Journal of Korean Library and Information Science Society
Vol. 56, No. 2

한국문학 분야 KORMARC 서지레코드를 활용한 저작 식별 개선 방안 연구

A Study on Improving Work Identification Using KORMARC Bibliographic Records in the Field of Korean Literature

나 상 오(Sangoh Na)*
강 우 진(Woojin Kang)**
이 종 욱(Jongwook Lee)***

〈 목 차 〉

Ⅰ. 서 론
Ⅱ. 개념적 배경 및 선행연구
Ⅲ. 연구 설계
Ⅳ. 연구 결과
Ⅴ. 결 론

요 약 : 본 연구는 국립중앙도서관에서 제공하는 한국문학 분야 KORMARC 서지레코드를 활용하여 저작 식별의 개선 방안을 제안하였다. 기존 연구에서는 서지개체를 식별할 때 레코드별 저작 세트 간 완전 일치를 기준으로 저작을 군집화하는 방식을 주로 사용하였다. 이러한 방식은 동일 저작이라도 표기 방식 등의 차이로 인해 서로 다른 저작으로 분리되는 문제가 발생하였다. 이 문제를 해결하기 위해 본 연구에서는 저자명과 표제를 추출하는 과정에서 한문-한글 변환, 서양인명의 성-이름 순서 변환, 저자 역할어 및 관제 제거 등의 전처리 과정을 수행하였다. 또한, 네트워크 분석 기법을 적용하여 저자명 및 표제를 기준으로 각각의 네트워크를 생성한 후, 네트워크 간 레코드 집합의 교집합을 식별하는 과정을 통해 저작을 식별하고 관련 레코드를 연결하였다. 연구 결과, 453,846건의 서지레코드에서 268,684개의 저작을 식별하였으며, 기존의 문자열 완전 일치 방식보다 더 정교한 저작 식별이 가능함을 확인하였다. 다만, 서지레코드의 입력 상태에 따라 저작 식별에 일부 한계가 나타났으며, 이를 보완하기 위한 후속 연구가 필요함을 알 수 있었다. 본 연구에서 제안한 방법은 저작 식별 과정을 개선함으로써 서지개체의 식별은 물론, 장기적으로는 서지레코드의 링크드데이터 변환에도 기여할 것으로 기대된다.

주제어 : 서지레코드, 저작 식별, 네트워크 분석, 한국문헌자동화목록형식

ABSTRACT : This study proposes an improved method for identifying works using KORMARC bibliographic records in the field of Korean literature, provided by the National Library of Korea. Previous research primarily clustered works based on exact matches between bibliographic record sets. However, this approach often led to the separation of identical works due to differences in notation and other variations. To address this issue, this study applied preprocessing techniques that included Chinese character-to-Korean conversion, reordering of Western names (surname first, then given name), and the removal of author role terms and added titles when extracting author names and titles. Additionally, a network analysis technique was employed to construct separate networks based on author names and titles. The process of identifying the intersection of record sets across these networks was then used to identify works and link related records. As a result, 268,684 works were identified from 453,846 bibliographic records, demonstrating a more refined work identification process compared to exact string matching methods. Nevertheless, some limitations were observed in identifying works due to errors in some bibliographic record inputs, highlighting the need for further research to address these issues. The proposed method is expected to improve the work identification process, thereby improving bibliographic entity identification and, in the long term, contributing to the transformation of bibliographic records into linked data.

KEYWORDS : Bibliographic Record, Work Identification, Network Analysis, KORMARC

* 경북대학교 대학원 문헌정보학과 석박사통합과정(gg4518@knu.ac.kr / ISNI 0000 0005 0490 1663) (제1저자)
** 경북대학교 대학원 문헌정보학과 박사과정(rkddnwls1234@knu.ac.kr / ISNI 0000 0005 0659 7247) (공동저자)
*** 경북대학교 문헌정보학과 부교수(jongwook@knu.ac.kr / ISNI 0000 0004 6830 6145) (교신저자)

• 논문접수: 2025년 2월 28일 • 최초심사: 2025년 3월 6일 • 게재확정: 2025년 3월 11일
• 한국도서관·정보학회지, 56(2), 109-132, 2025. http://dx.doi.org/10.16981/kliss.56.2.202506.109

Ⅰ. 서 론

국제도서관협회연맹(International Federation of Library Associations and Institutions, 이하 'IFLA')은 1998년 FRBR을 시작으로 FRAD, FRSAD로 이어지는, 서지개체와 그 관계에 대하여 정의하는 개념 모형을 발표하였다. 이후 목록 환경 변화의 영향으로 새로운 개념 모형에 대한 요구가 있었고, 2017년에는 기존 FR 모형을 통합한 형태의 LRM을 발표하였다. LRM은 RDA를 포함한 목록 규칙들의 개정은 물론, MARC와 같은 입력 포맷에도 영향을 미쳤다. 목록 환경 변화의 원인은 다양하지만 2010년대 이후로는 시멘틱웹의 발전과 함께 발견된 MARC의 한계와 링크드데이터로의 전환이 주로 언급되며, 이러한 요인들이 LRM 개발에 영향을 준 것으로 간주된다(노지현 외, 2023). 기존 MARC 레코드의 한계점으로 기계가 이해할 수 없는 문자열 값으로 구성되는 경우가 많아 데이터의 연결을 지향하는 시멘틱 웹에서 활용이 어렵다는 점이 지적이 있다. 이에 따라 MARC의 대체제에 대한 필요성이 대두되었으며, 링크드데이터에 기초하는 새로운 프레임워크가 등장하였다(이미화 외, 2022).

기존 목록을 링크드데이터로 변환하기 위해서는 MARC 기반 서지레코드에 반영된 서지개체의 구분과 식별이 필수적이다. 그 중에서도 목록 작성 시 하나의 기준 단위로서 관련 표현형, 구현형 등과 같은 정보를 집중화하는 '저작'의 식별이 특히 중요하다(이혜원, 2011). 또한, 저작은 저작 이름의 전거통제 기능을 하므로 효율적인 전거통제를 위해서 우선적으로 식별될 필요가 있다(이미화, 정연경, 2008).

서지개체라는 개념을 KORMARC 레코드에 적용하고 이를 식별하거나 저작 세트를 추출하는 연구는 FRBR 제 1그룹의 발표 이후 지속적으로 수행되어왔다. 서지개체를 MARC 레코드에 적용하고자 알고리즘을 개발한 Online Computer Library Center(이하 'OCLC')와 미국의회도서관(Library of Congress, 이하 'LC')의 연구부터 국내 서지레코드를 대상으로 저작 세트를 생성하고 각 레코드를 매칭하는 연구까지 다양하게 수행되었다(김현희 외, 2007; 노지현, 2008; 조재인, 2004).

국내에는 샘플 레코드를 대상으로 저작 식별 요소를 추출하고 클러스터링을 수행한 시도가 존재하는데, 이는 국내 목록 환경에 적합한 저작 식별 요소를 찾고 집중화한다는 점에서 큰 의의가 있다. 그렇지만 기존 연구들은 레코드별 저작 색인('저자명+표제') 매칭에서 '완전 일치'를 기준으로 하므로 실제로는 동일 저작이나 하나의 요소 값에서라도 부분적인 차이가 있으면 서로 다른 저작으로 분리되는 문제가 발생한다.

이에 본 연구에서는 저자명에서 역할어, 한문어 및 서양인명 표기로 인해 차이가 발생하는 문제와 표제에서 관제, 한문 표기로 인한 구분의 문제를 해결하고자 하였다. 또한 저자명과 연결한 레코드 집합과 표제와 연결한 레코드 집합 간의 중복되는 레코드 집합을 식별한 후 해당 집합에서 가장 빈번하게 출현하는 저자명과 표제를 결합하여 저작을 식별하는 방안을 제안하였다. 이러한

방법을 국립중앙도서관의 한국문학 분야 KORMARC 전체 서지레코드에 적용하였으며, 그 결과를 살펴보았다. 본 연구 결과는 목록 데이터의 링크드데이터로의 변환에 있어 저작의 정교한 식별 과정에 기여할 수 있을 것이다.

Ⅱ. 개념적 배경 및 선행연구

1. LRM 개체 및 저작 개념

LRM은 IFLA가 기존에 발표한 서지(FRBR), 전거(FRAD), 주제명(FRSAD) 개념 모형을 통합한 것으로 개체(entity), 속성(attribute), 관계(relationship)의 3가지 요소로 구성되어 있다. '개체'는 총 11개로 구성된다. 최상위 계층의 레(Res)는 물질적이거나 형식적인 사물과 개념, 추상적 개념을 모두 포함하는 최상위 개체로 서지 세계의 모든 것이 해당한다고 볼 수 있다. 나머지 개체는 모두 레의 직간접적인 하위 클래스로 간주된다.

저작(Work)은 창작물의 지적 혹은 예술적 내용이며, 표현형(Expression)은 저작의 내용을 전달하는 뚜렷하게 구분되는 기호의 결합을 의미한다. 구현형(Manifestation)은 지적, 예술적 내용과 물리적 형태 측면에서 공통 특성을 공유한다고 여겨지는 모든 수록매체의 집합이다. 개별자료(Item)는 지적 혹은 예술적 내용을 전달하기위해 신호를 수록한 객체를 의미한다. 에이전트(Agent)는 서지개체와 창작, 제작, 배포, 소장, 수정의 관계를 가지는 개인(Person)과 집합에이전트(Collective Agent)를 의미한다. 노멘(Nomen)은 개체와 이를 지칭하는 명칭과의 연계를 의미하며, 장소(Place)는 공간 범위를, 시간범위(Time-span)는 시간적 범위를 의미한다(Riva et al., 2017).

저작은 지적 또는 예술적 창작물이라는 점에서 그 개념이 추상적이며, 개체의 명확한 경계를 규정하기 어렵다. 이와 관련하여 Tiillett(2001)은 파생관계의 정도에 따라 저작과 표현형의 경계를 구분하여 저작의 개념을 구체화하였다. 축약판, 개정, 번역, 편곡 등의 가벼운 변형은 모두 동일 저작으로 구분하였으며, 개작 또는 장르 변경과 같은 변화가 있는 경우에는 새로운 저작으로 취급하였다.

2. 선행연구

가. 서지레코드의 FRBR 적용 연구

FRBR 모형의 등장 이후, 기구축된 MARC 레코드를 FRBR로 구현할 필요성이 인식됨에 따라,

이를 실현하기 위한 방안에 대한 OCLC와 LC 차원의 논의가 있었다(Hickey & Toves, 2005; Library of Congress, 2004). 수작업을 통하여 구현하는 것은 불가능하다는 판단 하에 자동화 알고리즘 설계의 필요성이 대두되었고, OCLC는 WorldCat 데이터를 FRBR로 구현하는 알고리즘을 설계하였다. OCLC의 알고리즘은 저작 세트 식별 알고리즘으로, NACO 전거파일을 활용하여 서지레코드상의 저자 및 서명을 비교하고 일치하는 경우 이를 서지레코드의 대표 저작 세트로 할당하였다. 이후, 대표 저작 세트를 전체 서지레코드와 비교하여 동일한 저작 세트가 나타나면 이를 동일 저작으로 간주하였다(OCLC, 2005; 2009). LC 또한 MARC21 레코드를 FRBR로 디스플레이하기 위한 도구를 고안하였다. OCLC와 유사하게 저작 단위 매칭을 우선적으로 실시하였으나 전거파일 기반의 대표 저작세트를 구축하는 과정은 생략되었으며, 서지레코드로부터 직접 저자명과 서명을 추출하고 상호 매칭하여 군집화하였다(Library of Congress, 2004).

OCLC와 LC의 알고리즘은 구축된 전거데이터 없이는 적용이 어렵다는 한계가 있다(이미화, 정연경, 2008; 조재인, 2004). 이에 국내 목록 환경에 적용하는 것이 어려움에 따라 선행연구에서는 어떠한 KORMARC 필드를 활용하여 서지개체를 식별할지 모색하고 알고리즘을 설계하는 등의 방향으로 연구가 진행되었다. 구체적으로 서지적 관계의 규명, 개체 식별을 위하여 기존 KORMARC 서지레코드에 FRBR 모형을 적용하는 연구들이 수행되었다. 이성숙과 이현주(2013)는 국악자료의 서지레코드를 대상으로 FRBR 모형 적용방안을 연구하였다. 국악자료의 서지적 관계 특성을 FRBR의 관계 기반으로 분석하였으며, 이를 바탕으로 FRBR 모형에 기반하며 국악자료의 특성을 반영하는 서지적 관계 유형을 제안하였다.

김정현(2015)은 FRBR 및 RDA 등에서 나타난 서지적 관계 유형을 분석하였고, 이를 근거로 국립중앙도서관의 사서오경 관련 서지레코드를 조사하여 KORMARC 레코드의 서지적 관계 기술 한계를 파악하였다. 이에 저작을 관계 유형별(예: 번역, 해설, 비평 등)로 추출하여 나타낼 필요성이 있음을 주장하였고, 전거형 접근점으로서 통일표제를 활용하여 관련 저작 간의 연결 장치로 활용할 것을 제안하였다.

기존의 서지레코드로부터 FRBR 제 1집단의 서지개체를 식별하려는 연구 또한 수행되었다. 김정현(2007)은 FRBR 모형의 저작 개념을 기반으로 KORMARC 형식 내 저작 유형을 분석하여 유용성을 확인하고자 하였다. 그 결과, 하나의 저작이 복수의 표현형과 구현형을 가지는 것과 같이 복잡한 서지적 관계가 존재할수록 유용하다는 것을 밝혔다.

김현희 외(2007)는 통합서지용 KORMARC 데이터베이스에 FRBR 모형의 적용 가능성을 검증하고자 하였다. 저작, 표현형, 구현형 수준으로 구분 및 그룹핑하는 알고리즘을 설계하여 107건의 음악자료 레코드에 적용하였고 성공적으로 FRBR화된 80건의 레코드를 대상으로 데이터베이스 실험 시스템을 구현하였다.

노지현(2008)은 FRBR 모형을 적용하기 위한 국내외 사례를 분석하고 FRBR 제 1집단에 주목

하여 KORMARC의 데이터로부터 개체 식별 요소를 파악하였다. 국립중앙도서관 서지레코드로부터 161건의 표본 레코드를 선정하여 서지개체별 클러스터링을 시도하였고, 전거통제가 적용되지 않은 목록 환경과 일관성 없이 기술된 KORMARC 레코드가 서지개체 식별을 위한 데이터 추출에 문제가 된다는 것을 확인하였다.

나. KORMARC 서지레코드 저작 식별 연구

조재인(2004)은 전거레코드가 적절히 구축되지 않은 국내 도서관 데이터베이스를 대상으로 저작 식별을 시도하였다. 기본표목으로 통일표제(130)가 있을 경우 이를 저작 식별 요소로 우선 활용하였다. 130 필드 부재 시 서지레코드의 표제(240/245/505/730/740 필드의 첫 번째 표제)와 저자명(100/110/111/700/710/711 필드)에서 '저자명+표제'를 추출하였다. 추출한 저자명과 표제는 전체 레코드와의 1:1 비교를 거친 후 일정 기준 조건 이상을 충족하는 경우에만 동일 저작으로 군집하되, 전문가의 적합성 검증을 거쳐 부적격 레코드를 제거하는 방안을 제시하는 등 대안 또한 모색하였다.

노지현(2008)은 저작 개체로의 클러스터링을 위하여 표본 레코드의 저자명(100/700/900)과 표제((240▼a/245▼a +▼h/740▼a) 데이터를 모두 추출하였고, 이를 각각 조합하여 '저자명+표제' 형태의 데이터를 생성하고 상호 일치하는 경우 동일 저작으로 간주하는 방식을 취하였다. 이때 해외 사례와 같이 전거레코드 기반의 대표 엔트리를 선정하는 것은 불가능했기 때문에 동일 저작으로 군집화되어야 하나 상이한 저작으로 취급된 복수 저작 문제가 있음을 확인하였다.

이미화와 정연경(2008)은 FRBR 알고리즘을 실재 목록 시스템에 적용하여 문제점과 해결 방안을 탐색하였다. 표제와 저자명 속성이 담긴 필드를 파악하여 알고리즘을 설계하였는데, 이때 저자명의 기본기입이 없는 경우 부출 저자를 '700 ▼a', '710 ▼a, ▼b', '711 ▼a'에서 2개까지 추출하였다. 이후 서지레코드의 군집화 과정에서 부출저자를 활용한 경우 2인의 저자 중 하나만 일치하더라도 동일 저작으로 취급하였다.

김정현 외(2015)는 국내외 FRBR 구현 알고리즘 개발 사례의 장단점을 비교하고, KORMARC 레코드에 적용하였다. 저작 클러스터링을 위하여 '저자명+표제' 형식으로 전거형 접근점을 작성할 수 있도록 알고리즘을 설계하였으며, 실험데이터 세트 적용 및 분석하였다.

저자명을 식별함에 있어 역할 용어의 부재를 극복하고자 한 연구도 있었다(윤재혁 외, 2020). 저자명 식별을 위하여 흔히 사용하는 기본표목(1XX) 및 부출표목(7XX) 이외에 표제 및 책임표시사항(245)의 식별기호 ▼d와 ▼e에 나타난 역할 용어를 반영하여 역할어사전을 구축하였다. 또한 '저자명+표제'의 형태로 매칭하여 저작을 식별하는 기존 연구와 달리, 저자명과 표제를 구분하고 각각 코사인 유사도를 활용하여 저작을 클러스터링하였다.

선행연구에서 활용된 저작 식별 요소를 요약한 결과는 〈표 1〉과 같다. 서지레코드에서 해당

요소들을 추출하여 저작 군집화를 위해 조합하는 경우 일반적으로 '저자명+표제' 형태의 세트로 생성하였다. '통일표제'가 존재하는 경우 이를 우선하여 저자명 없이 활용한 사례가 존재하며, 서지 레코드의 어떠한 필드에서도 저자명을 확인할 수 없는 경우 표제만을 활용하기도 하였다. 군집화 과정에서는 일반적으로 저작 세트의 텍스트 완전 일치를 기준으로 하였고, 이로 인하여 동일 저작이 서로 다른 저작 군집을 형성하는 문제가 발생하기도 하였다.

〈표 1〉 KORMARC 레코드에서의 저작 식별 요소

선행연구		저작 식별 요소	
KOMARC 통합서지용 개정(2014) 이전	조재인 (2004)	저자명	100, 110, 111, 700, 710, 711
		서명	240, 245, 505, 730, 740 필드의 첫 표제
		예시	〈Shakespeare, William / 햄릿〉
	김현희 외 (2007)	저자명	① 100▼a▼b▼c▼d, 110▼a▼b▼c▼d, 111▼a▼c▼d▼n 1XX 태그 부재 시, ② 700▼a▼b▼c▼d, 710▼a▼b▼c▼d, 711▼a▼c▼d▼n
		서명	240▼a▼d▼k▼m▼n▼p▼r, 243▼a▼d▼m▼n▼p▼r, 245▼a▼g▼k▼n▼p
			표제만을 활용하여 식별하는 경우: ① 130▼a▼d▼k▼m▼n▼p▼r ② 240▼a▼d▼k▼m▼n▼p▼r(또는 243▼a▼d▼m▼n▼p▼r, 245▼a▼g▼k▼n▼p)
		예시	-
	노지현 (2008)	저자명	100, 700, 900 / *500▼a 를 제외
		서명	240▼a, 245▼a +▼h, 740▼a
		예시	〈햄릿 / Shakespeare.〉, 〈햄릿 / Shakespeare, William.〉
	이미화, 정연경 (2008)	저자명	① 100 ▼a, 110 ▼a ▼b, 111 ▼a ② 700 ▼a, 710 ▼a ▼b, 711 ▼a ③ 507 ▼a
		서명	① 130 ▼a ② 240 ▼a ③ 507 ▼t, 245 ▼a
		예시	work_key(저작명)=서지 번호 → work_key(Little Prince/Saint)=54930292
KOMARC 통합서지용 개정(2014) 이후	김정현 (2015)	저자명	① 전거레코드 우선 ② 전거레코드 부재시 100, 110, 111 ▼a ▼d ③ 700, 710, 711 ▼a ▼d ▼e
		서명	240, 245, 246, 740 ▼a, ▼n, ▼p, ▼x 합집/선집의 경우: ① 740 , 부재 시 ② 505 ▼t 무저자의 경우: 245▼a 단독 활용 또는 130, 730 ▼a
		예시	-

Ⅲ. 연구 설계

1. 데이터 개요 및 전처리

본 연구에서는 국립중앙도서관의 국가서지과로부터 제공(제공일 2024. 3. 14.) 받은 한국십진분류법(KDC) 한국문학(810)에 해당하는 453,846건의 데이터를 활용하였다. 저자 및 표제 관련 필드를 추출하기에 앞서, 분석 대상 서지레코드에 작성된 필드 현황을 파악하였다. 453,846건의 레코드에 입력된 주요 필드의 분포는 〈표 2〉와 같다. 저자명과 표제의 추출에 주요하게 활용되는 표제와 책임표시사항(245) 필드는 모든 레코드에 입력되었음을 확인할 수 있다.

〈표 2〉 주요 표목, 표제 필드 작성 현황

표목 관련				표제 관련			
필드번호	필드명	레코드 수(건)	비율(%)	필드번호	필드명	레코드 수(건)	비율(%)
100	기본표목-개인명	38,495	8.48	240	통일표제	22	0.005
110	기본표목-단체명	7,466	1.65	245	**표제와 책임표시사항**	**453,846**	**100**
111	기본표목-회의명	1	0.0002	246	여러 형태의 표제	39,484	8.7
700	부출표목-개인명	398,827	87.88	505	내용주기	29,765	6.56
710	부출표목-단체명	25,917	5.71	730	부출표목-통일표제	68	0.01
711	부출표목-회의명	13	0.003	740	부출표목-비통제 관련/분출 표제	32,817	7.23
900	로컬표목-개인명	32,954	7.26				
910	로컬표목-단체명	3,036	0.67				
911	로컬표목-회의명	1	0.0002	총계		453,846	100

모든 서지레코드에 한국십진분류기호가 할당되어 있었다. 이 중 96.98%의 레코드에 하나의 분류기호가 할당되어 있었으며, 2개부터 최대 5개까지의 분류기호가 할당된 레코드 또한 존재하였다. 한국문학 분류 현황의 경우, 소설(43.75%), 시(26.78%) 순으로 높은 비중을 차지하고 있었고, 연설, 웅변(0.03%), 풍자 및 유머(0.47%)는 상대적으로 현저히 낮은 비중을 차지하고 있었다(〈표 3〉 참조).

〈표 3〉 한국문학 분류기호 할당 현황

한국문학(810)	한국문학 시(811)	한국문학 희곡(812)	한국문학 소설(813)	한국문학 수필(814)
46,513(10.25%)	121,529(26.78%)	4,996(1.1%)	198,579(43.75%)	33,526(7.39%)
한국문학 연설, 웅변(815)	한국문학 일기, 서간, 기행(816)	한국문학 풍자 및 유머(817)	한국문학 르포르타주 및 기타(818)	미사용(819)
157(0.03%)	8,856(1.95%)	2,145(0.47%)	41,441(9.13%)	10(0.002%)

다음으로 저작 식별 요소 추출에 앞서 분석 대상 서지레코드를 KORMARC 통합서지용 규칙에 따라 파싱하여 태그, 식별기호, 지시기호 등의 정보를 추출할 수 있도록 하였다. 이때 2개 이상의 저작이 하나의 도서에 수록되어 있는 전집, 총서 등은 다수의 표제와 저자명 조합을 생성하여 저작 간 식별에 어려움을 준다. 따라서, 원활한 분석 진행을 위하여 해당 레코드 35,960건은 분석에서 제외하였다. 전집, 총서 등의 식별은 다음과 같은 기준을 통하여 진행하였으며, 이를 통해 식별된 35,960건의 레코드는 분석에서 제외하였다.

2. 저작 식별 요소 추출

파싱된 서지데이터로부터 표제와 저자명 관련 필드 값을 추출하였다. 선행연구를 기반으로 저작 식별에 흔히 사용되는 기본표목(1XX), 표제와 표제관련필드(20X-24X), 부출표목(7XX) 필드로부터 정보를 추출하였으며, 이 외 로컬표목(9XX)에서도 식별기호와 지시기호 등을 참조하여 추출하였다.

가. 표제 추출

표제는 표목 필드인 통일표제(130/630/730/930), 표제와 표제관련필드인 통일표제(240), 종합통일표제(243)의 $a(통일표제)에서 우선 추출하고, 번역서인 경우 여러 형태의 표제(246) 필드에서 원표제를 함께 추출하였다. 또한 통일표제와 원표제가 미기입된 경우가 많으므로 표제와 책임표시사항(245)의 $a(본표제)와 $x(대등표제), 여러 형태의 표제(246)에서 제2지시기호가 ♭(표출어를 생성하지 않음)가 아닌 경우의 $a(본표제/간략표제)도 추출하였다. 추가적으로 표제에서 관제로 인하여 동일한 저작으로 묶이지 않는 상황을 방지하기 위해 지시기호를 참조하여 관제가 있는 경우('1' 또는 '2')는 이를 제거하였다.

나. 저자명 추출

저자명은 기본표목(1XX)과 부출표목(7XX), 표제와 책임표시사항(245)에서 추출하였다. 또한, 국립중앙도서관이 로컬표목(9XX)에 저자명의 이형 표현을 할당하고 있으므로 이를 함께 추출하였다(〈표 4〉 참조). 저자명의 이형 표현은 필명으로 더욱 유명한 경우나 이름 대신 호(號)와 같이 특수한 저자명으로 입력되어 있는 다양한 레코드 간의 연결고리로 작용할 수 있다. 예를 들어, 〈표 5〉의 예시 레코드들은 같은 저작임에도 상이한 저자명 즉, '김소월'과 '金廷湜'(김정식)으로 입력되어 있어 군집화될 수 없다. 이에 로컬표목(900) 필드의 이형 표현을 활용하여 이들 레코드를 연결하여 동일 저작으로 연결하였다.

〈표 4〉 개인명 추출 예시

필드	지시기호 1	지시기호 2	데이터
100	1		$aWerber, Bernard,$d1961-$0KAC199629675
700	1		$a나쓰메 소세키,$d1867-1916$0KAC199632386$4aut
900	1		$a하목수석,$g夏目漱石,$d1867-1916

〈표 5〉 로컬표목 활용 예시

예시 레코드 1			
태그	식별기호		필드
001			KMO200821373
			:
245	0	0	$a진달래꽃 :$b소월 시전집 /$d김정식 저 :$e이억영 화
			:
700	1		$a김정식
			:
900	0	0	$a이억영
900	1	0	$a김소월
예시 레코드 2			
태그	식별기호		필드
001			KMO201767733
			:
245	0	0	$a진달래꽃 :$b한국명시 /$d金廷湜 著
			:
700	1		$a김소월,$g金素月,$d1902-19340KAC2011019824aut
			:
900	0	0	$a소월,$g素月,d1902-1934
900	1	0	$a김정식,$g金廷湜,d1902-1934

표목의 개인명(100, 700, 900) 필드에서는 $a(개인명)와 $d(생몰년)를 주로 활용하였으며, 부가적으로 $e(역할어), $4(관계) 값을 확인하여 저자 여부(예: $4의 값이 'aut'일 경우 저자)를 식별할 수 있도록 하였다. 개인명 이외의 단체명, 회의명도 같은 방식으로 추출하였다. 저자명을 추출할 때 위에서 언급한 모든 필드의 값을 수집하는 것은 아니며, 245 필드의 책임표시 사항 입력 내용을 기준으로 표목(1XX, 7XX, 9XX)의 저자명을 추출하였으며, 이에 대한 구체적인 방식은 다음과 같다.

첫째, 245 필드에 첫 번째 책임표시만이 입력된 경우 별도의 처리 과정 없이 표목의 저자명을 모두 추출하였다. 이 경우 두 번째 이하 책임표시가 존재하지 않으므로 표목(1XX, 7XX, 9XX)에 이형 표현이 입력되어 있더라도 동일한 저자를 의미하기 때문이다.

둘째, 245 필드에 책임표시가 둘 이상 있는 경우 표목의 저자명(1XX, 7XX, 9XX의 $a)은

첫 번째 책임표시(＄d)와 비교하여 일치하는 것만 추출하였다.

셋째, 서양인명의 경우, 245 필드에 기입된 저자명이 표목 필드에 입력된 표기와 다를 수 있음을 고려하여, 지시기호를 확인하고 성으로 시작하는 이름은 '이름+성'으로 순서를 변환한 표기(예: 'Werber, Bernard'의 경우 'Bernard Werber')도 함께 추출하였다(〈표 6〉 참조).

〈표 6〉 성과 이름 순서 변환 예시 (서양인명)

필드	지시기호 1	지시기호 2	데이터
245	0	0	a(聖約)면죄부/d퀸스 S. 어메랑스 지음
700	1		a어메랑스, 퀸스 S.

변환 전	변환 후
a어메랑스, 퀸스 S.	퀸스 S., 어메랑스

또한 245 필드의 책임표시는 이름, 역할, 생몰년 등의 항목이 명확히 구분되지 않은 채 '개인/단체/회의명+역할어', '역할어+개인/단체/회의명', '개인/단체/회의명' 등의 다양한 형식으로 작성된다. 이에 저작 세트를 생성하였을 때 표목에 작성된 저자명과 함께 군집화될 수 있도록 245 필드에 작성된 책임표시를 표목 필드에서 추출된 명칭('성+이름' 순서 변경 포함)과 비교하여 역할어를 제외하고 저자명만 남기는 과정을 추가하였다(〈표 7〉 참조). 이때 제외한 역할어는 별도로 수집하여 사전을 구축하였으며, 해당 사전을 활용하여 역할어를 한 번 더 제거하는 과정을 추가하였다.

〈표 7〉 역할어 제거 및 사전 구축 예시

필드	지시기호 1	지시기호 2	데이터
245	0	0	a(趙廷來 大河小說)太白山脈/d趙廷來 著
700	1		a조정래

매칭 전	매칭 후
• 趙廷來 著 • 조정래	• 趙廷來/조정래[매칭] • 著는 역할어사전에 저장

한편 동양 인명의 경우 한문으로 작성된 사례가 존재하므로 이를 한글로 변환하는 작업을 수행하였다. 한문의 한글 변환 과정에는 한문의 한글 음이 필요하므로 '대한민국법원'과 'Unicode Inc.'로부터 인명용 한문과 유니코드 정보를 수집하여 한문과 그 한글 음의 정보를 수집하였다. 일부 서지데이터는 간체자가 활용되었으므로, 이를 포함한 총 11,000여 자를 수집 및 활용하였다. 변환 과정은 한문의 음이 여러 개 존재하는 상황(〈표 8〉 참조), 두음법칙으로 인해 음이 변하는 상황(〈표 9〉 참조), 간체자가 사용된 상황(〈표 10〉 참조)을 모두 고려하였다.

〈표 8〉 한문의 한글 변환 예시 1 (복수 음)

필드	지시기호 1	지시기호 2	데이터
245	0	0	a金太子傳/d金起東
700	1		a김기동
변환 전			변환 후
• 金起東			• 금기동 • 김기동[매칭]

〈표 9〉 한문의 한글 변환 예시 2 (간체자)

필드	지시기호 1	지시기호 2	데이터
245	0	0	a爸爸比花漂亮 /d赵恩美 著
700	1		a조은미
변환 전			변환 후
• 赵恩美			• 赵은미(번체자만 사용한 경우) • 조은미[매칭]

〈표 10〉 한문의 한글 변환 예시 3 (두음법칙)

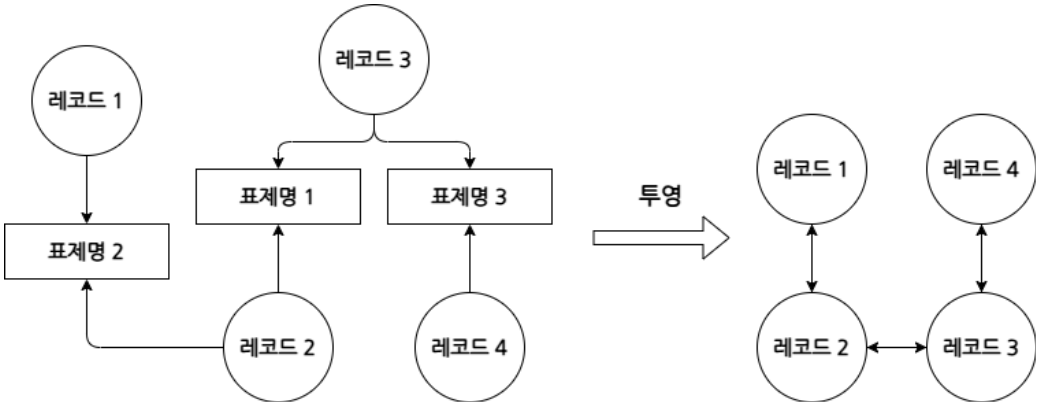
필드	지시기호 1	지시기호 2	데이터
245	0	0	a亂中日記 /d李舜臣 著
700	1		a이순신
변환 전			변환 후
• 李舜臣			• 李舜臣 • 리순신(두음법칙 적용하지 않은 경우) • 이순신[매칭]

3. 네트워크화 및 저작 식별

서지레코드에서 추출한 표제와 저자명을 기반으로 네트워크 분석 기법을 활용하여 저작을 군집화하였다. 네트워크 분석 기법은 노드와 엣지(관계)를 활용하여 객체 간 관계를 구조화하는 방법이다. 본 연구에서는 각 서지레코드(식별자)를 노드로 설정하고, 동일한 표제나 저자명을 가진 레코드들과의 엣지를 생성하였다.

네트워크의 경우, 레코드 식별자와 표제 또는 저자명이 연결된 이분 그래프(bipartite graph) 형태로 '저자명-레코드 네트워크'와 '표제-레코드 네트워크'를 생성하였다. 각 이분 그래프에서는 표제 또는 저자명이 레코드와 연결되며, 이를 기반으로 레코드 간에 간접적인 연결도 가능하다. 이후, 이를 투영(project)하여 레코드 간에 직접적인 연결을 형성하였다. 이때 '투영'이란 표제 또는 저자명을 공유하는 레코드들이 간접적으로 연결된 경우, 이를 직접 연결된 네트워크로 변환

하는 과정을 의미한다(〈그림 1〉 참조). 결과적으로, 동일한 표제 또는 저자명을 가진 레코드는 직접 연결되며, 이형 표현을 활용하여 다른 레코드를 거쳐 간접적으로 연결될 수 있다. 즉, 동일 혹은 이형 표현을 가진 표제/저자명 레코드들을 하나의 집합으로 모을 수 있도록 하였다.



〈그림 1〉 이분 그래프의 투영 과정

이분 그래프를 투영하여 구축한 '저자명-레코드 네트워크'와 '표제-레코드 네트워크'를 생성함으로써 각각 동일 저자와 표제에 대한 이형 표현을 쉽게 식별하고 비교할 수 있도록 하였다. 이와 같은 네트워크를 생성하면 다수의 레코드 간 연결성을 쉽게 파악할 수 있는데, 특히 직접적으로 연결되지 않은 레코드 간(예: 〈그림 1〉의 레코드 1과 4의 관계)에도 연결이 가능한 장점이 있다.

'표제-레코드 네트워크'와 '저자명-레코드 네트워크'는 각각 존재하며, 이들 두 네트워크를 대조하여 레코드 간 교집합(중첩)을 식별하는 과정을 거치면 동일한 저자와 표제를 공유하는, 즉 하나의 저작에 대한 관련 레코드를 식별할 수 있다. 이러한 방식을 활용하여 텍스트 완전 일치를 기준으로 저작을 식별할 때 발생하는 동일 저작이 다른 저작으로 구분되는 문제를 보완하였다. 최종 식별된 저작 군집의 명칭은 출현 빈도가 가장 많은 표제와 저자명을 조합하여 생성하였다.

Ⅳ. 연구 결과

1. 저자명 및 표제 식별 결과

저자명은 전체 레코드의 표목 필드(X00/X10/X11)(709,686건)에서 450,665건, 표제와 책임표시사항(245) 필드 417,886건에서 478,605건을 추출하였다. 이때 전체 레코드 수보다 60,719건 더 많은

저자명이 추출된 이유는 한문에 대한 한글 변환 표기를 생성하였기 때문이다. 표제는 전체 레코드에서 501,480건을 추출하였으며, 이 중 494,870건은 245 필드에서 추출하였다. 전체 레코드 수보다 많은 표제가 추출된 이유는 대등표제도 함께 추출하였기 때문이다.

위 작업을 통해 110,315건의 고유한 저자명과 294,611건의 고유한 표제를 추출하였다. 레코드 수 기준으로 상위 20개의 저자명과 표제는 아래 〈표 11〉 및 〈표 12〉와 같다. 각 저자명과 표제는 다른 언어 또는 표현 방식으로 작성된 다양한 이형 표현이 존재하는 것을 확인할 수 있었다.

〈표 11〉 상위 20위 저자명 (레코드 수 기준)

순위	저자명	레코드 수	레코드 내 출현 이형 표현	순위	저자명	레코드 수	레코드 내 출현 이형 표현
1	공지영	1,678	• Gong, Ji-yeong • Ji-yeong Gong • 孔枝泳	11	최인호	652	• 崔仁浩 • 최인 • 최정암 외 2개
2	박완서	1,299	• 朴婉緖 • 朴婉绪 외 7개	12	최도열	639	-
3	이문열	9,29	• 李文烈 • Yi, Mun-yeol • Lee, Mun-yeol • 이열 외 8개	13	박경리	584	• 朴景利 • 박금이 외 4개
4	황석영	830	• 黃晳暎 • Hwang, Seokyeong • Hwang, Sukyoug • 황수영 외 9개	14	김훈	579	• Gim, Hun • 金薰 외 7개
5	이외수	815	• 李外秀 • 李外水 • Yi, Oe-su 외 5개	15	고정욱	563	• 高廷旭 • Ko, Jeong-uk 외 5개
6	한비야	763	• 韩飞野	16	황성	561	• 전병문 외 4개
7	사마달	709	• 신동욱 • 司馬達 • Samadal 외 2개	17	황선미	550	• 黃善美 • ファン・ソンミ • Hwang, Seonmi 외 7개
8	김진명	696	• Gim, Jin-myeong • Kim, Jin-myeong • 金辰明 외 6개	18	이청준	547	• Yi, Cheong-jun • Lee, Cheong-jun • 李清俊 외 9개
9	조정래	696	• 趙廷來 • Jo, Jeong-rae • Cho, Jung-rae 외 5개	19	유안진	540	• 柳岸津 • 류안진 • Yu, Anjin 외 6개
10	신경숙	692	• Sin, Gyeong-suk • Shin, Kyoung-sook • 申京淑 외 7개	20	이광수	539	• 李光洙 • 춘원 • Yi, Gwang-su 외 14개

다만, 추출한 표제와 저자명 중 동일한 저작임에도 불구하고 다른 저작으로 구분되는 것으로 보이는 사례도 확인되었다. 이러한 사례로는 표제에 대괄호로 자료유형 표기(예: '[디지털]',

‘[디]’)가 작성되어 서로 다른 표제로 구분되는 것과 저자 역할어 표기가 다양하여 이들이 완전하게 제외되지 않는 경우(예: ‘외 1인’)가 존재했다. 또한 〈표 12〉의 ‘연탄길’ 표제의 이형 표현 ‘가슴 찡한 우리 이웃들의 이야기’는 245 필드에서 본표제와 상이한 언어 또는 문자를 기재하는 $x(대등표제)에 입력된 것으로 나타났다. 이는 식별기호에 적절하지 않은 내용이 입력된 사례로 저작을 식별하는 과정에 문제를 야기할 수 있으며, 저작 식별에 필요한 KORMARC 필드의 입력 품질 향상이 필요함을 시사한다.

〈표 12〉 상위 20위 표제 (레코드 수 기준)

순위	표제	레코드 수	레코드 내 출현 이형 표현
1	연탄길	280	• (만화)연탄길 • (감성에세이) 연탄길 • 煤炭路 • 가슴 찡한 우리 이웃들의 이야기 외 2개
2	한국전래동화	256	• (에니메이션)한국전래동화 • (the)goddess of farming, chachongbi • king haeburu, the founder of puyo 외 43개
3	maple story	254	• 메이플스토리 • (코믹) 메이플스토리 • 메이플 스토리 외 26개
4	멈추면, 비로소 보이는 것들	251	• (the) things you can see only when you slow down • 人生那么长, 停一下又何妨 • (the) things we can see only after we stop
5	가시고기	244	• (조창인 장편소설)가시고기 • グッドライフ • 刺鱼 외 3개
6	메이플스토리	241	• maple story • (코믹) 메이플스토리 • (판타지 동화) 메이플스토리 외 1개
7	(코믹) 메이플스토리	239	• 메이플스토리 • maple story
8	아버지	238	• (姜龍俊小說集)아버지 • (나의 삶 나의)아버지 • (만화)아버지 외 6개
9	문학	234	• (고등학교)문학 • (우공비 문해력) 문학 • (윤희재 전공국어) 문학 외 13개
10	지도 밖으로 행군하라	229	• marching off the map • (어린이를 위한) 지도 밖으로 행군하라 • 永不说放弃
11	marching off the map	229	• 지도 밖으로 행군하라 • [디지털대상] 지도 밖으로 행군하라 • 永不说放弃

순위	표제	레코드 수	레코드 내 출현 이형 표현
12	그 많던 싱아는 누가 다 먹었을까	220	-
13	홍길동전	219	• (우리고전)홍길동전 • (the) story of hong gil dong • 洪吉童傳 외 21개
14	쾡이부리말 아이들	219	-
15	상록수	219	• 상록수(常綠樹) • (중학생이 보는)상록수 • (심훈 장편소설)상록수 외 8개
16	구운몽	216	• (우리고전)구운몽 • 구운몽(九雲夢) • (김만중이 들려주는)구운몽 외 22개
17	아홉살 인생	210	• (위기철 소설)아홉살 인생 • 9歳の人生 • when turned nine 외 3개
18	아리랑	209	• (趙廷來 大河小說)아리랑 • (project)아리랑 외 5개
19	(고등학교)문학	196	• 문학
20	(조창인 장편소설) 가시고기	195	• 가시고기

2. 저자 역할어 추출 결과

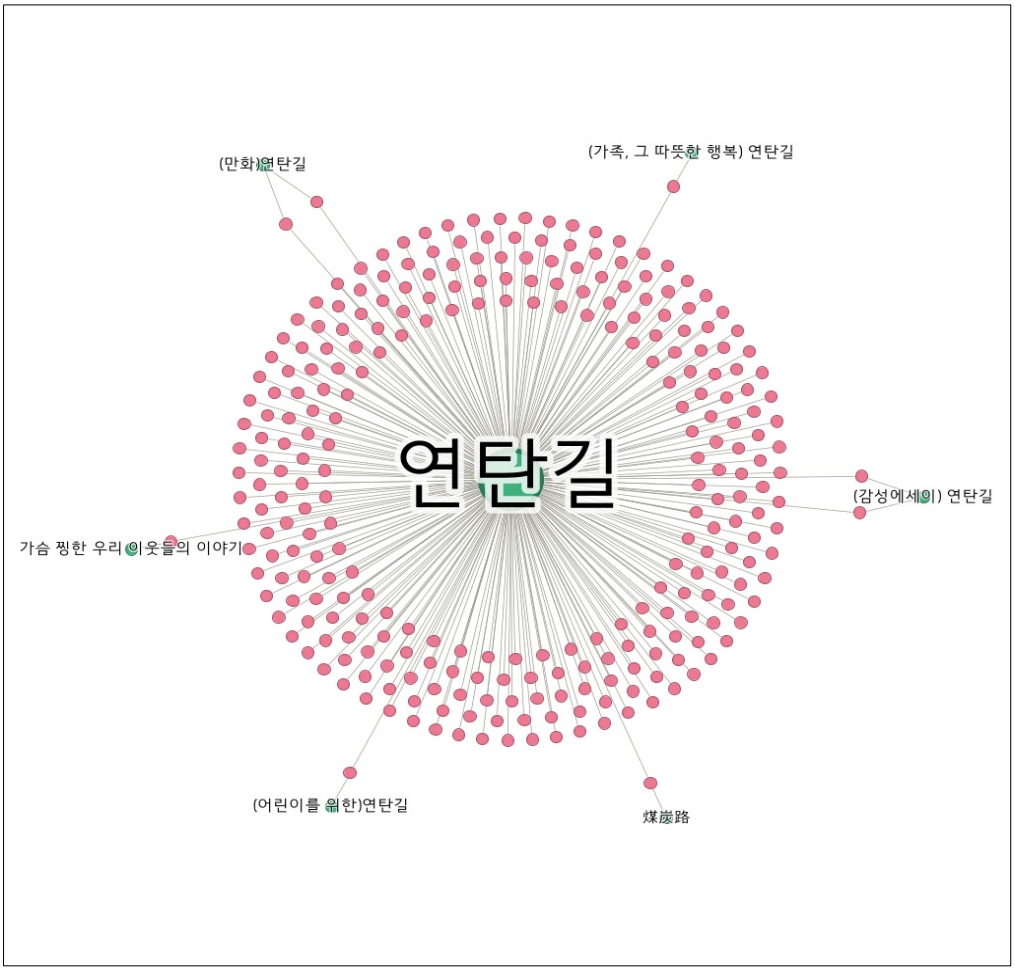
저자명 추출 시 역할어를 제외하는 과정을 포함하였음에도 일부 레코드에서 저자명과 역할어가 함께 추출되었다. 이에 역할어 자동 제거 과정에서 구축한 사전을 기반으로 이를 제거하는 작업을 추가적으로 수행하였다. 역할어를 종합하여 구축한 사전은 〈표 13〉과 같다. 저자 관련 역할어(예: 지음, 글씀)가 가장 빈번하게 사용되었으며, 그 중에서도 '지음'과 '지은이'가 자주 활용되고 있다. 식별된 역할어 중에는 한글이 아닌 언어로 작성된 사례 또한 존재하였다.

〈표 13〉 저자 역할어 추출 및 분류 결과

구분	역할어
저자	글, 글쓴이, 글쓴 사람, 글씀, 원작/原作, 쓰다, 씀, 작/作, 저/著, 작가, 저자/著者, 저작/著作, 지음, 지은이, 집필, 찬/撰 author, by, writer, written by, เขียน, tác giả, зохиолч 등
편집/출판	엮음, 엮은곳, 엮은이, 탈초, 펴냄, 펴낸이, 편/編, 편역/編譯, 편자/編者, 편저/編著, 편저자/編著者, 편집/編輯, 편집인/編輯人, 편찬/編纂, ed. by, edited by 등
번역	국역/國譯, 번역/飜譯, 번안, 역/譯, 역자/譯者, 역해/譯解, 옮긴이, 옮김, 평역/評譯, translated by 등
그림/사진	그린이, 그림, 극화, 동화, 사진, 삽화, 만화, 일러스트레이션/Illustration, 일러스트, 작화/作畵, 카툰, illustrated by 등
주석/해	교주/校注, 주석/注釋, 주해/註解, 해제, 해설/解說, 해설자 등
시	낭송, 동시, 동시집, 시/詩, 시인/詩人, 산문, 운문, poems by, poems of 등
극/희곡	구연, 각색, 각본/脚本, 극본, 극본작가, 기획, 대본, 시나리오, 윤색, 연출, story, story by 등

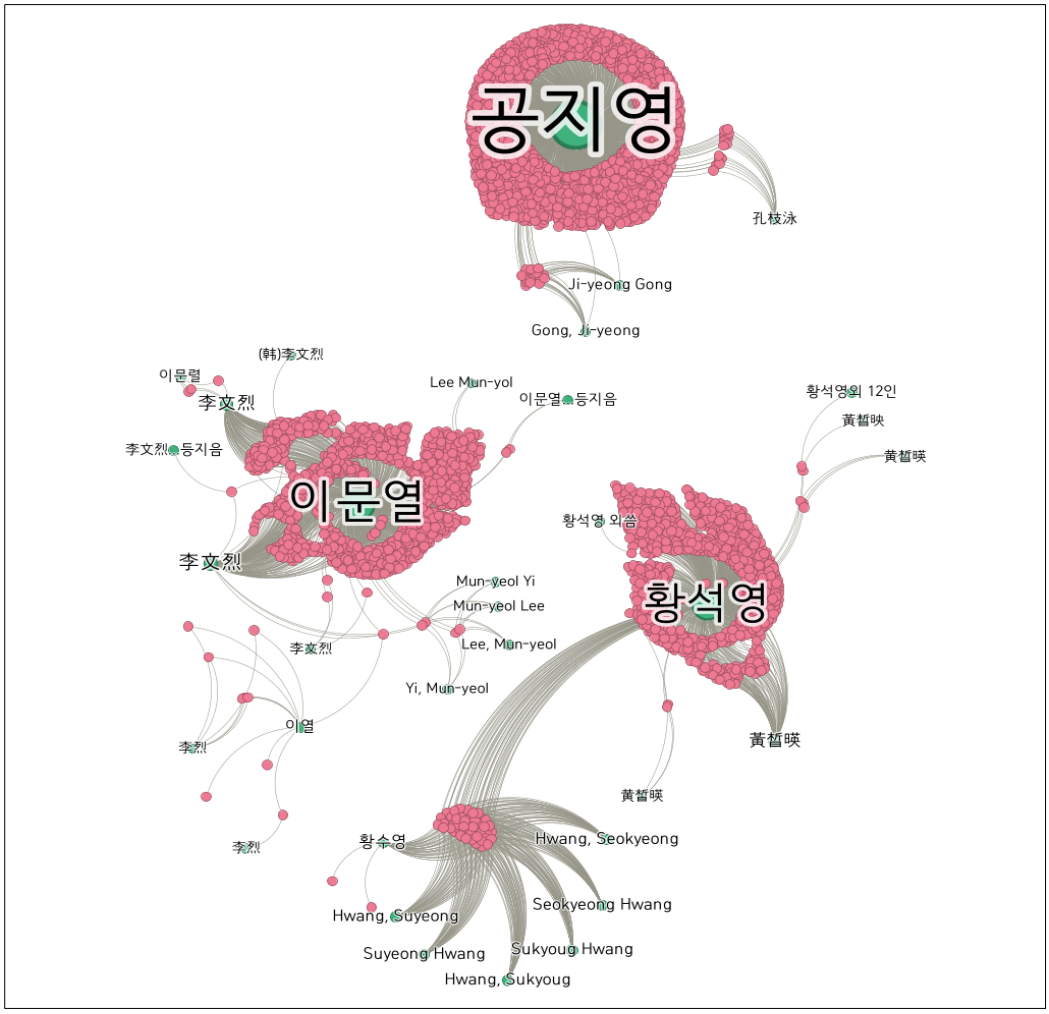
3. 저작 네트워크 생성

추출한 저자명과 표제를 활용하여 각각 이분 그래프를 생성하였다. '표제-레코드 네트워크'를 생성한 결과, 총 417,886건의 레코드(노드)가 2,090,351개의 관계(엣지)로 연결된 네트워크가 생성되었다. 동일 표제를 공유하는 레코드 집합은 242,503개이며, 그 가운데 58,550개 집합에 2건 이상의 레코드가 연결된 것으로 나타났다. '표제-레코드 네트워크'의 경우, 대체로 국문, 영문, 한문으로 작성된 표제를 가진 레코드들이 묶여 하나의 군집을 이룬 것을 확인할 수 있었다. 예를 들어, 가장 많은 레코드가 연결된 표제인 '연탄길'은 '가슴 찡한 우리 이웃들의 이야기', '煤炭路'와 연결되어 군집을 이룬 것을 확인할 수 있다(〈그림 2〉 참조).



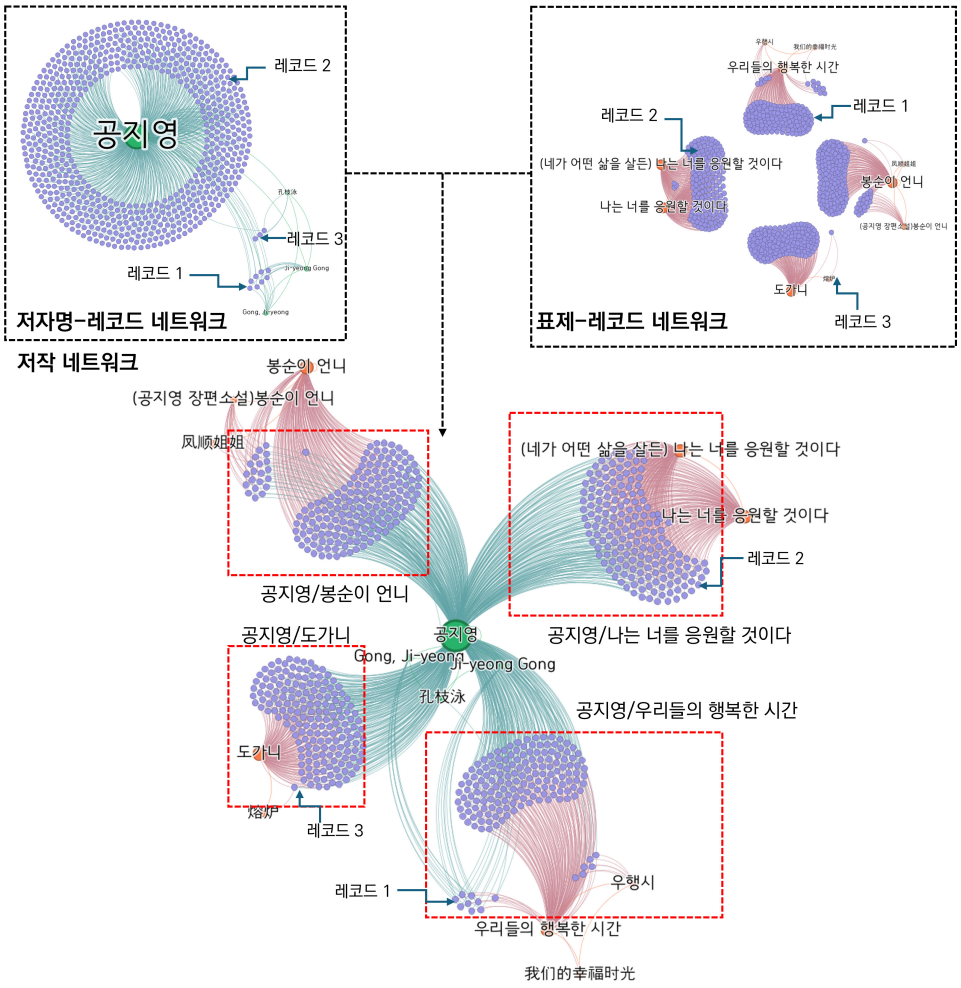
〈그림 2〉 표제-레코드 네트워크 예시

다음으로 '저자명-레코드 네트워크'를 생성한 결과 417,886건의 레코드(노드)가 18,858,526건의 관계(링크)로 연결되었다. 저자명 군집으로 간주되는 레코드 집합은 64,857개 식별되었으며, 그 가운데 33,747개 군집이 2건 이상의 레코드와 연결된 것으로 나타났다. 저자명은 추출 과정에서 '한문의 한글 변환', '성+이름' 순서 변경 등의 작업으로 인하여 이형 표현이 다양하게 생성되었다. 이에 '표제-래코드 네트워크'에 비해 레코드 간 연결이 쉬워지면서 2개 이상 연결된 레코드 집합의 수가 많이 나타났다. 또한, 저자의 영문명, 한문명이 함께 묶여서 나타남을 확인할 수 있었다. 가장 많은 레코드가 연결된 저자 '공지영'은 '孔枝泳', 'Gong, Ji-yeong', 'Ji-yeong Gong'과 연결되었다(〈그림 3〉 참조).



〈그림 3〉 저자명-레코드 네트워크 예시

다음 단계로 '표제-레코드 네트워크'와 '저자명-레코드 네트워크' 간 중복되는 레코드 집합을 식별하였다(〈그림 4〉 참조). 〈그림 4〉의 예시와 같이 저자 '공지영'을 공유하는 네트워크와 각 표제를 공유하는 네트워크에서 중복되는 레코드를 식별하여 '저작 네트워크'를 구성하였으며, 이를 기반으로 저작을 식별할 수 있다. 그 결과 총 417,886개 레코드(노드)가 1,543,613개의 관계(엣지)로 연결된 네트워크가 생성되었다. 동일 저작으로 간주할 수 있는 레코드 집합은 268,684개 식별되었으며, 이는 곧 417,886건의 레코드가 268,684개 저작으로 군집화되었음을 의미한다. 식별된 저작 중 2건 이상의 레코드를 하나의 저작으로 집중시킨 사례 즉, 군집 내 2건 이상의 레코드를 포함하고 있는 저작은 56,749개인 것으로 나타났다.



〈그림 4〉 저작 네트워크 예시

집합 내 레코드 수 기준 상위 20개 저작을 순서대로 식별한 결과는 〈표 14〉와 같다. 가장 많은 레코드가 군집화된 저작은 '이철환/연탄길'로 280개의 레코드가 하나의 저작으로 군집화되었음을 확인할 수 있다. 해당 군집 내 레코드는 대부분 국문 문자 자료였으며, 이외에 만화 자료, 한국어 원작의 중국어 번역본, 녹음자료 등이 함께 군집화되어 있었다. '조창인/가시고기'의 경우 '가시고기'라는 원표제 및 중국어 번역본 표제('刺鱼') 외에도 일어 제목('グッドライフ')이 포함된 것을 확인할 수 있는데, 이는 표제 추출 과정에서 여러 형태의 표제 필드(246) 필드를 포함한 결과 함께 군집화된 것이다. 또한 로컬표목을 활용한 결과, 특정 작가의 호와 다른 이름을 함께 군집화할 수 있었다(예: 심훈, 염상섭). '김만중/구운몽'의 경우 구운몽과 구운몽을 토대로 작성한 에세이, 비평 작품도 함께 식별되는 것을 확인할 수 있었다.

〈표 14〉 상위 20개 저작 식별 결과

순위	저작	주요 저자명	주요 표제	레코드 건 수
1	이철환 /연탄길	• 이철환 • 李喆奐	• 연탄길 • 煤炭路 • 가슴 찡한 우리 이웃들의 이야기 등	280
2	송도수 /메이플스토리	• 송도수	• (코믹) 메이플스토리 • 메이플스토리 • maple story 등	252
3	혜민 /멈추면, 비로소 보이는 것들	• 혜민 • 慧敏	• 멈추면, 비로소 보이는 것들 • 人生那么长, 停一下又何妨 • (the) things we can see only after we stop	249
4	조창인 /가시고기	• 조창인 • チョ・チャンイン • 赵昌仁	• 가시고기 • グッドライフ • good life • 刺鱼	243
5	한비야 /지도 밖으로 행군하라	• 한비야 • 韩飞野	• 지도 밖으로 행군하라 • marching off the map • 永不说放弃 등	241
6	박경리 /토지	• 박경리 • 朴景利	• 토지 • 土地	237
7	박완서 /그 많던 싱아는 누가 다 먹었을까	• 박완서	• 그 많던 싱아는 누가 다 먹었을까	220
8	김중미 /괭이부리말 아이들	• 김중미	• 괭이부리말 아이들	219
9	심훈 /상록수	• 심훈 • 沈熏 • 심대섭 • Sim, Hun • 해풍	• 상록수 • 常緑樹	216
10	위기철 /아홉살 인생	• 위기철 • ウィ ギチョル • 卫奇哲	• 아홉살 인생 • 9歳の人生 • when turned nine • 九岁人生	206

순위	저작	주요 저자명	주요 표제	레코드 건 수
11	김정현 /아버지	• 김정현	• 아버지	190
12	공지영 /봉순이 언니	• 공지영 • 孔枝泳	• 봉순이 언니 • 凤顺姐姐	187
13	공지영 /도가니	• 공지영 • 孔枝泳	• 도가니 • 熔炉	176
14	공지영 /나는 너를 응원할 것이다	• 공지영	• 나는 너를 응원할 것이다	176
15	한비야 /그건사랑이었네	• 한비야	• 그건, 사랑이었네	174
16	공지영 /우리들의 행복한 시간	• 공지영 • 孔枝泳 • Gong, Ji-yeong	• 우리들의 행복한 시간 • 我们的幸福时光 • 우행시	170
17	염상섭 /삼대	• 염상섭 • 廉想涉 • 횡보	• 삼대 • 三代	169
18	김진명 /무궁화꽃이 피었습니다	• 김진명 • Gim, Jin-myeong	• 무궁화꽃이 피었습니다	167
19	김만중 /구운몽	• 김만중 • 金万重 • 金萬重	• 구운몽(九雲夢) • "kim man jungs the cloud dream of the nine : a modern korean translation with a critical essay" • "kim man jungs the cloud dream of the nine : the original text with exgetical notes and a biographical essay" • "the story of guunmong : korean classic rewritten by kang won-hee, writer of childrens books"	165
20	조정래 /太白山脈	• 조정래 • 趙廷來	• 太白山脈	163

Ⅴ. 결 론

영미권의 목록에서는 전거데이터 구축이 잘 이루어져 있는 반면에 국내 목록의 경우 기본표목이 필수적으로 입력되지 않으며, 전거통제 또한 미흡한 실정이다. 이에 국내 목록 환경에 적합한 서지개체 식별 과정이 필요하다는 것은 기존 연구를 통해 입증되었으며, 관련 연구가 지속적으로 이루어져 왔다. 그러나 대규모 서지레코드를 대상으로 저작 식별을 위한 핵심 정보(표제, 저자명)를 수집하고 정제하는 과정에 대한 연구는 상대적으로 부족하며, 저작 군집화 방식에서도 텍스트의 완전 일치를 기준을 적용함에 따라 필드 입력 방식(언어, 관제, 이형 표현, 입력 순서 등)의 차이로 인해 동일 저작이 서로 다른 저작으로 분리되는 문제가 발생하였다.

이에 본 연구에서는 표제 및 저자명 추출과 전처리 과정을 개선하는 방안을 제안하고, 이를 기반으로 표제와 저자명 기반의 네트워크를 구축하여 저작을 식별하는 방법을 제안하였다. 이러

한 방법을 국립중앙도서관의 한국문학 분야 KORMARC 서지레코드 453,846건에 적용한 결과, 기존 연구에서 나타난 저작 식별의 한계점을 보완할 수 있음을 확인하였다.

연구의 주요 내용을 살펴보면, 먼저 표제 추출 시 관제를 제거하는 단계를 포함하였고, 한문이 포함된 사례를 다수 확인됨에 따라 이를 한글로 변환한 후 함께 추출할 수 있도록 하였다. 저자명 추출 과정에서는 서양인명의 '성-이름' 순서 변환, 한문 입력값의 한글 변환을 처리할 수 있도록 설계하였다. 또한, 저자 역할어를 식별하여 사전을 구축하고, 이를 제거함으로써 저자명의 식별성을 향상시켰다. 이러한 전처리 과정을 실제 서지레코드에 적용한 결과, 저자명 표기 순서, 작성 언어의 차이, 저자 역할어, 표제내 관제 등의 요인으로 인해 동일 저작이 상이한 저작으로 분리되는 문제를 방지할 수 있음을 확인할 수 있었다.

네트워크 구축 과정에서는 이분 그래프 기반의 투영 네트워크 분석 기법을 적용하였다. 이를 통해 특정 저자명 및 표제의 이형 표현을 효과적으로 파악하고, 직접적인 연결이 없는 경우에도 동일한 저작 집합으로 포함될 수 있도록 하여 저작 식별 정확도를 높였다. 그 결과, 표제 또는 저자명이 완전히 일치하지 않더라도 연결을 공유하는 노드(서지레코드)가 존재하는 경우 동일한 저작으로 식별할 수 있었다. 다시 말해, 표제 및 저자명이 서로 다른 언어나 표기 방식으로 입력된 경우에도 하나의 군집으로 묶을 수 있어 저작 식별의 완성도를 높일 수 있었다. 최종적으로 417,886건의 서지레코드를 268,684개 저작으로 군집화하였다. 이와 같이 서지레코드에서 저작 관련 정보를 추출하고 정제하는 일련의 과정과 네트워크화 기법을 적용한 저작 식별 과정은 국내 목록의 서지개체 식별 과정에 활용될 수 있으며, 궁극적으로 링크드데이터 변환에도 기여할 수 있을 것으로 기대된다.

다만, 본 연구에서는 전집과 총서와 같이 다수 저작을 포함하는 일부 레코드를 분석에서 제외하였으며, 표제에 자료유형이 포함된 경우 저작 식별이 원활하게 수행되지 않는 한계가 있다. 또한, KORMARC 레코드의 필드, 지시기호, 식별기호 등의 값이 올바르게 입력되지 않은 사례도 존재하는 것으로 나타났다. 따라서 후속 연구에서는 저작 식별을 위한 KORMARC 레코드 개선 방안을 모색하고, 동일한 저작임에도 부분적인 표기 방식의 차이로 인해 상이한 저작으로 구분되는 문제를 해결하는 연구가 필요할 것이다.

참 고 문 헌

가족관계의 등록 등에 관한 규칙. 대법원규칙 제3140호.

김정현 (2007). 한국어 서지레코드에 있어 FRBR 모형의 유용성에 관한 연구. 한국문헌정보학회지, 41(4), 295-314.

김정현 (2015). 사서오경의 서지적 관계 특성에 따른 FRBR 적용에 관한 연구. 한국도서관·정보학회지, 46(2), 317-336. https://doi.org/10.16981/kliss.46.2.201506.317

김정현, 이성숙, 이유정 (2015). KORMARC 서지레코드의 FRBR 알고리즘 개발에 관한 연구. 한국도서관·정보학회지, 46(1), 1-23. https://doi.org/10.16981/kliss.46.1.201503.1

김현희, 유영준, 박서은 (2007). FRBR 모형의 KORMARC 데이터베이스로의 적용 가능성에 대한 실험적 연구: 음악자료를 중심으로. 한국도서관·정보학회지, 38(2), 185-202. https://doi.org/10.16981/kliss.38.2.200706.185

노지현 (2008). KORMARC 레코드에 대한 FRBR 모델의 적용 실험: 국립중앙도서관 서지레코드를 사례로 하여. 한국도서관·정보학회지, 39(2), 291-312. https://doi.org/10.16981/kliss.39.2.200806.291

노지현, 이미화, 이은주 (2023). 목록이론의 이해와 적용. 서울: 한국도서관협회.

윤재혁, 도슬기, 오삼균 (2020). 저자역할용어사전 구축 및 저작군집화에 관한 연구. 정보관리학회지, 37(2), 197-223. https://doi.org/10.3743/KOSIM.2020.37.2.197

이미화, 이은주, 노지현 (2022). LRM 이후 목록 동향과 KORMARC 통합서지용에서의 수용 방안. 한국비블리아학회지, 33(1), 25-45. https://doi.org/10.14699/kbiblia.2022.33.1.025

이미화, 정연경 (2008). 저작 클러스터링 분석을 통한 FRBR의 목록 적용에 관한 연구. 정보관리학회지, 25(3), 65-82. https://doi.org/10.3743/KOSIM.2008.25.3.065

이성숙, 이현주 (2013). 한국전통음악의 서지적 관계 특성에 따른 FRBR 모형 적용방안. 사회과학연구, 24(2), 399-421. https://doi.org/10.16881/jss.2013.04.24.2.399

이혜원 (2011). 목록의 집중기능을 향상시키는 '원형' 개념에 관한 연구. 한국비블리아학회지, 22(3), 91-107.

조재인 (2004). FRBR 알고리즘 분석 및 KORMARC 데이터베이스 적용 방안. 한국문헌정보학회지, 38(3), 5-21.

Hickey, T. B. & Toves, J. (2005, April). FRBR Work-set Algorithm. Dublin, Ohio: OCLC Online Computer Library Center, Inc. Available: https://www.oclc.org/content/dam/research/activities/frbralgorithm/2005-04.pdf

Hickey, T. B. & Toves, J. (2009, August). FRBR Work-set Algorithm: version 2.0. Dublin, Ohio: OCLC Online Computer Library Center, Inc. Available: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf

Library of Congress (2004). FRBR display tool version 2.0. Available: http://www.loc.gov/marc/marc-functional-analysis/tool.html

Riva, P., Le Boeuf, P., & Žumer, M. (2017). IFLA library reference model: A conceptual

model for bibliographic information. Hague: IFLA

Tillett, B. B. (2001). Relationship in the Organization of Knowledge. Kluwer Academic Publishers.

Unicode, Inc. (2024, July 31) Unicode Han Database (Unihan). Available: https://www.unicode.org/reports/tr38/

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

Cho, Jane (2004). Study on the FRBR algorithm and application of KORMARC database. Journal of the Korean Society for Library and Information Science, 38(3), 5-21.

Kim, Hyun-Hee, Yoo, Yeong-Jun, & Park, Suh-Eun (2007). An experimental study on the FRBR model adaptation to KORMARC database: focusing on music materials. Journal of Korean Library and Information Science Society, 38(2), 185-202. https://doi.org/10.16981/kliss.38.2.200706.185

Kim, Jeong-Hyen (2007). A study on the utility of FRBR model in korean bibliographic record. Journal of the Korean Society for Library and Information Science, 41(4), 295-314.

Kim, Jeong-Hyen (2015). A study on the adoption of the FRBR according to the bibliographic relationships of five classics and four books. Journal of Korean Library and Information Science Society, 46(2), 317-336. https://doi.org/10.16981/kliss.46.2.201506.317

Kim, Jeong-Hyen, Lee, Sung-suk, & Lee, You-Jeong (2015). A study on the development of FRBR algorithm for KORMARC bibliographic record. Journal of Korean Library and Information Science Society, 46(1), 1-23. https://doi.org/10.16981/kliss.46.1.201503.1

Lee, Hyewon (2011). A study on a concept of 'prototype' for enhancing the collocation function of catalog. Journal of the Korean Biblia Society for Library and Information Science, 22(3), 91-107.

Lee, Mihwa & Chung, Yeon-Kyoung (2008). A study of FRBR implementation to catalog by using work clustering. Journal of the Korean Society for Information Management, 25(3), 65-82. https://doi.org/10.3743/KOSIM.2008.25.3.065

Lee, Mihwa, Lee, Eun-Ju, & Rho, Jee-Hyun (2022). Cataloging trends after LRM and

its acceptance in KORMARC bibliographic format. Journal of the Korean Biblia Society for Library and Information Science, 33(1), 25-45. https://doi.org/10.14699/kbiblia.2022.33.1.025

Lee, Sung-Sook & Lee, Hyun-Ju (2013). A study on the adoption of the FRBR model according to the bibliographic relationships of Korean classical music. Journal of Social Science, 24(2), 399-421. https://doi.org/10.16881/jss.2013.04.24.2.399

Regulations of on Registration of Family Relations. Rule No. 3140.

Rho, Jee-Hyun (2008). An application of FRBR model to KORMARC records. Journal of Korean Library and Information Science Society, 39(2), 291-312. https://doi.org/10.16981/kliss.39.2.200806.291

Rho, Jee-Hyun, Lee, Mihwa, & Lee, Eun-Ju (2023). Cataloging Theory and Practice. Seoul: Korean Library Association.

Yun, Jaehyuk, Do, Seulki, & Oh, Sam Gyun (2020). Designing a FRBR work grouping algorithm of bibliographic records using a role term dictionary of authors. Journal of the Korean Society for Information Management, 37(2), 197-223. https://doi.org/10.3743/KOSIM.2020.37.2.197