록 보장하기 위해 취해지는 다양한 조치
Johnston (2017)	데이터 발견 및 검색을 가능하게 하고, 데이터 품질을 유지하고, 가치를 추가하며, 인증, 보관, 관리, 보존 및 표현을 포함한 활동을 통해 시간이 지남에 따라 재사용을 제공하는 활동. 데이터 큐레이션의 목표는 연구 결과를 원래 목적 이상으로 유용하게 만들고, 완전성을 보장하며, 장기적인 이용 가능성을 촉진하는 방식으로 연구 결과를 준비하는 것

2.2 데이터 큐레이션 관련 국내 연구

데이터 큐레이션과 관련된 국내 선행연구의 주제는 데이터 큐레이션의 모델, 프레임워크, 가이드라인에 관한 연구와 데이터 큐레이션을 위한 사서 교육 프로그램 및 직무 분석 연구, 그리고 문헌정보학 이외 분야에서의 데이터 큐레이션 서비스 방안 연구 크게 세 가지로 나눠 볼 수 있다.

첫 번째 데이터 큐레이션의 모델, 프레임워크, 가이드라인에 관한 연구에서 최동훈 외(2017)는 커뮤니티 협업 환경에서의 데이터 큐레이션 모델을 제시하였다. 과학 데이터가 재사용되기 위해서는 충분한 맥락정보와 메타데이터가 필요한데, 이는 해당 분야 전문가와 데이터 재사용자, 데이터 사서 등으로 구성된 커뮤니티가 함께 생성해야 한다. 이 모델에서는 데이터와 문헌 각각의 메타데이터와 문맥정보가 DOI로 상호 연결된다. 진보라와 윤유라(2017)는 데이터 수명주기에 따른 데이터 큐레이션 가이드라인 생성을 위해 문헌연구와 사례분석을 수행하였다. 한나은(2023)은 기존에 제안된 5가지 디지털 또는 데이터 큐레이션 모델에서 공통되는 요인을 추출하여 새로운 연구데이터 큐레이션 모델을 수립하였다. 여기에는 데이터 수명주기 10단계와 각 단계에서 발생할 수 있는 잠재적인 문제가 포함된다. 이정미(2020)는 대학도서관의 데이터 큐레이션 프레임워크와 이를 서비스 할 때의 고려사항을 제시하였다.

두 번째 사서 교육 프로그램 및 직무 분석 연구에서 김진희 외(2019)는 사서들의 연구지원을 위한 데이터 큐레이션 교육 프로그램을 개발

하기 위해 문헌연구, 사서 및 전문가 요구분석, 초점집단 면접을 수행하였다. 교육 내용으로는 데이터 큐레이션의 필요성, 데이터의 개념과 유형, 데이터 개방성, 데이터 관리 계획, 리포지터리를 활용한 데이터 큐레이션 등 9개 핵심주제가 포함되었다. 이유경과 정은경(2015)은 국외 데이터 큐레이터 구인 공고와 국내 데이터 관리 실무자 5명과의 심층면담을 통해 데이터 큐레이터의 핵심 직무를 규명하였다. 4가지 범주로 정리된 직무에는 커뮤니케이션 능력, IT 지식을 기반으로 한 데이터 관리 시스템 구축 및 운영 능력, 이용자 교육 및 관련 서비스 도구 제공 능력 등이 포함된다.

세 번째 문헌정보학 이외에서 데이터 큐레이션 서비스 방안에 관한 연구로 이상헌(2020)은 기록관리 측면에서 공공데이터 기록 활용의 활성화를 위해 공공데이터포털과 혜안시스템의 사례를 분석하였고, 이제욱(2022)은 스포츠 산업 발전을 위해 스포츠 공공데이터 큐레이션 서비스를 제안하며, 이를 위한 법적, 정책적 개선방안을 제시하였다. 이현조 외(2022)는 생산, 유통, 소비 등 농업 전 과정에서 데이터를 활용하는 '디지털 농업'을 위해 농업 데이터의 수집, 전처리, 저장, 분석을 수행하는 데이터 큐레이션 서비스 방안을 제안하였다.

이상을 정리하면 데이터 큐레이션 관련 국내 연구는 주로 문헌정보학 분야에서 수행되었다. 문헌정보학 외에서는 기록관리, 스포츠, 농업 등 분야에서 연구가 수행되었으나 아직은 연구의 수나 다양성 면에서 미미한 편이었다. 본 연구에서는 Scopus와 WoS에 수록된 1,797건의 데이터 큐레이션 관련 연구를 분석하여 국외 연구의 주요 주제와 연구 트렌드 등을 살펴보는 것을 목표로 한다.

2.3 데이터 연구 동향 분석

디지털화와 데이터의 폭발, 데이터 기반 의사결정, 데이터의 경제적 가치 등 사회 전반적으로 데이터의 중요성이 커짐에 따라 데이터 관련 연구도 증가하고 있다. 이러한 데이터에 관한 연구 동향 연구가 국내외 여러 차례 수행되었다.

먼저 데이터 큐레이션의 상위 범주인 디지털 큐레이션과 관련하여 국내외를 대상으로 연구동향을 분석한 연구가 있다. 김판준(2015)은 국외 디지털 큐레이션 연구동향을 위해 LISTA에서 추출한 374편의 논문을 대상으로 디스크립터 프로파일링 기법을 활용하여 지적구조를 규명하고 각 디스크립터의 생산성과 성장성을 분석, 이를 바탕으로 향후 문헌정보학에서 다루어야 할 연구과제를 제시하였다. 연구 결과, 가장 많이 생산된 주제는 디지털도서관, 디지털 보존, 정보자원관리 3개 영역이며, 성장성이 높은 주제는 데이터베이스 관리, 관리, 큐레이터직, 교육과정 4개로 나타났다. Kim(2014)은 국제 디지털 큐레이션 저널(International Journal of Digital Curation)에 게재된 219편의 논문을 대상으로 연도별 생산성과 초록의 텍스트를 분석하고 공저, 인용, 연구자금 현황을 조사하였다. 분석 결과, 가장 자주 등장한 10개 구문은 디지털 보존, 연구데이터, 데이터 관리 등이었으며, 75%의 연구가 2인 이상 저자의 연구였고, 86% 이상이 한 번 이상 인용되었으며 40%는 연방 정부의 보조금을 지원받았다. 박민석과 이지수(2024)는 국내 디지털 큐레이션 연구

동향 분석을 위해 39건의 논문을 대상으로 키워드 네트워크 분석 등을 수행하였다. 분석결과, 연결중심성이 가장 높은 디지털 큐레이션을 중심으로 기관, 콘텐츠, 도서관, 큐레이션, 보존, 지식 등이 가장 영향력이 높게 나타났다. 응집 그룹 분석에서는 디지털 큐레이션 콘텐츠 및 정보서비스 그룹과 기관별 큐레이션 정책 및 모델 그룹, 인문학 빅데이터 해석 그룹, 연구데이터 큐레이션 모델 그룹, 디지털큐레이션 역량 그룹 등으로 군집이 형성되었다.

한편 데이터 큐레이션의 유사 개념인 연구데이터, 연구데이터 관리 연구(RDM)와 관련해서도 연구 동향 분석이 이루어졌다. Lee et al.(2024)은 RDM에 관한 연구가 전 세계적으로 어떻게 발전하고 있으며, 국가별로 어떻게 다른지 조사하기 위해 WoS에서 추출한 403개의 논문을 대상으로 계량서지학적 접근 방식을 통해 연구동향을 분석하였다. 연구결과, 연도별로 2000년대 초반 북미에서부터 연구가 시작되고, 이후 유럽, 아시아 등으로 확산되는 양상이 나타났다. 국가별로 독일, 미국, 영국이 가장 많은 발행국이었으며 세 국가의 공통 주제는 리포지터리, 데이터 관리, 데이터 공유, 오픈사이언스로 나타났다. 403편 중 15회 이상 인용된 29편 논문을 대상으로 동시인용분석을 실시했고, 동시인용 네트워크를 생성하였다. 5개의 클러스터는 RDM의 기본, RDM 서비스, 연구데이터 공유의 원칙과 가이드라인, 과학자들의 데이터 공유, 연구데이터의 적절한 관리와 공유로 나타났다. 한상우(2023)는 연구데이터 관련 국내 연구 동향 분석을 위해 RISS에서 수집한 58건의 논문을 대상으로 키워드 네트워크를 분석하였다. 키워드 빈도수, 연결정도 중심성, 매개중심성에서 연구데이터 공유, 연구데이터 관리, 데이터 리포지터리 등이 높게 나타났고, 응집 구조 또한 해당 키워드들 위주로 형성되었다.

그 밖에도 오픈데이터, 공공데이터, 데이터 거버넌스, 데이터 정책, 빅데이터, 공간 빅데이터 등 다양한 개념을 대상으로 연구 동향 분석이 이루어졌다. 이혜경과 이용구(2023)는 오픈데이터 및 링크드 오픈데이터의 연구 동향과 지적 구조를 파악하기 위해 Scopus에서 저자 키워드로 검색한 6,543건의 논문을 대상으로 키워드 네트워크 분석을 하였다. 분석 결과, 오픈데이터와 관련해서는 열린정부, 공공데이터 관련 키워드가 상위에 위치했고, 링크드 오픈데이터와 관련해서는 기계적인 컴퓨터 관련 연구 비율이 높았다. 박대영 외(2021)는 토픽 모델링을 활용하여 공공데이터와 관련한 국내와 국외의 연구동향을 비교, 분석하였다. 이를 위해 국내는 KCI, 국외는 Springer DB를 통해 논문을 수집하고, 키워드 빈도수, LDA 토픽모델링, 토픽 시계열 분석 등의 텍스트 기법을 활용하여 분석하였다. 연구 결과 국내는 공공분야 정책 연구가 주를 이룬 반면 국외는 의료, 건강 관련 연구가 다수 수행된 것으로 나타났다. 정선경(2022)은 국내 데이터 거버넌스 연구동향 분석을 위해 KCI 논문을 대상으로 빈도분석, 키워드 네트워크, LDA 토픽모델링 등을 수행하였다. 연구결과, 최빈 키워드는 정보, 빅데이터, 관리, 정책, 정부 등이며 네트워크에서는 데이터 산업정책, 데이터 커버넌스 성과 등이 중심적인 역할을 하고, 토픽모델링을 통해 정책, 플랫폼, 법률, 구현 4개의 토픽을 도출하였다. 유화선과 정도범(2023)은 국외 데이터 정책 연구동향 분석을 위해 Web of Science에

서 수집한 논문을 대상으로 LDA 토픽모델링 및 의미연결망 분석을 수행하였다. 연구결과, 개인정보보호, 정부역할, 데이터 분석 방법론, 데이터 활용, 데이터 수집, 데이터 영향력 6개의 토픽이 도출되었다. 이원상과 손소영(2015)은 공간 빅데이터 연구 현황을 파악하기 위해 Scopus 등재 논문 1,621개를 대상으로 LDA 토픽모델링과 키워드 네트워크 분석을 수행하였다. 총 40개의 토픽을 19개로 분류였는데, 생태학 관련 주제가 17.5%로 가장 많이 발생했고, 기상학, 도시환경에서의 재해, 삼림관리 등이 뒤를 이었다. 연도별로 공간 빅데이터를 위한 IT 연구, 지리통계학 분야 등은 최근 연구가 급등한 반면, 삼림 관련 위성 이미지 분석 등은 빈도가 다소 감소했다. 이들 토픽으로 네트워크 분석을 한 결과, 토픽 간 연결은 많지 않았으

〈표 2〉 데이터 관련 연구 동향 분석 선행 연구

연구자	연구주제	국내/국외 (DB 또는 저널명)	분석대상 연도	논문 건수	분석 방법
김판준(2015)	디지털 큐레이션	국외(LISTA)	1987~2014	374	기초분석, 디스크립터 프로파일링
Kim(2014)	디지털 큐레이션	국외(IJDC)	2006~2013	219	기초분석, 키워드 빈도분석
박민석과 이지수(2024)	디지털 큐레이션	국내(RISS, KCI, DBPia)	2009~2023	39	기초분석, 키워드 빈도분석 공저 및 키워드 네트워크, 응집그룹 분석
Lee et al. (2024)	연구데이터 관리 (RDM)	국외(WoS)	2000~2022	403	성과분석, 키워드 분석, 인용분석, 동시인용분석, 인용문헌분석, 대응분석
한상우(2023)	연구데이터	국내(RISS)	2000~2023	58	기초분석, 키워드 빈도분석, 키워드 네트워크, 응집그룹
이혜경과 이용구(2023)	오픈데이터	국외(Scopus)	1999~2023	5,589	기초분석, 키워드 네트워크, 군집분석
박대영 외(2021)	공공데이터	국내(KCI)/ 국외(Springer)	2007~2020	1,437/ 9,607	기초분석, 키워드 빈도분석, LDA 토픽모델링
정선경(2022)	데이터 거버넌스	국내(KCI)	2009~2021	158	기초분석, 키워드빈도분석, 키워드 네트워크, LDA 토픽모델링
유화선과 정도범(2023)	데이터 정책	국외(WoS)	2011~2022	1,109	기초분석, 키워드 빈도분석, TF-IDF, LDA 토픽모델링, 키워드 네트워크
이원상과 손소영(2015)	공간 빅데이터	국외(Scopus)	~2014	1,621	기초분석, TF-IDF, LDA 토픽모델링, 키워드 네트워크
Mohammadi와 Karami(2022)	빅데이터	국외(WoS)	2012~2017	36,821	기초분석, LDA 토픽모델링, 동시출현 단어분석

나 기상 및 환경, 경작 및 도시 관련, 이미지 처리, 소셜 및 도시환경에서의 서비스 관련 연구 등 크게 4가지 방향으로 주제 연결이 발생하였다. Mohammadi와 Karami(2022)는 학문 분야를 넘나드는 빅데이터 연구 동향을 살피기 위해 WoS의 36,000편 이상의 빅데이터 관련 논문을 토픽 모델링과 동시출현단어 기법을 통해 분석하였다. 그 결과, 빅데이터 연구는 대규모 데이터의 저장, 수집, 분석과 관련한 여러 주제로 나타났으며 주로 컴퓨터 과학 쪽에서 수행되었다. 그러나 빅데이터 기법과 관련하여 교육학, 도시정보학, 경영학 등 여러 학문 분야로 연구가 확장되고 있음이 나타났다.

이상의 데이터 관련 연구 동향 분석 연구를 정리하면 〈표 2〉와 같다. 데이터를 대상으로 한 여러 연구 동향 분석이 수행되었지만, 연구지원을 위해 수집부터 정리, 보존, 재사용까지 데이터를 관리하며 최종적으로 연구 데이터의 재사용을 위해 이뤄지는 활동인 '데이터 큐레이션'과 관련한 연구 동향 분석은 아직까지 수행되지 않은 것으로 조사되었다. 따라서 본 연구에서는 Scopus와 WoS에서 'data curation'으로 검색하여 추출한 1,797건의 학술자료의 제목, 키워드, 초록을 바탕으로 데이터 큐레이션 연구 동향을 분석하고자 한다.

3. 분석 대상 및 방법

3.1 데이터 추출 및 전처리

본 연구의 목적은 데이터 큐레이션의 국외 연구 동향을 분석하는 것이다. 분석 대상으로는 과학, 기술, 의학, 예술, 인문학 등 포괄적인 분야의 저널과 논문 정보를 수록한 데이터베이스인 Scopus와 WoS를 선택하였다. 2024년 7월 5일 Scopus에서 검색 필드를 'Keywords'로 두고 "data curation"을 검색한 결과 1,849건의 학술정보를 추출하였다. 이어 WoS Core Collection에서는 검색 필드를 'Author Keywords'에 두고 마찬가지로 "data curation"으로 검색하여 학술정보 436건을 추출하였다. DOI가 있는 경우 DOI를 기준으로, 없는 경우 표제와 저자명을 기준으로 중복된 논문 403건을 제거한 1,882건의 초록을 읽었다. 저자 키워드에 데이터 큐레이션이 포함되어 있으나 〈표 1〉에서 정의된 데이터 큐레이션과 관련성이 떨어지는 논문 85건이 있었다. 예를 들어 의학 분야에서 기관 절개술을 받은 환자의 탈관(Decannulation) 절차와 관련된 프로토콜의 내용 타당성을 검증하는 연구나, 간호 현장에서 사용되는 용어를 표준화하여 간호사와 의료팀 간의 의사소통을 개선하는 연구 등은 본 연구가 정의하는 데이터 큐레이션과 관련성이 적다고 판단하여 분석 대상에서 제거하였다. 이러한 과정을 거쳐 최종적으로 1,797개의 학술자료가 남게 되었다. 여러 정보 중에서도 'Author Keywords'와 'Title', 'Abstract'를 분석 대상으로 선택하였고 각 행별로 이 세가지 열을 병합하여 분석을 위한 새로운 열을 생성하였다.

전처리를 비롯한 거의 모든 분석은 Google Colab 환경에서 Python 코딩으로 진행되었다. 텍스트 전처리 과정은 크게 정제, 정규화, 품사 태깅 및 선택 3단계로 이루어졌다. 정제 과정에서는 불용어 제거를 위해 NLTK의 Stopwords와 연구자가 작성한 불용어 리스트를 이용하였

다. 불용어 리스트에는 검색어로서 모든 텍스트에 포함되어 해석에 유의미한 도움이 되지 않는 'data', 'curation'과 초록 작성 시 통상적으로 포함되는 단어인 'background', 'methodology', 'conclusion', 'study', 'article' 등을 포함시켰다. 'research'와 'analysis'등 초록의 통상적인 언어로 사용된 경우와 연구 주제로서 사용된 경우가 혼재된 단어를 불용어로 포함시킬지 여부를 판단하기 위해 불용어 제거 전 키워드 빈도수 분석을 실시하였다. 'research'의 경우 유사어인 study(1,123회), paper(665회), article(556회)보다 훨씬 많은 2,235회가 등장하여 단순한 초록 용어보다는 연구 주제로서 사용된 경우가 많다고 판단하여 불용어 리스트에 포함하지 않았다. 추가적인 판단을 위해 저자 키워드만을 대상으로 빈도수 분석을 수행하였다. 'analysis'는 data analysis(30회), sequence analysis(27회), protein analysis(14회) 등 다양한 단어와 함께 연구 주제로 사용되고 있어 단순 초록 용어로만 보기 어렵다고 판단하여 역시 불용어로 포함하지 않았다. 이와 같은 불용어 작업 이후 특수문자, 숫자 등을 제거하였다. 다음으로 정규화 작업에서는 NLTK의 WordNetLemmatizer를 통해 단어를 기본형으로 변환하는 표제어 추출 작업을 하였고 복수형을 모두 단수형으로 변환하였다. 또 'AI'는 'Artificial intelligence'으로, 'NLP'는 'Natural language processing'으로 변환하는 등 동의어 정리 작업을 수행하였다.

마지막으로 품사 태깅 및 선택을 위해 NLTK를 활용해 품사를 구별하였다. 전처리된 단어를 살펴보았을 때 분석에 도움이 될 것으로 판단되는 명사와 형용사만을 최종 분석 대상으로 선택하였다.

3.2 분석 방법

본 연구는 국외 데이터 큐레이션 연구 동향을 분석하기 위해 Scopus와 WoS로부터 추출한 1,797건의 학술 정보를 분석 대상으로 하였다. 토픽 모델링 중에서도 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)은 문서가 잠재 주제들의 랜덤 혼합으로 표현되며, 각 주제는 특정 단어들에 대한 확률 분포로 특징지어진다는 가정을 기반으로 하는 분석 방법이다. 문서의 주제 분포를 결정하는 매개변수 α는 디리클레 분포를 따르며, 각 주제에 속한 단어 분포는 다항 분포 β에 의해 결정된다. LDA는 기존 토픽 모델링 방법인 pLSI, Unigram Mixture 모델에서 제기된 오버피팅 문제를 효과적으로 해결하며, 문서 모델링, 문서 분류, 협업 필터링 등 다양한 응용 분야에서 우수한 성능을 보인다(Blei et al., 2003). 데이터 연구 동향 연구에서도 국내외 공공데이터(박대영 외, 2021), 국내 데이터 거버넌스(정선경, 2022), 국외 데이터 정책(유화선, 정도범, 2023), 국외 공간 빅데이터(이원상, 손소영, 2015), 국외 빅데이터(Mohammadi & Karami, 2022) 등 여러 연구에서 LDA를 채택하고 있다. 따라서 본 연구에서도 국외 데이터 큐레이션 연구 동향 분석 방법으로 LDA 토픽 모델링을 채택하였다.

다음 LDA 토픽 모델링으로 도출한 토픽의 키워드를 대상으로 네트워크 분석을 진행하였다. 키워드 네트워크의 중심성 지수를 통해 특정 연구 분야에서 중요한 키워드를 확인할 수 있으며, 키워드 간의 관계를 분석함으로써 텍스트에 내재된 의미를 파악할 수 있다. LDA와 마찬가지로 데이터 관련 연구동향 분석에

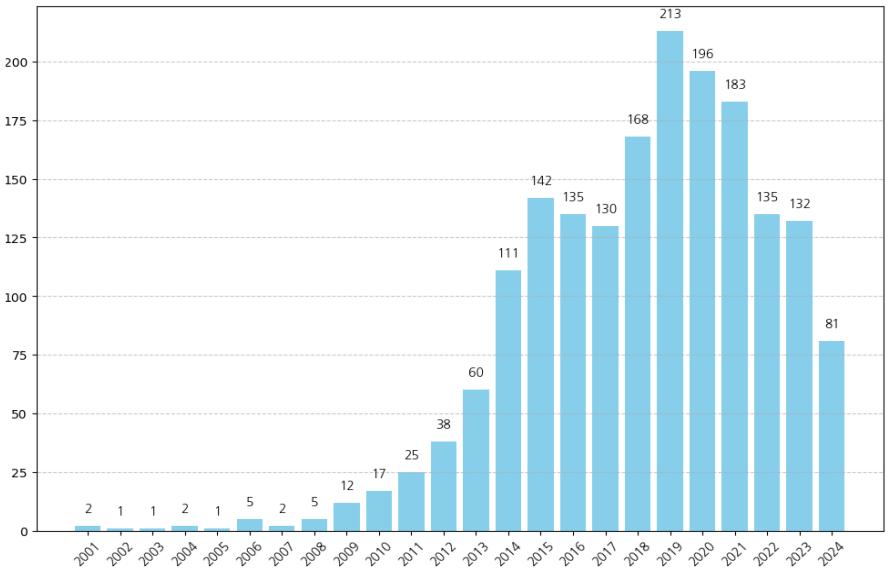
서 키워드 네트워크 분석이 다수 활용되었다(박민석, 이지수, 2024; 유화선, 정도범, 2023; 이원상, 손소영, 2015; 이혜경, 이용구, 2023; 정선경, 2022; 한상우, 2023; Mohammadi & Karami, 2022).

키워드 네트워크는 각 키워드 쌍의 동시출현 빈도를 계산하여 이 빈도로부터 키워드 간의 유사도를 계산하여 구할 수 있다(이수상, 2012, 99). 본 연구에서는 가중 데이터값의 크기를 잘 반영할 수 있는 피어슨 상관계수(이수상, 2012, 162)를 통해 유사도를 계산하였다. 이후 가중 네트워크를 대상으로 개발된 중심성 척도인 이재윤(2006; 2013)의 최근접이웃중심성(NNC), 평균연관성(AVGSIM), 삼각매개중심성(TBC)을 적용하였다. 최근접이웃중심성(NNC)은 다른 노드에 의해 최근접 이웃으로 꼽히는 정도를 나타내는 지역 중심성 중의 하나다. 평균연관성(AVGSIM)은 해당 노드가 다른 노드들과 평균적으로 얼마나 가까운가를 나타내는 척도로 전역 중심성을 나타내는 척도가 된다. 삼각매개중심성(TBC)은 대상 키워드가 다른 키워드 사이를 결속시켜주는 정도를 측정하는 척도로 역시 전역 중심성 중에 하나다(이재윤, 2006; 2013). 중심성 계산을 위해 피어슨 상관계수로 산출된 각 노드의 유사도를 정방행렬로 변환한 후, 이재윤의 WNET(이재윤, [n.d.])을 활용하여 중심성을 도출하였다.

4. 분석 결과

4.1 기초분석

텍스트 분석에 앞서, 분석 대상 논문의 발행연도와 문서유형 그리고 발행 학술지명을 조사하였다. 먼저 연도별 출판건수는 〈그림 1〉과 같



〈그림 1〉 데이터 큐레이션 분야 연도별 출판건수

다. 2001년 데이터 큐레이션과 관련한 논문이 처음 출간된 이후, 2009년부터 매년 증가하여 2015년 142건이 출판되었다. 이후 증감을 반복하다 2019년 가장 많은 213건이 출간된 이후 최근에는 연 100건대 정도의 연구가 꾸준히 수행되고 있다.

다음으로 Scopus와 WoS의 데이터 큐레이션 관련 학술자료의 유형은 〈표 3〉과 같다. 논문(Article)이 1,080으로 가장 많고 이어 Conference paper(488개), Review(94개), Book chapter(54개), Note(22개), Proceedings paper(21개), Editorial(16개) 등 순이다.

이어 ISSN을 기준으로 산출한 발행 상위 학술지 목록은 〈표 4〉와 같다. 가장 많은 학술지는

〈표 3〉 유형별 학술자료 개수

유형	개수	유형	개수
Article	1,080	Editorial	16
Conference paper	488	Data paper	11
Review	94	Letter	5
Book chapter	54	Short survey	3
Note	22	Meeting abstract	3
Proceedings paper	21		

〈표 4〉 발행 상위 학술지 목록

순위	학술지명	주요 연구분야	개수
1	Database: the journal of biological databases and curation	Biology, Biocuration, Database	130
2	Nucleic Acids Research	Chemistry, Computational biology	86
3	Studies in Health Technology and Informatics	Health informatics	55
4	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	Computer Science	39
5	PLoS ONE	Biology, Medicine, Physics	37
6	CEUR Workshop Proceedings	Computer Science	36
7	Journal of Biomedical Informatics	Biomedical informatics	30
8	Proceedings of the Association for Information Science and Technology	Information Science	27
9	Proceedings of the ACM/IEEE Joint Conference on Digital Libraries	Digital libraries, Information Science	23
9	Scientific Data	Research data	23
11	Proceedings of the ASIST Annual Meeting	Digital libraries, Information Science	21
11	Scientific Reports	Natural sciences, Medicine, Engineering	21
13	Methods in Molecular Biology	Molecular Biology	20
14	Data Science Journal	Data science, Data analytics	16
14	WORLDWIDE COMMONALITIES AND CHALLENGES IN INFORMATION LITERACY RESEARCH AND PRACTICE	Information literacy	16
16	Bioinformatics	Bioinformatics, Computational Biology	15
17	PLoS Computational Biology	Computational Biology	11

Database: the journal of biological database and curation(130개)으로 생물학, 생물정보학 등을 연구하는 학술지다. 이어 2위인 Nucleic Acids Research(86개)는 DNA와 RNA 등 핵산 연구를 주로 다루는 학술지로 화학, 계산생물학 등을 주 연구분야로 하고 있다. 3위인 Studies in Health Technology and Informatics(55개)는 의료정보학, 4위인 Lecture Notes in Computer Science(39개)는 컴퓨터과학, PLoS ONE은 과학 일반, CEUR Workshop Proceedings은 컴퓨터과학, Journal of Biomedical Informatics는 의료정보학 등으로 국외 데이터 큐레이션 연구는 과학 분야, 그중에서도 생물정보학, 의료정보학, 생명과학, 컴퓨터과학 분야 등에서 활성화 되어 있음을 확인할 수 있다. 정보학 및 문헌정보학 분야의 학술지는 8위의 Proceedings of the Association for Information Science and Technology(27개), 9위의 Proceedings of the ACM/IEEE Joint Conference on Digital Libraries(23개), 11위의 Proceedings of the ASIST Annual Meeting(21개) 등으로 포함되어 있다.

4.2 키워드 빈도분석

전처리된 키워드의 60위까지의 빈도는 〈표 5〉와 같다. 최빈도 단어는 'research(2,235회)'로 나타났다. research가 포함된 실제 연구를 살펴본 결과, 'medical research', 'biomedical research', 'clinical research' 등 의학, 생의학, 임상의학 연구와 'research data management', 'research data', 'research infrastructure' 등 연구지원의 다양한 측면에서 사용되고 있었다. 다음으로 자주 등장하는 단어인 'information'은 데이터 큐레이션 연구가 정보 처리, 정보 시스템 등의 측면에서 다뤄지고 있음을 보여준다.

〈표 5〉 상위 단어 빈도

순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도	순위	키워드	빈도
1	research	2,235	16	science	687	31	health	527	46	collection	444
2	information	2,030	17	user	656	32	community	520	47	framework	442
3	database	1,464	18	new	635	33	disease	517	48	access	432
4	model	1,451	19	high	617	34	scientific	499	49	set	429
5	annotation	1,284	20	clinical	597	35	practice	498	50	task	424
6	process	1,269	21	digital	589	36	open	497	51	genome	417
7	system	1,186	22	resource	588	37	repository	492	52	text	407
8	analysis	1,180	23	protein	572	38	medical	491	53	work	398
9	gene	1,152	24	use	565	39	metadata	483	54	standard	395
10	human	1,095	25	large	554	40	researcher	479	55	application	394
11	quality	970	26	source	547	41	library	473	56	biological	389
12	management	931	26	network	547	42	challenge	460	56	project	389
13	tool	801	28	service	540	43	sequence	457	58	term	387
14	image	767	29	different	538	44	time	455	59	ontology	382
15	genetic	709	30	available	532	45	software	450	60	structure	381

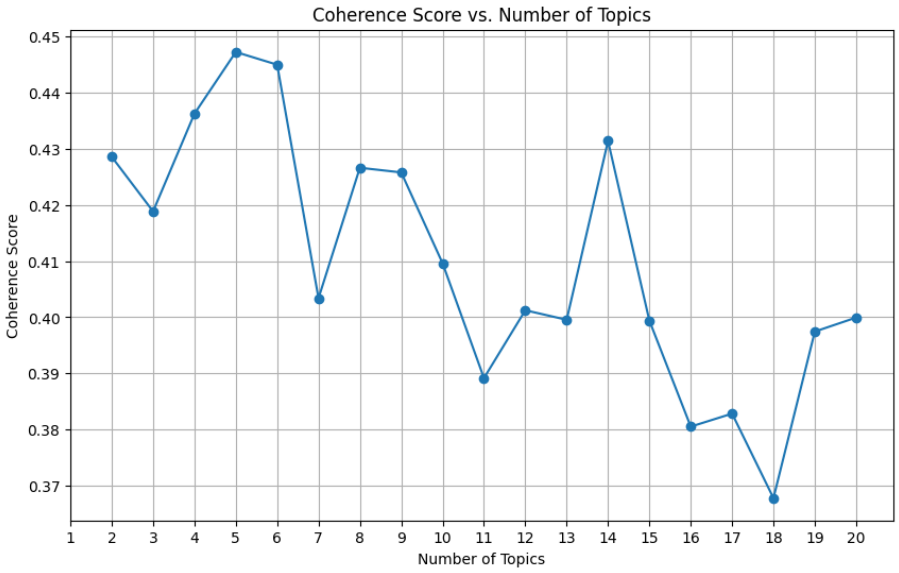
이어 'database'는 큐레이션된 데이터를 DB로 구축하거나 DB의 관리와 관련된 연구가 많음을 보여주고, 'model'은 데이터 큐레이션 관련 모델 연구를, 'annotation'을 통해서는 데이터 주석과 관련한 연구가 많음을 알 수 있다. 'system', 'analysis', 'quality', 'management'는 데이터의 품질 관리를 위한 시스템이나 데이터 분석과 관련한 연구를, 'gene', 'human', 'genetic', 'clinical' 등의 단어는 주로 생물학, 생의학, 의학 연구 시 발생하는 데이터와 관련하여 큐레이션 연구가 이뤄지고 있음을 알 수 있다.

4.3 LDA 토픽 모델링 분석

본 연구에서는 데이터 큐레이션 관련 연구 동향을 파악하기 위해 텍스트에서 데이터 내 숨겨진 주제를 자동으로 식별할 수 있는 토픽 모델링을 분석 방법으로 채택하였다. 토픽 모델링 중에 널리 쓰이는 방법 중 하나인 LDA 알고리즘을 통해 분석하였다. 분석에 앞서 적절한 토픽 수를 결정하기 위해 토픽 응집도(topic coherence)를 계산하였다. 토픽 응집도는 각 토픽에서 상위 비중을 차지하는 단어들이 의미적으로 유사한지를 나타내는 척도로, 만일 토픽이 단일 주제를 잘 표현한다면 값이 높게 나타난다. 〈그림 2〉는 토픽 수에 따른 응집도를 나타낸 그래프이다. 토픽 개수가 5일 때 응집도가 0.447로 가장 높게 나타났다. 따라서 토픽의 수를 5개로 정하고 LDA 토픽 모델링을 수행하였다.

토픽 분석 결과는 〈표 6〉과 같다. 토픽의 명칭은 도출된 토픽과 키워드를 바탕으로 해당 키워드가 포함된 연구의 제목, 키워드, 초록을 다시 살펴보며 연구자가 명명하였다.

토픽 1에는 'health', 'cell', 'analysis', 'information', 'cancer', 'quality', 'research', 'clinical', 'patient',



〈그림 2〉 토픽 수에 따른 응집도

〈표 6〉 데이터 큐레이션 연구 토픽과 핵심 키워드 및 비중

토픽(토픽명)	핵심 키워드	비중
Topic 1 (임상 의료 데이터의 품질 제고와 분석)	health, cell, analysis, information, cancer, quality, research, clinical, patient, record	9.15%
Topic 2 (빅데이터 관리와 처리 시스템의 효율성 향상)	system, process, analysis, quality, information, user, model, big, source, management	15.09%
Topic 3 (과학 데이터의 관리와 디지털 리포지터리)	research, management, information, science, digital, model, practice, repository, system, researcher	36.24%
Topic 4 (의료 및 생물학적 데이터의 주석과 모델링)	annotation, image, model, text, information, clinical, process, human, network, medical	18.9%
Topic 5 (유전자 및 단백질 데이터베이스 연구)	database, gene, information, genetic, human, protein, annotation, process, analysis, disease	19.85%

'record' 등의 키워드가 포함되어 있다. 이 토픽은 환자의 종양 치료와 같은 임상 연구에서 건강 기록 및 세포 관련 데이터를 다루며, 이러한 데이터의 품질 향상에 관한 연구를 포함하고 있음을 시사한다. 따라서 토픽 1을 '임상 의료 데이터의 품질 제고'로 명명하였다.

토픽 2는 'system', 'process', 'analysis', 'quality', 'information', 'user', 'model', 'big', 'source', 'management' 등의 키워드로 구성되어 있다. 이 토픽은 빅데이터 환경에서 데이터 관리와 처리, 분석 시스템의 품질 관리에 관한 다양한 연구를 다루고 있다. 특히, 데이터의 수집, 정제, 관리 과정에서의 효율성 향상과 이용자의 요구에 맞춘 데이터 관리 전략이 주요 주제로 다뤄진다. 따라서 이 토픽을 "빅데이터 관리와 처리 시스템의 효율성 향상"으로 명명하였다.

토픽 3은 'research', 'management', 'information', 'science', 'digital', 'model', 'practice', 'repository', 'system', 'researcher' 등의 키워드로 구성되어 있다. 이 토픽은 과학적 연구와 데이터 관리에서의 디지털 시스템과 모델의 활용, 연구자의 데이터 관리를 위한 리포지토리 구축과 관련된 연구를 다루고 있다. 따라서 이 토픽을 "과학 데이터 관리와 디지털 리포지토리"로 명명하였다.

토픽 4는 'annotation', 'image', 'model', 'text', 'information', 'clinical', 'process', 'human', 'network', 'medical' 등의 키워드로 구성되어 있다. 이 토픽은 의료 및 생물학적 데이터를 포함한 다양한 도메인에서의 이미지와 텍스트 주석 작업, 데이터 처리 및 모델링 방법론을 다룬다. 따라서 이 토픽을 "의료 및 생물학적 데이터의 주석과 모델링"으로 명명하였다.

토픽 5에는 'database', 'gene', 'information', 'genetic', 'human', 'protein', 'annotation', 'process', 'analysis', 'disease' 등이 포함되어 있다. 이 토픽은 유전자 및 단백질 데이터베이스와 관련된 연구를 다루며, 유전자 및 단백질의 상호작용, 생물학적 데이터의 분석 및 주석 작업 등을 다룬다. 이에 이 토픽을 "유전자 및 단백질 데이터베이스 연구"로 명명하였다.

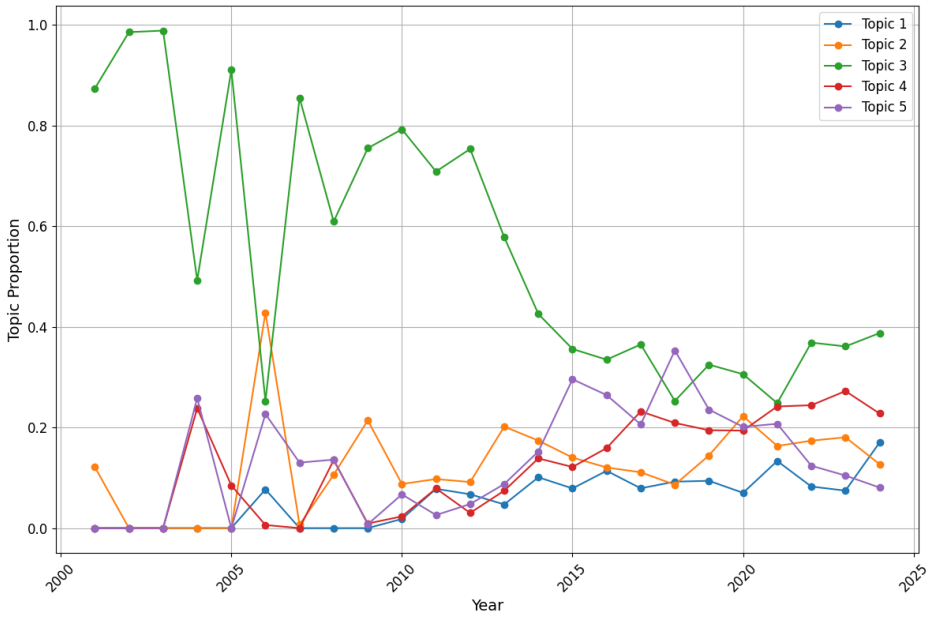
도출된 토픽들이 전체 데이터 큐레이션 연구에서 얼마만큼의 비중을 갖고 있는지 살펴보기 위해 토픽별 비중을 분석하였다. 분석 결과, 토픽 3(과학 데이터의 관리와 디지털 리포지터

리)가 36.24%로 가장 높은 비중을 차지했으며 다음으로 토픽 5(유전자 및 단백질 데이터베이스 연구) 19.85%, 토픽 4(의료 및 생물학적 데이터의 주석과 모델링) 18.9%, 토픽 2(빅데이터 관리와 처리 시스템의 효율성 향상) 15.09%, 토픽 1(임상 의료 데이터의 품질 제고와 분석) 9.15% 순이었다.

〈그림 3〉은 시간의 흐름에 따른 각 토픽의 비중 변화를 나타낸다. 토픽 3(과학 데이터의 관리와 디지털 리포지터리)는 데이터 큐레이션 연구 초기인 2000년대 초반부터 2015년까지 다른 토픽에 비해 월등히 높은 비중을 차지했으나, 2015년 이후에는 타 토픽과 비슷한 수준으로 감소하였다. 그러나 최근 들어 해당 토픽에 대한 연구가 다시 활성화되는 추세를 보인다. 토픽 1(임상 의료 데이터의 품질 제고와 분석)은 전체 토픽 중에서는 비중이 낮지만, 최근 성장하고 있는 연구 주제임을 확인할 수 있다. 토픽 2(빅데이터 관리와 처리 시스템의 효율성 향상)는 2006년에 모든 토픽 중 가장 높은 비중을 보이며 일시적으로 연구가 활발하였으나, 최근에는 감소 추세를 보이고 있다. 토픽 4(의료 및 생물학적 데이터의 주석과 모델링)는 2017년부터 일정한 연구량을 유지하고 있는 분야로 판단된다. 마지막으로, 토픽 5(유전자 및 단백질 데이터베이스 연구)는 2018년에 모든 토픽 중 가장 높은 비중을 차지하였으나, 그 이후로는 연구가 감소하고 있는 주제이다.

4.4 키워드 네트워크 분석

본 연구에서는 LDA 토픽 분석을 통해 5개의 토픽이 도출되었으며, 각 토픽별로 발생 확률이 높은 10개의 단어, 총 50개의 단어를 확인



〈그림 3〉 토픽 트렌드 분석

할 수 있었다(〈표 6〉 참조). 이 중 중복되는 단어 17개를 제외한 총 33개의 단어를 대상으로 키워드 네트워크 분석을 실시하여 삼각매개중심성(TBC), 평균연관성(AVGSIM), 최근접 이웃중심성(NNC)을 산출한 결과는 〈표 7〉과 같다. 전역 중심성인 삼각매개중심성(TBC)과 평균연관성(AVGSIM)에서 'analysis'가 가장 높은 지수를 나타내었다. 실제 논문을 확인해 보면, 'analysis'는 큐레이션된 데이터를 활용하는 측면에서 분석 방법이나 분석 시스템 등으로 폭

〈표 7〉 데이터 큐레이션 토픽 키워드의 중심성 분석 결과

키워드	rTBC	키워드	AVGSIM	키워드	rNNC
analysis	0.65927	analysis	0.49054	research	0.15625
system	0.64516	research	0.48764	gene	0.09375
information	0.61895	information	0.482	system	0.09375
research	0.59073	system	0.47406	health	0.0625
human	0.54435	human	0.44711	management	0.0625
source	0.52823	quality	0.4387	medical	0.0625
quality	0.51008	health	0.43846	database	0.0625
process	0.45766	source	0.43815	user	0.0625
genetic	0.40927	management	0.42617	cancer	0.0625
management	0.39919	medical	0.42357	information	0.03125
network	0.39113	genetic	0.4229	source	0.03125
cell	0.3871	process	0.41834	genetic	0.03125
database	0.38105	database	0.4117	science	0.03125
health	0.37702	patient	0.40494	network	0.03125
user	0.34879	cell	0.40406	model	0.03125
patient	0.33065	clinical	0.40222	disease	0.03125
medical	0.3246	science	0.40085	text	0.03125
science	0.30847	network	0.40039	big	0.03125
model	0.30444	user	0.39307	record	0.03125
image	0.29637	model	0.37599		
digital	0.29032	digital	0.37474		
clinical	0.28831	practice	0.37369		
practice	0.26613	image	0.3722		
text	0.20766	cancer	0.37187		
gene	0.20766	disease	0.36951		
disease	0.20161	text	0.34264		
cancer	0.19758	gene	0.3379		
big	0.13306	researcher	0.33593		
researcher	0.12702	big	0.31234		
protein	0.10081	record	0.28038		
record	0.07863	protein	0.26988		
repository	0.06452	repository	0.229		
annotation	0.02419	annotation	0.13063		

넓게 논의되고 있음을 확인할 수 있었다. 이어 'system', 'research', 'information', 'human' 등의 단어도 네트워크 전역적으로 높은 영향력을 가진 용어로 나타났다. 반면, 'protein', 'record', 'repository', 'annotation' 등의 단어는 네트워크 전반에서 매우 낮은 영향력을 가진 것으로 나타났다. 이와 같은 결과를 통해 데이터 분석, 데이터 큐레이션 시스템, 인간을 대상으로 한 데이터 연구 등은 데이터 큐레이션 관련 연구에서 전역적으로 중요한 주제지만, 단백질이나 의료 기록, 데이터 리포지터리, 데이터 주석 등은 전역적으로 영향력이 낮은 주제임을 알 수 있다.

지역 중심성인 최근접이웃중심성(NNC)에서는 'research'가 높은 지수를 나타냈으며, 'gene', 'system' 등도 상위권에 포함되었다. 'gene'의 경우 삼각매개중심성(TBC)과 평균연관성(AVGSIM)에서 하위권에 위치했으나, 최근접이웃중심성(NNC)에서는 2위에 올라 유전 연구는 전반적인 연구 주제라기 보다는 독립적인 주제 영역을 갖고 있는 것으로 나타났다.

5. 결론 및 시사점

본 연구의 목적은 국외 데이터 큐레이션의 연구 동향을 분석하는 것이다. 이를 위해 Scopus와 WoS에서 'data curation'을 검색하여 중복 제거와 관련 없는 논문을 제외한 총 1,797건의 논문, 학술대회 발표자료 등의 학술 정보를 분석 대상으로 선정하였다. 연구 방법으로는 대량의 텍스트에서 의미 있는 정보를 추출할 수 있는 텍스트 마이닝 기법을 채택하였으며, 키워드 빈도 분석과 LDA 토픽 모델링을 수행하여 주요 연구 주제를 도출하였다. 도출된 토픽의 키워드를 바탕으로 가중 네트워크 중심성 분석을 실시하였다. 연구 결과는 다음과 같다.

첫째, 국외 데이터 큐레이션 연구는 2001년 처음 등장하였고 2009년부터 매년 증가하기 시작하여 2019년 213건으로 가장 많이 수행되었다. 최근에는 연 100건대 정도의 연구가 꾸준히 수행되고 있다.

둘째, 데이터 큐레이션 연구는 『Database: the journal of biological databases and curation』, 『Nucleic Acids Research』, 『Studies in Health Technology and Informatics』와 같은 학술지에서 주로 발표되었으며, 이는 생물정보학, 의료정보학, 생명과학, 컴퓨터과학 분야에서 데이터 큐레이션 연구가 활발히 이루어지고 있음을 시사한다.

셋째, 전처리된 키워드의 분석 결과, 최빈도 단어는 'research'로, 주로 의학, 생의학, 임상의학 및 연구지원 측면에서 사용되고 있었다. 'information'은 정보 처리와 관련하여 자주 등장하며, 'database', 'model', 'annotation' 등은 데이터베이스와 데이터 큐레이션 모델, 주석 연구와 관련이 있었다. 또한 'gene', 'human', 'genetic', 'clinical' 등의 단어는 생물학 및 의학 연구에서의 데이터 큐레이션과 밀접한 관련이 있음을 보여준다.

넷째, LDA 토픽 모델링 분석을 통해 데이터 큐레이션 연구에서 다섯 가지 주요 토픽이 도출되었다. '과학 데이터 관리와 디지털 리포지터리'가 가장 큰 비중을 차지하며, '유전자 및 단백질 데이터베이스 연구'와 '의료 및 생물학적 데이터의 주석과 모델링'이 뒤를 이었다. 각 토픽의 비중은 시간에 따라 변동이 있었으며,

특히 '과학 데이터 관리와 디지털 리포지터리'는 2000년대 초반부터 높은 비중을 유지하다가 최근 다시 증가하는 추세를 보였다. '임상 의료 데이터의 품질 제고'는 최근 증가하는 연구 주제로 확인되었다.

다섯째, 키워드 네트워크 분석 결과, 'analysis'가 전역 중심성에서 높은 지수를 기록하며 데이터 분석이 데이터 큐레이션 연구에서 중요한 주제임을 나타냈다. 이어 'research', 'system', 'information'도 네트워크 전역에서 높은 영향력을 보였다. 반면 'protein'과 'annotation' 등은 낮은 영향력을 가진 것으로 나타났다. 지역중심성에서는 'research'가 가장 높게 나타났으며, 이어 'gene', 'system' 등이 상위권에 올랐다. 'gene'의 경우 전역중심성에는 낮은 순위를 보였으나 지역중심성에서는 높은 순위에 올라 독립적인 주제로 중요한 위치를 차지하고 있음을 보여주었다.

이러한 연구 결과를 통해 다음과 같은 시사점을 도출할 수 있었다. 첫째, 데이터 큐레이션 연구는 최근 몇 년간에도 꾸준히 100건 이상의 연구가 수행되고 있다는 점에서, 이 분야의 연구가 지속적으로 성장하고 있음을 보여준다.

둘째, 데이터 큐레이션 연구는 문헌정보학 뿐 아니라 생물정보학, 의료정보학, 생명과학, 컴퓨터과학 등 다양한 분야에서 이루어지고 있고 이는 데이터 큐레이션을 연구할 때 다양한 학문과의 융합이 필요함을 시사한다.

셋째, 국외 데이터 큐레이션 연구는 5개의 토픽 중 4개의 토픽(1, 3, 4, 5)이 모두 과학과 관련된 것으로 나타났다. 데이터 큐레이션의 상위 범주인 디지털 큐레이션 연구 동향 분석에서 디지털 도서관, 디지털 보존, 정보자원관리(김판준, 2015), 디지털 보존, 연구데이터, 데이터 관리(Kim, 2014), 기관, 콘텐츠, 도서관, 큐레이션, 보존 등(박민석, 이지수, 2024) 문헌정보학의 주제가 주요 주제로 나타난 것과 비교되는 부분이다.

넷째, 국내 연구에서는 주로 문헌정보학 영역에서 데이터 큐레이션 연구가 이루어졌고, 그 중에서도 모델, 프레임워크, 가이드라인 연구가 주로 이루어졌다. 국내 연구에서도 국외 연구와 마찬가지로 다른 학문 분야로의 데이터 큐레이션 연구 확산이 필요하며, 문헌정보학 영역에서도 국외에서 자주 다뤄지는 주석과 같은 메타데이터 연구나, 분석 방법, 시스템 연구와 같은 분야로 연구 주제를 확장할 필요가 있다.

본 연구는 그동안 수행되지 않았던 데이터 큐레이션 국외 연구 동향을 분석했다는 점에서 의의가 있다. 그러나 본 연구의 한계는 다음과 같다. 불용어 처리 과정에서 자주 사용되는 단어들이 초록 작성시 일반적으로 포함된 단어인지 혹은 연구 주제와 관련된 것인지 판단하기 위해 불용어 제거 전과 후로 나누어 빈도 분석을 수행하고, 저자 키워드만을 대상으로 빈도 분석을 수행하여 살펴보는 등 노력을 기울였으나, 연구자의 주관적인 견해를 완전히 배제할 수 없어 좀 더 객관적인 연구를 위해 향후 보완해야 할 사항으로 판단된다.

국내 데이터 큐레이션 연구가 충분히 수행되었을 때, 국내 연구 동향 분석을 통해 국외 연구와 비교가 진행된다면 국내 연구와 국외 연구 간의 차이점과 공통점을 보다 명확하게 이해할 수 있을 것이며, 이를 바탕으로 국내 데이터 큐레이션 연구 발전 방향을 보다 잘 모색할 수 있을 것이다.

참 고 문 헌

과학기술정보통신부 (2023). 국가연구데이터 관리 및 활용 촉진에 관한 법률 제정안.

과학기술정보통신부, 한국과학기술기획평가원 (2020). 국가연구개발사업 연구관리 표준매뉴얼.

김진희, 최서연, 임철일, 함윤희 (2019). 연구지원을 위한 데이터 큐레이션 사서교육 프로그램 개발. 교육문화연구, 25(6), 757-779. https://doi.org/10.24159/joec.2019.25.6.757

김판준 (2015). 디지털 큐레이션 연구동향 분석과 과제: 문헌정보학 분야를 중심으로. 정보관리학회지, 32(1), 265-295. https://doi.org/10.3743/KOSIM.2015.32.1.265

박대영, 김덕현, 김건욱 (2021). 토픽 모델링 기반의 국내외 공공데이터 연구 동향 비교 분석. 디지털융복합연구, 19(2), 1-12. https://doi.org/10.14400/JDC.2021.19.2.001

박민석, 이지수 (2024). 체계적 문헌고찰을 통한 국내 디지털 큐레이션 연구동향 분석. 한국기록관리학회지, 24(2), 41-63. https://doi.org/10.14404/JKSARM.2024.24.2.041

유화선, 정도범 (2023). 텍스트 마이닝을 활용한 데이터 정책 연구동향 분석. 한국콘텐츠학회논문지, 23(3), 17-26. https://doi.org/10.5392/JKCA.2023.23.03.017

이민정 (2023). 미국의 「오픈사이언스의 해」 선포와 정책적 시사점 (KISTEP 브리프 59). 한국과학기술기획평가원.

이상헌 (2020). 공공데이터 기록관리 활성화와 큐레이션 활용방안 연구: 공공데이터포털, 혜안시스템을 중심으로. 기록과 정보·문화 연구, (11), 115-153. https://doi.org/10.23035/kaics.2020.1.11.115

이수상 (2012). 네트워크 분석 방법론. 서울: 논형.

이원상, 손소영 (2015). 공간빅데이터 연구 동향 파악을 위한 토픽모형 분석. 대한산업공학회지, 41(1), 64-73. https://doi.org/10.7232/JKIIE.2015.41.1.064

이유경, 정은경 (2015). 데이터 큐레이터의 핵심 직무 요건 고찰에 관한 연구. 한국비블리아학회지, 26(3), 129-150. https://doi.org/10.14699/kbiblia.2015.26.3.129

이재윤 (2006). 계량서지적 네트워크 분석을 위한 중심성 척도에 관한 연구. 한국문헌정보학회지, 40(3), 191-214.

이재윤 (2013). Tnet과 WNET의 가중 네트워크 중심성 지수 비교 연구. 정보관리학회지, 30(4), 241-264. https://doi.org/10.3743/KOSIM.2013.30.4.241

이재윤 [n.d.]. WNET (Weighted Network analysis): PFNet, PNNC, and Weighted Network Centralities (v.0.4.1) [Computer software].

이정미 (2020). 교수학습활동 지원 개선을 위한 대학도서관의 데이터 큐레이션 연구. 한국문헌정보학회지, 54(1), 175-195. https://doi.org/10.4275/KSLIS.2020.54.1.175

이제욱 (2022). 스포츠 공공데이터 큐레이션 서비스 활성화를 위한 법적 연구. 스포츠와 법, 25(4), 195-205. http://doi.org/10.19051/kasel.2022.25.4.195

이현조, 조한진, 채철주 (2022). 디지털 농업을 위한 디지털 농업 데이터 큐레이션 서비스 방안 연구. 한국컴퓨터정보학회논문지, 27(2), 171-177. https://doi.org/10.9708/jksci.2022.27.02.171

이혜경, 이용구 (2023). 동시출현단어 분석을 이용한 오픈 데이터 분야의 지적 구조 분석. 정보관리학회지, 40(4), 429-450. https://doi.org/10.3743/KOSIM.2023.40.4.429

정선경 (2022). 텍스트 마이닝을 활용한 데이터 거버넌스 연구 동향 분석: 2009년~2021년 국내 학술지 논문을 중심으로. 디지털융복합연구, 20(4), 133-145. https://doi.org/10.14400/JDC.2022.20.4.133

진보라, 윤유라 (2017). 데이터 큐레이션 구현을 위한 통합적 가이드라인 연구. 예술인문사회융합멀티미디어논문지, 7(6), 767-776. https://doi.org/10.35873/ajmahs.2017.7.6.072

최동훈, 박재원, 김병규, 신진섭 (2017). 커뮤니티 주도적 과학 데이터 큐레이션 협업 환경의 개발. 한국콘텐츠학회논문지, 17(9), 1-11. https://doi.org/10.5392/JKCA.2017.17.09.001

한나은 (2023). 활동이론을 중심으로 한 연구데이터 큐레이션 개념 모델 제안. 한국도서관·정보학회지, 54(1), 167-190. https://doi.org/10.16981/kliss.54.1.202303.167

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

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Hemphill, L., Pienta, A., Lafia, S., Akmon, D., & Bleckley, D. A. (2022). How do properties of data, their curation, and their funding relate to reuse?. Journal of the Association for Information Science and Technology, 73(10), 1432-1444. https://doi.org/10.1002/asi.24646

Johnston, L. R. (2017). Curating Research Data Volume One: Practical Strategies for Your Digital Repository. Chicago: Association of College and Research Libraries.

Johnston, L. R., Curty, R., Braxton, S. M., Carlson, J., Hadley, H., Lafferty-Hess, S., Luong, Hoa., Petters, Jonathan L., & Kozlowski, W. A. (2024). Understanding the value of curation: A survey of US data repository curation practices and perceptions. PloS One, 19(6), e0301171. https://doi.org/10.1371/journal.pone.0301171

Kim, J. (2014). Growth and trends in digital curation research: The case of the international journal of digital curation. Proceedings of the American Society for Information Science and Technology, 51(1), 1-4. https://doi.org/10.1002/meet.2014.14505101074

Lee, J. Y., Syn, S. Y., & Kim, S. (2024). Global research trends in research data management: A bibliometrics approach. Journal of Librarianship and Information Science, 09610006241239083.

https://doi.org/10.1177/09610006241239083

Mannheimer, S. (2024). Scaling Up: How Data Curation can Help Address Key Issues in Qualitative Data Reuse and Big Social Research. Cham: Springer International Publishing AG.

Marsolek, W., Wright, S. J., Luong, H., Braxton, S. M., Carlson, J., & Lafferty-Hess, S. (2023). Understanding the value of curation: A survey of researcher perspectives of data curation services from six US institutions. PloS one, 18(11), e0293534. https://doi.org/10.1371/journal.pone.0293534

Mohammadi, E. & Karami, A. (2022). Exploring research trends in big data across disciplines: A text mining analysis. Journal of Information Science, 48(1), 44-56. https://doi.org/10.1177/0165551520932855

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., Bonino da Silva Santos, L., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C. 't, Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

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

Choi, Dong Hoon, Park, Jae Won, Kim, Byung kyu, & Shin, Jin Sup (2017). Development of collaborative environment for community-driven scientific data curation. The Journal of the Korea Contents Association, 17(9), 1-11. https://doi.org/10.5392/JKCA.2017.17.09.001

Han, Na eun (2023). Proposal of a conceptual model for research data curation based on activity theory. Journal of Korean Library and Information Science Society, 54(1), 167-190. https://doi.org/10.16981/kliss.54.1.202303.167

Han, Sang woo (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. https://doi.org/10.16981/kliss.54.4.202312.393

Jeong, Sun Kyeong (2022). The study on data governance research trends based on text mining: Based on the publication of Korean academic journals from 2009 to 2021. Journal of Digital Convergence, 20(4), 133-145. https://doi.org/10.14400/JDC.2022.20.4.133

Jin, Bo Ra & Youn, You Ra (2017). A study on the guidelines for the development of data curation policy. Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, 7(6), 767-776. https://doi.org/10.35873/ajmahs.2017.7.6.072

Kim, Jin Hee, Choi, Seo Yeon, Lim, Cheol Il, & Ham, Yoon Hee (2019). Development of a data curation training program for research support for librarians. Journal of Education & Culture, 25(6), 757-779. https://doi.org/10.24159/joec.2019.25.6.757

Kim, Pan Jun (2015). An analytical study on research trends of digital curation: Focused on library and information science. Journal of the Korean Society for Information Management, 32(1), 265-295. https://doi.org/10.3743/KOSIM.2015.32.1.265

Lee, Hye Kyung & Lee, Yong Gu (2023). Intellectual structure analysis on the field of open data using co-word analysis. Journal of the Korean Society for Information Management, 40(4), 429-450. https://doi.org/10.3743/KOSIM.2023.40.4.429

Lee, Hyun Jo, Cho, Han Jin, & Chae, Choel Joo (2022). A study on Digital Agriculture Data Curation Service Plan for Digital Agriculture. Journal of the Korea Society of Computer and Information, 27(2), 171-177. https://doi.org/10.9708/jksci.2022.27.02.171

Lee, Jae Yoon (2006). Centrality measures for bibliometric network analysis. Journal of the Korean Society for Library and Information Science, 40(3), 191-214.

Lee, Jae Yoon (2013). A comparison study on the weighted network centrality measures of tnet and WNET. Journal of the Korean Society for Information Management, 30(4), 241-264. https://doi.org/10.3743/KOSIM.2013.30.4.241

Lee, Jae Yoon [n.d.]. WNET (Weighted Network Analysis): PFNet, PNNC, and Weighted Network Centralities (v.0.4.1) [Computer software].

Lee, Je Wook (2022). Strategy to activate sports public data curation service. Sports Entertainment and Law, 25(4), 195-205. http://doi.org/10.19051/kasel.2022.25.4.195

Lee, Jung Mee (2020). A study on data curation of university libraries for improving teaching and learning support. Journal of the Korean Society for Library and Information Science, 54(1), 175-195. https://doi.org/10.4275/KSLIS.2020.54.1.175

Lee, Min Jung (2023). The U.S. Declaration of the "Year of Open Science" and Its Policy Implications (KISTEP Brief 59). Korea Institute of Science and Technology Evaluation and Planning.

Lee, Sang Hyuen (2020). A study on the activation of public data record management and the

application of curation: Focusing on public data portal, Hyean system. The Korean Journal of Archival, Information and Cultural Studies, (11), 115-153. https://doi.org/10.23035/kaics.2020.1.11.115

Lee, Soo Sang (2012). Network Analysis Methodology. Seoul: Nonhyeong.

Lee, Won Sang & Sohn, So young (2015). Topic model analysis of research trend on spatial big data. Journal of the Korean Institute of Industrial Engineers, 41(1), 64-73. https://doi.org/10.7232/JKIIE.2015.41.1.064

Lee, You Kyong & Chung, Eun Kyung (2015). An investigation on core competencies of data curator. Journal of the Korean Biblia Society for Library and Information Science, 26(3), 129-150. https://doi.org/10.14699/kbiblia.2015.26.3.129

Ministry of Science and ICT & Korea Institute of Science and Technology Evaluation and Planning (2020). Standard Manual for Research Management in National R&D Projects.

Ministry of Science and ICT (2023). Draft Legislation on the Promotion of National Research Data Management and Utilization.

Park, Dae Yeong, Kim, Deok Hyeon, & Kim, Keun Wook (2021). Topic modeling-based domestic and foreign public data research trends comparative analysis. Journal of Digital Convergence, 19(2), 1-12. https://doi.org/10.14400/JDC.2021.19.2.001

Park, Min Seok & Lee, Ji Soo (2024). A systematic review of trends of domestic digital curation research. Journal of Korean Society of Archives and Records Management, 24(2), 41-63. https://doi.org/10.14404/JKSARM.2024.24.2.041

Yu, Hwasun & Jung, Do Bum (2023). A research trend analysis of data policy using text mining. The Journal of the Korea Contents Association, 23(3), 17-26. https://doi.org/10.5392/JKCA.2023.23.03.017