한국문학 분야 KORMARC 서지레코드의 표현형 식별 연구*

A Study on the Identification of Expression in KORMARC Bibliographic Records in the Field of Korean Literature

나 상 오 (Sangoh Na)**
이 종 욱 (Jongwook Lee)***

목 차

초 록

본 연구는 KCR4 체제하에 구축된 KORMARC 서지레코드로부터 KCR5의 핵심 개체인 표현형(Expression)을 식별하는 방법을 설계하고 적용하였다. 국립중앙도서관의 한국문학 서지레코드 453,846건 가운데 복수의 레코드로 구성된 56,749개 저작, 205,951건을 대상으로 분석한 결과 RDA 수용에 따라 신설된 표현형 관련 필드(336, 377, 381)는 기입률이 저조하여 활용이 어려운 반면, 책임표시사항(245), 부출표목(700, 710), 판사항(250), 언어부호(041) 등 범용 필드에는 표현형을 구분할 단서가 자연어 형태로 존재하였다. 이를 정규표현식과 통제어휘 변환으로 정제하여 내용유형, 언어, 기타식별특성 순으로 적용한 결과 103,783개의 표현형 클러스터가 식별되었으며, 전체 파생의 95.89%가 기타식별특성 단계에서 발생하였다. 이는 기존 대규모 서지데이터를 개체-관계 구조로 전환하기 위한 실증적 방안을 제시하였다는 점에서 의의를 지닌다.

ABSTRACT

This study designed and applied a method for identifying the Expression entity required by KCR5 from KORMARC bibliographic records created under KCR4. Since re-cataloging accumulated large-scale databases under new rules is not feasible, this study utilized values already recorded in legacy data. From 453,846 Korean literature records held by the National Library of Korea, 56,749 works comprising 205,951 records with multiple bibliographic records were analyzed. Field analysis showed that expression-related fields introduced through RDA (336, 377, 381) had extremely low input rates and could not be used independently, whereas general-purpose fields such as 245, 700, 710, 250, and 041 contained natural-language clues for distinguishing expressions. After normalizing these values through regular expressions and controlled-vocabulary mapping, content type, language, and other distinguishing characteristics were applied sequentially, yielding 103,783 expression clusters, with 95.89% of all derivations occurring at the third stage. This study presents an empirical approach for converting large-scale legacy bibliographic data into an entity-relationship structure, and may serve as foundational research for Linked Data implementation and the assignment of expression-level identifiers.

키워드: 서지레코드의 기능상 요건, 표현형 식별, 한국목록규칙 제5판, 한국문헌자동화목록형식
FRBR, Expression Identification, KCR5, KORMARC

---

* 이 논문은 2025학년도 경북대학교 연구년 교수 연구비에 의하여 연구되었음.
** 경북대학교 문헌정보학과 석박사통합과정(gg4518@knu.ac.kr / ISNI 0000 0005 0490 1663) (제1저자)
*** 경북대학교 문헌정보학과 부교수(jongwook@knu.ac.kr / ISNI 0000 0004 6830 6145) (교신저자)
논문접수일자: 2026년 7월 27일 최초심사일자: 2026년 8월 2일 게재확정일자: 2026년 8월 10일
한국문헌정보학회지, 60(3): 443-463, 2026. http://dx.doi.org/10.4275/KSLIS.2026.60.3.443

1. 서 론

서지데이터는 문헌 정보자원의 표제, 저자, 판본, 주제어 등 자원의 고유한 속성을 체계적으로 기술한 메타데이터로 간주된다(Riva et al., 2017). 이는 도서관 환경에서 주로 기계가독형목록(이하 MARC)으로 관리되고 있으며, 필드별로 데이터를 입력하는 구조상 정보자원에 대한 상세한 기술(description)에는 최적화되어 있다(국립중앙도서관, 2023). 하지만 MARC 구조는 서지레코드를 정보자원별로 각각 기술하며 데이터 필드간 연결이 제한적인 관계로 정보자원 간 복잡한 의미적 관계를 표현하거나 다른 정보자원과의 연결에는 한계를 지니고 있다.

이러한 한계를 극복하기 위한 논의는 서지 세계를 개체와 그 관계로 파악하려는 시도로 이어졌다. 국제도서관연맹(이하 IFLA)이 개발한 '서지레코드의 기능상 요건'(이하 FRBR)과 이를 통합한 '도서관 참조 모형'(이하 LRM)이 대표적이며, 이들 개념모형은 국제목록원칙규범의 제정은 물론 영미목록규칙(AACR2)이 자원기술과 접근(이하 RDA)으로 전면 개정되는 데 영향을 주었다. 한편 데이터 기술 프레임워크 측면에서는 단순 문자열 기반의 목록에서 벗어나 데이터 간 의미적 관계를 컴퓨터가 이해하고 연결하는 링크드데이터 체제로의 이행이 진행되었으며, 미국 의회도서관이 개발한 BIBFRAME이 그 대표적 사례에 해당한다(Kroeger, 2013; Kim et al., 2021).

국제적인 서지 기술 환경의 패러다임 변화에 대응하여 한국도서관협회 또한 FRBR 및 RDA를 수용한 '한국목록규칙 제5판'(이하 KCR5)을 제정하였다(한국도서관협회, 2025). KCR5의 핵심은 FR 개념모형에 기반한 규칙 구조를 통해 정보자원을 기술하도록 안내한다는 점이다. 즉, WEMI 구조라고도 불리는 저작(Work), 표현형(Expression), 구현형(Manifestation), 개별자료(Item)라는 네 가지 '개체'에 대하여 독립적으로 기술하며, 이들 간 관계를 조직하는 데 중점을 둔다(한국도서관협회, 2025).

특히, 지적 또는 예술적 창조물인 '저작'과 저작이 기호, 음향, 이미지 등으로 실현된 형태인 '표현형'은 흩어져 있던 다양한 번역본과 판본들을 개체 단위로 집중하여, 이용자의 정보접근성을 향상시킬 수 있다는 점에서 중요하다(이혜원, 2011). 또한 KCR5에 따르면 표현형 속성을 기록하는 목적은 하나의 저작에 대한 둘 이상의 표현형을 구분하여 식별하고, 이용자의 자원 선정 요구에 부응할 수 있도록 하는 것이다. 따라서 국내 도서관 목록 환경에서 정보자원의 저작과 표현형 식별은 IFLA에서 정의하는 이용자 과업 중 식별(identify)과 선정(select)의 효율성을 높이는 데 도움이 될 수 있다.

예를 들어 정보자원의 내용적 측면과 물리적 측면에 대한 정보가 한 데 혼재되어 기술되는 현행 KORMARC 구조에 따르면, 특정 정보자원에 대한 검색결과를 구현형 중심의 개별 서지레코드로서 제공해줄 뿐이지만, 개체 식별이 이루어지면 특정 장르 또는 언어 등에 해당하는 서지레코드를 저작 또는 표현형 단계로 계층화하여 보여주고 비교적 이용자 요구에 적합한 정보자원으로 안내할 수 있게 된다.

이에 FRBR 개체 식별의 중요성을 인식하고 기존 MARC 형태 서지레코드로부터 정보자원의 내용적 측면을 식별하고자 한 연구는 국내

외에서 지속적으로 수행되어왔다. 대표적으로 OCLC와 미의회도서관의 FRBR 적용에 대한 논의가 있으며, 국내에서도 표제와 저자명 등 KORMARC 레코드의 데이터를 기반으로 개체를 식별하려는 노력이 있었다(Hickey & Toves, 2005; 2009; Library of Congress, 2004; 조재인, 2004; 노지현, 2008; 김정현 외, 2015). 다만, 대부분의 연구는 저작 개체의 식별에 집중하고 있으며 부분적으로 표현형을 함께 식별한 연구는 있으나, 표현형 자체의 속성과 식별에 집중한 연구는 드문 실정이다. 그 이유는 과거 KCR4 체제 아래에 작성된 기존 KORMARC 서지레코드에는 표현형 구별에 필요한 핵심 요소(예: 번역자, 언어, 판사항 등)가 자연어 형태로 다양한 필드에 분산되어 기계적인 정보 추출에 어려움이 있기 때문으로 보인다.

이에 본 연구에서는 기존 구현형 중심으로 기술된 KORMARC 서지레코드에 파편화 되어 있는 표현형 속성 관련 필드를 파악하고, 한국문학 분야의 약 45만 건에 달하는 대규모 서지데이터로부터 해당 필드의 현황을 분석한 뒤, 활용하기 적합할 것으로 판단되는 필드를 추출 및 정제하여 표현형 개체를 식별하고자 한다. 이 과정에서 서지레코드의 주제와 요목 수준(문학장르)을 기준으로 그 추이를 살펴보고자 한다. 본 연구의 결과는 새로운 목록규칙 도입으로 서지 기술 방식이 변화하는 시점에, 기존 규칙 기반의 서지레코드를 개체-관계 구조로 변환하는 데 필요한 방안을 모색하는 데 있어 시사점을 제공할 수 있을 것으로 기대된다.

2. 개념적 배경 및 선행연구

2.1 FRBR 모형과 WEMI 구조

현대 서지 기술 환경의 핵심이 되는 개념 모형은 서지레코드의 기능상 요건(Functional Requirements for Bibliographic Records, FRBR)이다. FRBR 모형은 서지 세계를 구성하는 핵심 개체(Entity)를 저작(Work), 표현형(Expression), 구현형(Manifestation), 개별자료(Item)의 4단계(WEMI)로 구분한다. '저작'은 지적 혹은 예술적 창작물로서 추상적인 개념 그 자체를 의미하며, '표현형'은 저작이 문자, 숫자, 음표, 소리 등의 기호를 통해 지각할 수 있는 형태로 실현된 것을 뜻한다. 동일한 저작이라 하더라도 번역, 개정, 편곡, 낭독 등 내용 혹은 형식의 변형이 가해질 경우 새로운 표현형으로 간주된다. '구현형'은 표현형이 물리적 또는 디지털 매체로 구현된 출판물 형태를 의미하며, '개별자료'는 지적 혹은 예술적 내용을 전달하기 위해 신호를 수록한 객체로 구현형의 단일 복사본을 지칭한다. 개체는 관계(Relationship)로 연결되며, WEMI 개체는 저작이 표현형으로 '실현'되고, 표현형이 구현형으로 '구현'되며, 구현형이 개별자료로 사례가 되는 기본 관계를 통하여 위계적으로 연결된다(Riva et al., 2017). 이 외에도 각 개체는 창작, 제작, 배포, 소장 등의 책임관계를 통해 개인, 가계, 단체와 연결되고, 저작 간에는 파생, 전체-부분, 선후 등의 서지적 관계가 설정된다.

이 가운데 표현형은 WEMI 구조 내에서 저작의 실질적 변이를 반영하는 핵심 개체로 볼 수 있다(이미화, 2019). 동일한 저작이더라도

번역, 개작, 개정 등의 과정이 진행되면 새로운 표현형으로 간주되며, 동일 저작 내 표현형을 적절히 분리하였을 때 정보자원은 탐색하기 더욱 적합한 형태로 계층화된다. 예를 들어, 하나의 외국 문학 저작에 대해 복수의 번역본이 출판된 경우, 번역자에 따라 표현형을 구분하여 제시하면 이용자는 단순히 동일 표제의 자료가 나열된 결과를 훑는 대신 특정 번역본을 지정하여 선택할 수 있다. 반면, 물리적 매체의 차이는 표현형이 아닌 구현형의 변화로 구분되는데, 기존 MARC 구조에는 이러한 개체별 속성이 명확히 분리되어 있지 않은 채 다양한 필드에 여러 개체의 속성이 함께 기술되어 있다.

한편 표현형을 어떤 속성으로 기술할 것인가에 대한 규정은 개념모형에 따라 변화해 왔다. FRBR은 표현형의 속성을 표제, 형식, 일자, 언어, 기타식별특성을 비롯하여 자료유형별 속성까지 제시하였으나, LRM에서는 속성을 유형, 크기, 이용 대상자, 언어 등으로 축소하고 세부 사항은 구축 환경에 위임하는 방향으로 정비되었다(Riva et al., 2017). 표현형 구별에 필요한 요소는 이를 수용한 목록규칙을 통해 구체화된다.

서지레코드에는 개체별 속성이 여러 필드에 혼재되어 있으며, 표현형을 구성하는 속성 또한 모형 또는 규정에 따라 그 범위가 상이하다. 따라서, 기존의 MARC 서지레코드로부터 WEMI 개체를 식별하기 위해서는 과거 구현형 중심으로 혼재되어 입력된 속성들로부터 FRBR의 구분에 적합하게 데이터를 처리하고 추출하는 과정이 필요하다. 또한, MARC 기반 서지레코드의 리더와 필드 가운데 어느 부분에 개체 관련 속성이 포함되는지, 그리고 그 값이 어떻게 구성되어 있는지를 파악해야 한다.

2.2 한국목록규칙 제5판(KCR5)

국제적인 서지 제어 패러다임이 FRBR 기반의 개체-관계 모형으로 전환되고 AACR2가 RDA로 전면 개정됨에 따라, 국내에서도 이에 대응하여 2025년 KCR5가 제정되었다. KCR5의 가장 큰 특징은 과거 KCR4까지 유지되어 온 '레코드 중심의 기술'에서 벗어나, WEMI 모델을 전면적으로 수용한 '개체 중심의 기술'로의 전환을 선언했다는 점이다(한국도서관협회, 2025).

KCR5는 저작, 표현형, 구현형, 개별자료, 개인, 가계, 단체 등 각 개체의 속성을 독립적으로 기술하고, 이들 간의 관계를 명시적으로 연결하는 방식을 채택하고 있다. 기존 KCR4에서 자원의 유형별로 기술 방식을 제시하던 것과 달리, 자원의 내용적 측면과 물리적 측면을 명확히 구분하여 기술하도록 안내하고 있다. 내용적 측면으로는 저작 및 표현형, 물리적 측면으로는 구현형 및 개별자료로 구분하여 기술하도록 안내하며 저작과 표현형에 대한 속성을 다수 추가함으로써 내용적 측면을 강조하고 있다(김정현, 2025; 한국도서관협회, 2003; 2025).

KCR5는 표현형 속성을 기술하는 요소로 내용유형, 표현형의 일자, 표현형의 언어, 표현형의 기타 식별특성, 표현형의 식별기호, 표현형의 내용을 제시하며, 이 가운데 표현형의 내용을 제외한 나머지를 모두 핵심요소로 규정하고 있다. 또한 KCR5는 이러한 속성을 통제된 값으로 기록할 것을 전제한다는 점에서 이전 규칙과 구별된다. 이러한 속성은 개별 요소로 기록될 뿐 아니라 접근점의 일부로도 기록될 수 있는데, 표현형의 전거형접근점은 해당 표현형

이 실현한 저작의 전거형접근점에 언어와 기타 식별특성, 일자 등의 속성을 부가하여 구성된다(한국도서관협회, 2025).

KCR5는 각 개체별로 속성을 분리하여 기술할 것을 명시함으로써 정보자원 간 계층적 구조를 명확히 하고 이용자로 하여금 탐색의 효율을 높일 수 있도록 고안되었다. 이러한 KCR5의 기술 규칙이 적용된다면, 이용자는 본인이 원하는 특정 번역본이나 판본을 정확히 구별하여 선정하고 획득하는 정보자원 탐색 환경을 제공받을 수 있으리라 예측된다. 다만 이러한 기술체계가 실제로 기능하기 위해서는 기존의 개별 서지레코드로부터 저작과 표현형이 먼저 식별될 필요가 있다.

2.3 선행연구

2.3.1 MARC 서지레코드의 FRBR 적용 연구

FRBR 및 LRM 발표에 따라 나타난 국제적 목록 환경의 변화를 국내 환경에 반영하기 위한 연구가 수행된 바 있다. 이미화 외(2022)는 LRM 이후 국제적 목록 동향을 반영하여 KORMARC 통합서지용에서의 수용 방안을 마련하였다. LRM이 발표된 2017년부터 2021년까지의 MARC 21 토론문서를 분석하여 KORMARC에 신규 수용이 필요한 필드를 식별하고 그 적용 방안을 제안하였으며, 특히 표현형 기술에 관련하여 370(관련 장소), 386(창작자/기여자 특성), 388(창작기간) 등의 필드가 저작뿐만 아니라 표현형에도 적용될 수 있음을 밝히고, 대표표현형 기술을 위한 387 필드 신설을 고려할 것을 제안하였다.

KCR5가 제정된 이후에는 본격적으로 저작, 표현형, 구현형 개체의 속성을 KORMARC 서지레코드 및 전거레코드에 어떻게 기술할 것인지에 관한 연구가 수행되고 있다. 구체적으로 이미화 외(2026)는 KCR5의 저작과 표현형의 속성 및 전거형접근점을 KORMARC에 기술하는 구체적인 방안을 제안하였다. 서지레코드에는 구현형과 개별자료의 속성을 기술하고 전거레코드에는 저작과 표현형의 속성 및 전거형접근점을 기술하는 구현 시나리오를 제시하였다. 특히 표현형의 핵심 속성인 내용유형(336), 표현형의 언어(377, 041, 008/35-37), 기타 식별특성(250, 251, 381)에 대한 KORMARC 필드 매핑을 제안하고, 표현형의 전거형접근점이 저작의 전거형접근점에 언어(▼l), 기타식별특성(▼s), 일자(▼f) 등의 식별기호를 추가하여 구성될 수 있음을 제시하였다.

또한 노지현 외(2026)는 구현형에 초점을 맞추어 KCR5의 구현형 기술규칙을 KORMARC 통합서지용에 적용하기 위한 방안을 제안하였다. KCR5에서 KCR4의 자료 일반명칭(GMD)과 특수명칭(SMD)이 삭제되고, 내용유형(336), 매체유형(337), 수록매체유형(338)으로 대체되었음을 밝히며, 해당 필드를 통해 구현형의 유형을 보다 구조적으로 기술할 수 있음을 제시하였다. 아울러 구현형 레코드에서 저작과 표현형으로의 관계를 1XX 및 240 필드의 조합으로 기록하는 방안을 제안하였다.

이상의 연구들은 KCR5의 개체 중심 기술 체계를 KORMARC에 구현하기 위한 구조적 매핑 지침을 제시하였다. 본 연구는 이들 표현형 속성과 KORMARC 필드의 대응 관계를 표현형 식별의 기반으로 삼았다. 이러한 매핑은 향후 작성될 레코드에 어떤 필드를 어떻게 입력할 것

인가에 관한 규범에 해당하며, 이미 축적된 데이터에 해당 필드가 실제로 어떻게 입력되어 있는지를 규명하지는 않았다. 이에 본 연구는 선행연구에 제안된 필드의 입력 수준을 파악하고, 표현형 식별에의 유용성을 살펴보고자 하였다.

2.3.2 WEMI 개체 식별 연구

서지레코드에서 FRBR/LRM의 개체를 식별하고 집중화하려는 시도는 해외에서 먼저 시작되었다. 대표적으로 OCLC는 FRBR Work-Set Algorithm을 고안하였으며, MARC 레코드의 저자명과 통일표제(130/240)를 핵심 키로 활용하여 저작 수준의 클러스터링을 수행하였다(Hickey & Toves, 2005; 2009). 미국 의회도서관(Library of Congress) 또한 유사한 방식으로 MARC21 레코드를 FRBR로 디스플레이하기 위한 도구를 개발하여 서지레코드를 저작 수준으로 클러스터링한 바 있다(Library of Congress, 2004). 그러나 이들 연구는 주로 저작(Work) 수준의 식별에 집중하였으며, 표현형(Expression) 수준까지 구분하는 실증적 연구는 상대적으로 부족한 편이다.

국내에서는 조재인(2004)이 FRBR 모형의 표현형 계층에 주목하여, 이를 국내 목록 체계에 수용하기 위한 방안을 모색하였다. 이 연구에서는 통일표제(130/240)가 저작과 표현형을 식별하는 핵심 수단임을 강조하고, FRBR의 효과적 구현을 위해 통일표제의 적극적 도입과 전거 데이터 구축이 선행되어야 함을 강조하였다. 그러나 이는 개념적 수용 방안에 해당하며, 실제 서지데이터로부터 표현형을 식별하는 실증적 분석이 수행되지는 않은 한계가 있다.

나상오 외(2025)는 국립중앙도서관의 KORMARC 서지레코드를 대상으로 저작 식별 알고리즘을 적용하여 저작 수준의 클러스터링을 수행하고, 한자-한글 변환, 인명 정규화, 네트워크 기반 교차 클러스터링 등의 방법론을 통해 저작 식별의 정확도를 향상시켰다. 대규모의 KORMARC 데이터를 활용한 저작 식별 실증 연구로 FRBR 개체인 '저작'의 식별에 유의미한 개선을 이루었으나 '표현형'을 식별하는 단계까지는 나아가지 못한 한계가 있다.

종합하면, KCR5의 개체별 기술 방안에 대한 이론적 연구와 저작 수준의 식별에 대한 실증적 연구는 성과를 보였으나, 표현형 수준의 식별에 관한 실증적 연구는 여전히 부족한 상태이다. 이는 표현형 기술을 위해 마련된 필드가 기존 서지레코드에 거의 입력되어 있지 않아 특정 필드를 적정 식별 기준으로 삼기 어려웠기 때문으로 판단된다. 이에 본 연구는 표현형을 위해 마련된 필드 외에 245▼e와 700▼e의 역할어, 250의 판사항 같은 범용 필드와 자연어 값을 활용하고자 하였다. 또한, 이러한 자연어 데이터를 활용하기 위해서는 KCR5의 표현형 속성에 대응하도록 값을 정규화하는 과정이 필요하므로, 전처리 절차를 속성별로 설계하고 그 결과를 문학의 요목(장르)별로 비교함으로써 어떤 필드가 어느 자료 유형에서 표현형 식별에 기여하는지를 실증적으로 확인하고자 하였다.

3. 연구 설계

3.1 데이터 개요

본 연구에서 사용하는 데이터는 국립중앙도서

관의 국가서지과로부터 제공(2024. 3. 14.)받은 한국십진분류법(KDC) 기준 810(한국문학)에 해당하는 KORMARC 서지레코드이다. 확장자 'mrc' 형태의 데이터 파일에는 총 453,846건의 서지레코드가 포함되어 있으며, 각 레코드는 일반적인 MARC 레코드와 동일한 리더(Leader), 제어필드(00X), 데이터필드(01X-9XX)의 구조를 갖는다.

한국문학 데이터를 분석 샘플로 선정한 이유는 동일 저작이 다양한 판본, 번역, 해석, 주석 등의 형태로 반복 출판되는 특성이 나타나 하나의 저작 아래 복수의 표현형이 존재할 가능성이 높은 분야이기 때문이다. 또한, 한문으로 작성된 고전의 현대 한국어 번역 및 한국문학의 외국어 번역본과 같은 언어 기반 표현형 분기는 물론, 시, 희곡, 소설, 수필 등의 장르 차이로 인한 표현형 분기가 발생할 빈도가 높은 주제 영역이기도 하다.

본 데이터 세트를 활용하여 저작을 식별한 선행연구(나상오 외, 2025)가 수행된 바 있다. 해당 연구에서는 총 268,684개의 저작 군집 내 417,886개의 서지레코드가 식별되었다. 이 중 복수의 서지레코드로 구성된 저작 군집은 56,749개이며, 이에 속하는 서지레코드는 205,951건이다.

3.2 KORMARC 필드 선정 및 현황 분석

앞서 살펴본 KCR5의 표현형 속성 가운데, 본 연구에서는 이전 목록규칙(KCR4) 기반으로 작성된 서지레코드를 활용함에 따라 핵심요소와 일반요소의 필수 기술 여부 등을 고려하지 않고 선행연구 및 KORMARC 통합서지용에 근거하여 표현형 속성과 관련된 필드를 파악하였다.

특히, 선행연구(이미화 외, 2026)에는 저작과 표현형의 속성이 KORMARC 서지레코드의 어떤 필드에 기술되는 것이 적절한 지 제시된 바 있다. 본 연구는 이러한 선행연구와 KCR5의 표현형 속성에 대한 정의를 기반으로, 표현형 식별을 위해 추출 및 분석해야 할 KORMARC 서지레코드 필드를 〈표 1〉과 같이 선정하였다. 선행연구에 제시되지 않은 필드 또한 KORMARC 통합서지용의 전체 필드를 검토하여 표현형 식별에 유효할 가능성이 높다고 판단되면 선정하였다. 구체적으로는 서로 다른 표현형임을 확인할 수 있는 기여자 정보(예: 번역자, 편역자, 삽화가 등)가 포함된 '표제와 책임표시사항'(245) 필드 ▼e(두 번째 이하 책임표시)와 '부출표목'(7XX) 필드 역할어와 관계(▼e, ▼4)를 추가로 선정하였으며, 동일 저작의 다른 표현형을 직접적으로 지시하는 번역저록(767), 이판저록(775) 필드까지 고려하였다.

다음 매핑표를 근거로 205,951건의 서지레코드를 조사하여, 해당 필드들이 실제로 얼마나 입력되어 있으며 어떠한 값의 형태로 존재하는지 현황을 살펴보았으며, 그 결과는 〈표 2〉와 같다. 단, 표현형식(348, 500, 546)을 제외한 '표현형의 내용' 관련 필드는 현황 분석에서 배제하였다. 해당 필드는 내용유형에 따라 사용 여부가 달라 문자자료 중심인 본 데이터에서는 기입률이 0%에 수렴하기 때문이다.

필드 입력 현황을 살펴본 결과, 표현형 식별을 위해 활용 가능한 KORMARC 서지레코드의 핵심 필드 입력 빈도가 현저히 낮은 것으로 나타났다. 특히 표현형 식별에 주요한 역할을

〈표 1〉 표현형 속성과 KORMARC 통합서지용 필드 매핑

표현형 속성	관련 KORMARC 통합서지용 필드		출처
내용유형	• 리더/06(레코드 유형) • 336(내용유형)		이미화 외 (2026)
표현형의 일자	• 046(특별한 연도부호) ▼k(시작일), ▼l(종료일) • 388(창작 기간)		
표현형의 언어	• 008/35-37(부호화정보필드 - 공통/언어) • 041(언어부호) ▼a(본문언어)/▼h(원저작의 언어) • 1XX(기본표목)/7XX(부출표목)/240(통일표제)/243(종합통일표제) ▼l(저작의 언어) • 377(관련언어) • 546(언어주기)		
표현형의 기타 식별특성	• 250(판사항) • 251(버전 정보) • 381(저작 또는 표현형의 기타 구별 특성)		
	• 245(표제와 책임표시사항) ▼e(두 번째 이하 책임표시) • 700(부출표목-개인명) ▼e(역할어), ▼4(관계) • 767(번역저록) • 775(이판저록)		KORMARC 통합서지용
표현형의 식별기호	• 024(기타표준식별자)		
	• 130/240(통일표제) ▼0(전거레코드 제어번호 또는 표준번호), ▼1(Real World Object URI)		이미화 외 (2026)
표현형의 내용	내용 요약	520	
	캡처 정보	370, 388, 033, 518	
	표현형식	348, 500, 546	
	접근보조수단	341, 546	
	삽화/색상/음향	300 ▼b, 340	
	화면비율	345 ▼c, ▼d	
	연주수단	048, 382, 130/240/243▼m	
	재생시간	008/18-20, 300▼a, 306▼a, 505 ▼a	
	지도 정보 (축척, 도법표시 등)	034, 지도 255▼a/▼b, 342, 정지화상 507, 500	
	수상	586	
	주기	500, 508, 511 등	

〈표 2〉 표현형 속성을 포함한 KORMARC 필드 현황

KORMARC 필드	기입 건수	기입률(%)	입력값 예시	활용 가능성
008	**205,947**	**99.99**	**970711s1996 ulka j 000a kor**	**O**
리더/06	**205,946**	**99.99**		**O**
024	0	0.00	-	-
041	**6,607**	**3.21**	**akor ajpn aeng**	**O**
046	0	0.00	-	-

KORMARC 필드	기입 건수	기입률(%)	입력값 예시	활용 가능성
130	2	0.00	a춘향전	-
240	6	0.00	aRabbit's tail. l한국어 0KAT202307322	-
242	15	0.01	a종의 기원. ykor	-
243	0	0.00	-	-
245	**205,951**	**100.00**	**a녹색바벨탑 / d박태엽 지음**	**O**
250	**13,762**	**6.68**	**a影印本**	**O**
251	0	0.00	-	-
336	**46**	**0.02**	**atext btxt 2rdacontent**	**O**
348	0	0.00	-	-
377	0	0.00	-	-
381	0	0.00	-	-
388	0	0.00	-	-
500	23,184	11.26	a권말부록으로 "先祖考石農府君家狀" 등 수록	-
508	39	0.02	a제작진: 음향, 최은정	-
511	334	0.16	a목소리출연: 이재범, 김연아, 박종희	-
546	2,599	1.26	a한베대역본임	-
700	**174,234**	**84.60**	**a최인자, e번역**	**O**
710	**8,968**	**4.35**	**a한국고전신서편찬회, e편**	**O**
711	2	0.00	a문학의 해 기념 특별세미나(d1996)	-
730	46	0.02	a느리게 살아서 즐거운 나날들	-
767	1,599	0.78	t我的忐忑人生 z9787532151899 w(011001)CMO201...	-
775	15	0.01	t나의 쓸모: the notes z9791196854409 w(011001)	-

하는 '내용유형'과 관련된 필드를 살펴봤을 때 내용유형(336) 필드는 기입률이 0.02%에 불과한 반면, 리더/06의 경우 99.99% 입력되어 활용 가능성이 충분할 것으로 기대되었다. '표현형의 언어'와 관련된 언어(377) 및 기타 식별특성(381)의 경우 단 한 건도 입력되지 않은 것으로 확인되었으나, 언어부호(041)와 부호화정보 필드(008/35-37)는 충분히 활용 가능할 수준으로 기입률이 높게 나타났다.

'표현형의 기타식별특성'에 해당하는 판사항 및 기여자 정보가 포함된 필드의 경우, 표제와 책임표시사항(245), 개인명 부출표목(700) 필드는 기입률이 각각 100%, 84.60%로 압도적이었으며, 단체명 부출표목(710) 또한 4.35% 가량 입력되어 활용 가능할 것으로 예측되었다. 판사항(250) 또한 6.68% 입력된 것으로 나타나며 높은 기입률을 보이진 않았지만 판의 변형에 따른 표현형을 식별하는 데 활용할 수 있을 것으로 판단되었다. 245, 250, 700, 710 필드의 데이터는 정형화된 기호 또는 통제어휘가 아닌 자연어 형태로 역할어(옮김, 번역 등)가 불규칙하게 혼재되어 있었으며 이를 활용하기 위한 대책이 요구되었다. 언어 정보를 포함하는 008/35-37의 경우 거의 모든 레코드(99.99%)에 기입되어 있었으며, 언어부호(041)는 3.21% 입력되어 있었다.

이 외 주기사항(5XX)의 다수 필드에 정보가 입력된 것을 확인할 수 있었지만, 실제 데이

터 값을 확인한 결과 통제어휘가 아닌 것은 물론 불규칙적인 문장 형태로 기술된 것이 대부분인 탓에 표현형 식별에 직접적으로 활용하기는 어려울 것이라 판단하여 배제하였다.

현황 분석 결과는 기존 축적된 KORMARC 서지레코드로부터 정형화된 값이 입력되는 특정 필드(336, 377, 381 등)만을 활용하여 표현형을 식별하는 것이 사실상 불가능함을 시사하며, 여러 자연어 데이터가 혼재된 필드(245, 700 등)를 적극적으로 활용해야 함을 의미한다. 이에 본 연구에서는 내용유형(리더/06, 336), 언어(008/35-37, 041), 기타식별특성(245, 250, 700, 710)과 관련된 정보를 수록하고 있는 필드에서 데이터를 추출하고 표현형을 식별하고자 하였다.

3.3 KORMARC 필드 데이터 추출 및 정제

앞선 현황 분석을 토대로 표현형 속성별(내용유형: 리더/06, 336, 언어: 008/35-37, 041, 기타식별특성: 245, 250, 700, 710)로 필드를 구분하고 데이터를 추출하였다. 추출 대상 필드 중에는 통제어휘가 아닌 자연어 형태로 기술된 데이터가 상당 부분 존재하는 것으로 확인되었으며, 표현형을 식별하기 위해 이를 통제어휘 또는 기호 형태로 다음과 같은 정제 과정을 거쳤다.

첫째, 내용유형과 관련하여 336 필드를 우선 탐색하여 '텍스트', '구어' 등으로 입력된 자연어 데이터를 'TEXT', 'SPOKEN_WORD'와 같은 통제어휘로 변환하였다. 이때 336 필드가 누락되었다면 모든 서지레코드에 입력되어 있는 리더/06 값을 참조하여 결측치를 보완하였다. 리더/06의 코드값 또한 'a'(문자자료)는 'TEXT', 'i'(음악 이 외 녹음자료)는 'SPOKEN_WORD', 'c'(악보)는 'SCORE', 'j'(음악 녹음자료)는 'PERFORMED_MUSIC', 'e'(지도자료)는 'MAP' 등으로 변환하였다(〈표 3〉 참조).

둘째, 언어 정보를 파악하기 위해 041 ▼a를 통하여 서지레코드의 본문 언어를 추출하고 3자리 언어코드(ISO 639-2)에 맞추어 통일하였으며, 041 필드가 누락된 경우에는 008/35-37을 참조하여 보완하였다. 이때 041 ▼h가 존재한다면 해당 서지레코드가 번역본임을 의미하므로, 이 경우에는 별도의 태그를 통하여 번역본임을 체크하였다(〈표 4〉 참조).

셋째, 기타식별특성을 포함하는 필드의 데이터는 '판차'와 '기여자 정보'를 포함하는 필드를 구분하여 정제하였다. 판차 정보가 담긴 250 필드의 데이터에는 괄호 '[ ]' 혹은 '쇄' 표시와 같은 내용 변화가 없는 인쇄 차수가 섞여 있었다. 이러한 데이터로부터 표현형 식별에 필요한 부

〈표 3〉 내용유형 관련 필드(336, 리더/06) 정제 예시

KORMARC 데이터 예시	정규화 적용 필드	적용 로직	정제 결과 예시
리더/06: a, 336: 없음	리더/06	통제어 매핑, 결측치 보완	TEXT
336: atext btxt 2rdacontent	336	통제어 매핑	TEXT
리더/06: i, 336: 없음	리더/06	통제어 매핑, 결측치 보완	SPOKEN_WORD
리더/06: e, 336: 없음	리더/06	통제어 매핑, 결측치 보완	MAP

〈표 4〉 언어 관련 필드(041, 008/35-37) 정제 예시

KORMARC 데이터 예시	정규화 적용 필드	적용 로직	정제 결과 예시
008/35-37: kor, 041: 없음	008	주 언어 추출, 결측치 보완	KOR, 번역 False
041: aeng hkor	041	원본 언어 감지	ENG, 번역 True
041: ajpn	041	주 언어 추출	JPN, 번역 False
041: akor achi	041	다국어 추출	KOR, CHI, 번역 False

분만을 활용하기 위하여, 정규표현식을 통해 특수기호 및 '쇄' 표시를 제거하였다. 이후 '개정', '수정', '증보' 등의 단어가 발견될 시 이를 모두 'REVISED'로 통일하고, '감독판' 또는 '확장판'은 'SPECIAL_CUT'으로 통일하였다. 또한 숫자 형식(제2판, 2nd ed)은 숫자만 추출하여 'EDITION_2'와 같이 규격화하였다.

기여자 정보를 담고 있는 245, 700, 710 필드의 경우, 우선 인명의 정규화를 진행하였다. 구체적으로 영문 표기는 모두 소문자로 변환하였으며, 온점(.) 및 반점(,)을 제거하여 동일 인물이 별개의 인물로 식별되는 것을 방지하였다. 또한 ▼e에 기재된 '옮김', '번역', '역자', '편곡' 등의 역할어는 정형화된 통제어휘로 매핑하여 변환하였다(〈표 5〉 참조).

3.4 표현형 식별 과정 설계 및 적용

표현형의 식별에 앞서 수행되어야 하는 저작 식별의 경우, 동일한 데이터 세트를 활용하여 저작 개체를 식별한 나상오 외(2025)의 결과를 차용하였다. 복수의 서지레코드를 포함하여 둘 이상의 표현형으로 세분될 가능성이 있는 56,749개의 저작을 표현형 식별 대상으로 선정하였으며, 해당 저작 클러스터가 포함하고 있는 서지레코드 총 205,951건을 대상으로 표현형 식별을 수행하였다.

표현형을 식별하기 위하여 앞선 필드 추출 과정에서 정제한 데이터를 활용하여 표현형 세분화를 수행하였다. 각 표현형 속성별 파생 정도를 알아보기 위하여 내용유형, 언어, 기타식별특성을 독립적으로 활용하는 과정과 이를 모두 순차적으로 활용하는 과정을 구분하여 수행하였다.

표현형 속성을 순차적으로 적용한 것은 KORMARC 데이터에서 빈번하게 발생하는 결측치 문제를 통제하기 위함이다. 세 속성의 값을 하나의 키로 결합하여 일괄 매칭할 경우, 어느 한 필드가 비어 있는 서지레코드는 값이 존재하는 서지레코드와 서로 다른 조합을 갖게

〈표 5〉 기타식별특성 관련 필드(245, 250, 700, 710) 정제 예시

KORMARC 데이터 예시	정규화 적용 필드	적용 로직	정제 결과 예시
250: a개정판	250	키워드 매핑	판차: REVISED
250: a2판 5쇄	250	쇄차 제거, 숫자 추출	판차: EDITION_2
250: a[수정판]	250	특수기호 제거, 키워드 매핑	판차: REVISED
700: a홍길동, e옮김	700	공백 제거, 역할어 매핑	기여자: 홍길동(TRANSLATOR)

되어 동일한 표현형이 별개의 군집으로 분리된다. 이에 본 연구에서는 내용유형, 언어, 기타식별특성의 순서로 군집을 단계적으로 세분하되, 각 단계에서 해당 필드값이 결측인 레코드는 새로운 군집을 형성하지 않고 상위 군집에 유지되도록 처리하였다. 이를 통해 필드값에 명시적인 차이가 확인되는 경우에만 군집이 분화되도록 하여, 동일한 표현형이 분리되는 현상을 방지하였다.

4. 연구 결과

4.1 서지레코드의 특성 분석

선행연구(나상오 외, 2025)에 따르면 KDC 제6판 기준 한국문학(810)에 해당하는 417,886건의 서지레코드에서 총 268,684개의 저작이 식별되었다. 이 가운데 복수 서지레코드로 구성된 저작 클러스터의 수는 56,749개였다. 즉, 표현형이 구분될 여지가 있는 저작 클러스터는 전체 저작 중 21.12%에 해당하였으며, 포함된 서지레코드 수는 205,951개(49.28%)로 확인되었다. 복수 서지레코드로 구성된 저작 현황은 〈표 6〉과 같다. 2~5개 가량의 서지레코드로 구성된 저작이 전체 저작 클러스터의 90%에 달하는 반면, 51개 이상의 서지레코드가 포함된 대규모의 저작 클러스터는 0.33%에 불과한 것으로 나타났다.

저작 식별의 대상이 된 서지레코드는 시, 희곡, 소설, 수필, 연설 등의 다양한 형식으로 존재한다. 이는 KDC의 요목 수준으로 구분하여 파악할 수 있으며 205,951건의 레코드로부터 요목별 분포를 확인한 결과, 소설이 48.77%로 절반에 가까운 비중을 차지하고 있었으며, 시(21.84%), 르포르타주(9.52%), 수필(8.58%) 등이 뒤를 이었다(〈표 7〉 참조).

문학 형식별 서지레코드 수 대비 저작 수를 비교해 보면, 각 장르가 지닌 저작 집중도 차이가 뚜렷하게 확인된다. 가장 높은 비중을 차지하고 있는 장르인 소설(813)의 경우 100,434건의 서지레코드가 24,417개의 저작으로 집중되어 저작 1개당 평균 약 4.11개의 서지레코드를 포함하는 것으로 나타났다. 이들 중에는 조창

〈표 6〉 표현형 식별 대상 저작의 서지레코드 현황

동일 저작 내 서지레코드 수	저작 수	총 서지레코드 수	비율(%)
2개	31,752	63,504	55.95
3개	11,744	35,232	20.69
4개	5,160	20,640	9.09
5개	2,619	13,095	4.62
6~10개	3,745	27,097	6.60
11~20개	1,098	15,507	1.93
21~50개	446	13,441	0.79
51~100개	130	9,073	0.23
101개 이상	55	8,362	0.10
계	56,749	205,951	100.00

〈표 7〉 요목별 서지레코드 현황

KDC 요목	저작 수	서지레코드 수	비율 (%)	
			저작	서지레코드
810 (한국문학 일반)	4,876	15,078	8.59	7.32
811 (시)	15,405	44,970	27.15	21.84
812 (희곡)	589	1,686	1.04	0.82
813 (소설)	24,417	100,434	43.03	48.77
814 (수필)	4,886	17,664	8.61	8.58
815 (연설, 웅변)	19	87	0.03	0.04
816 (일기, 서한, 기행)	1,188	5,248	2.09	2.55
817 (풍자, 해학)	367	1,179	0.65	0.57
818 (르포르타주 및 기타)	5,000	19,600	8.81	9.52
819 (기타 문학)	2	5	0.00	0.00
계	56,749	205,951	100.00	100.00

인의 '가시고기'와 박경리의 '토지'와 같이 240건 가량의 서지레코드를 포함하고 있는 사례도 존재하였다.

이에 반해, 시(811)는 44,970건의 레코드가 15,405개의 저작으로 집중되어 저작당 평균 2.92개의 서지레코드가 집중된 것으로 확인되었으며, 희곡(812) 역시 1,686건의 레코드가 589개의 저작으로 묶여 평균 2.86개라는 상대적으로 낮은 집중도를 보이는 것을 확인할 수 있었다.

반면, 5,248건의 서지레코드가 포함되어 상대적으로 낮은 비중을 차지하는 일기 및 서한(816) 장르는 총 1,188개의 저작이 식별되며, 저작당 평균 4.42개의 서지레코드를 포함하는 것으로 나타났다. 이는 전체 장르 대비 높은 저작 집중도를 보인 것으로, 역사적 사료 가치가 있는 고전 일기나 편지류가 잦은 주해, 번역, 개정본 출간을 통해 다수의 서지레코드(예: 박지원 '열하일기', 이순신 '난중일기')로 실현된 탓으로 예측된다. 이러한 저작의 집중도 분석은 표현형 및 구현형으로 실현된 서지레코드를 다수 포함하고 있는 특정 장르(소설, 서한/일기 등)의 서지학적 특성을 명확히 보여준다.

4.2 한국문학 요목별 표현형 속성 분포

표현형 식별에 있어 요목(문학장르)별로 어떠한 표현형 속성에 주로 영향을 받는지 파악하기 위하여, 앞서 선정한 KORMARC 필드의 요목 수준별 분포를 파악하였다(〈표 8〉 참조). 내용유형 관련된 필드의 기입률은 서로간 상반된 양상을 보인다. 336의 경우 요목별로 0건부터 최대 28건에만 입력되어 있는 반면, 리더/06의 경우 장르를 막론하고 100%에 수렴하는 입력률을 보인다.

언어의 경우 008/35-37은 전 요목에서 100%에 수렴하는 반면, 041은 요목에 따라 0.00%에서 8.90%까지 뚜렷한 편차를 보인다. 041의 기입률은 희곡(812, 8.90%)과 한국문학 일반(810, 7.10%)에서 상대적으로 높고, 수필(814, 1.00%)과 풍자, 해학(817, 1.02%)에서 낮으며, 연설, 웅변(815)에서는 전혀 나타나지 않는다. 041은 원저작의 언어가 존재하는 경우에 기술되며 생

〈표 8〉 장르별 표현형 속성 필드의 출현 현황

KDC 한국문학	내용유형		언어		기타식별특성			
	리더/06	336	008/35-37	041	245	250	700	710
810 (한국문학-일반)	15,078 (100.00%)	2 (0.01%)	15,077 (99.99%)	1,070 (7.10%)	15,078 (100.00%)	2,314 (15.35%)	11,726 (77.77%)	2,116 (14.03%)
811 (시)	44,970 (100.00%)	6 (0.01%)	44,970 (100.00%)	2,184 (4.86%)	44,970 (100.00%)	2,340 (5.20%)	39,761 (88.42%)	1,557 (3.46%)
812 (희곡)	1,686 (100.00%)	0 (0.00%)	1,686 (100.00%)	150 (8.90%)	1,686 (100.00%)	120 (7.12%)	1,406 (83.39%)	50 (2.97%)
813 (소설)	100,432 (100.00%)	28 (0.03%)	100,432 (100.00%)	2,408 (2.40%)	100,434 (100.00%)	6,542 (6.51%)	84,961 (84.59%)	3,750 (3.73%)
814 (수필)	17,663 (99.99%)	1 (0.01%)	17,664 (100.00%)	176 (1.00%)	17,664 (100.00%)	1,261 (7.14%)	13,877 (78.56%)	415 (2.35%)
815 (연설, 웅변)	87 (100%)	0 (0.00%)	87 (100.00%)	0 (0.00%)	87 (100.00%)	2 (2.30%)	80 (91.95%)	0 (0.00%)
816 (일기, 서한, 기행)	5,248 (100.00%)	2 (0.04%)	5,248 (100.00%)	174 (3.32%)	5,248 (100.00%)	449 (8.56%)	4,568 (87.04%)	184 (3.51%)
817 (풍자, 해학)	1,179 (100.00%)	0 (0.00%)	1,179 (100.00%)	12 (1.02%)	1,179 (100.00%)	41 (3.48%)	729 (61.83%)	108 (9.16%)
818 (르포르타주 및 기타)	19,598 (99.99%)	7 (0.04%)	19,599 (99.99%)	433 (2.21%)	19,600 (100.00%)	693 (3.54%)	17,122 (87.36%)	788 (4.02%)
819 (기타 문학)	5 (100.00%)	0 (0.00%)	5 (100.00%)	0 (0.00%)	5 (100.00%)	0 (0.00%)	4 (80.00%)	0 (0.00%)

략될 수 있는 필드이므로, 이러한 편차는 각 요목에 번역 자료가 포함된 비중을 반영한 것으로 해석된다.

기타식별특성 관련 필드에서는 요목별 차이가 가장 뚜렷하게 드러난다. 245는 전 요목에서 100% 기입되어 그 자체로는 요목별 차이가 없는 반면, 250, 700, 710은 요목에 따라 상이한 분포를 보인다. 250(판사항)은 한국문학 일반(810)에서 15.35%로 가장 높으며, 이는 두 번째로 높은 일기, 서한, 기행(816, 8.56%)의 두 배에 가까운 수치이다. 700(부출표목-개인명)은 대부분의 요목에서 78%~92% 수준으로 고르게 높게 나타나나, 풍자, 해학(817)에서만 61.83%로 두드러지게 낮다. 710(부출표목-단체명) 역시 한국문학 일반(810)이 14.03%로 압도적이며, 그 외 요목은 대체로 4% 미만에 머무른다.

이상의 결과를 종합하면, 한국문학 일반(810)은 모든 표현형 속성 필드가 고르게 입력되어 있어 여러 속성에서 동시에 표현형 파생이 나타날 것으로 예상된다. 이는 해당 요목에 문학 전집이나 기관 편찬물이 다수 포함되어 판차 갱신과 단체 기여가 빈번한 데 따른 것으로 보인다. 시(811), 소설(813), 연설, 웅변(815), 르포르타주 및 기타(818)는 700이 84~92%에 이르는 반면 나머지 필드의 기입률이 모두 낮아, 기타식별특성의 기여자 정보에 의존하여 표현형 파생이 나타날 것으로 예상된다. 희곡(812)은 700이 83.39%로 유사한 수준을 유지하는 가운

데 041이 8.90%로 전 요목 중 가장 높아, 기여자 정보와 언어가 함께 작용할 것으로 보인다. 반면, 풍자, 해학(817)은 700이 61.83%로 가장 낮고 250 또한 3.48%에 그쳐 표현형을 구분할 근거 자체가 가장 부족하며, 표현형 식별이 상대적으로 어려운 요목에 해당한다.

이와 같은 요목별 표현형 속성 필드 분포는 표현형 파생을 유발하는 주요 기점(개정판, 다국어 번역, 단체의 지적 기여 등)이 장르에 따라 어떻게 차별화되는지를 이해하는 데 도움이 된다. 또한 요목별로 활용할 주요 필드를 고려할 수 있도록 도우며, 주요 필드의 기입률이 전반적으로 낮은 요목에 대해서는 식별의 난도를 미리 예측하여 대비할 수 있도록 한다.

4.3 표현형 식별 결과

대규모 저작 클러스터로부터 표현형이 세분화되는 과정을 확인하기 위해 내용유형, 언어, 기타식별특성의 표현형 속성별로 독립적인 표현형 파생을 살펴본 이후, 이를 순차적 적용하여 최종 표현형을 식별하였다. 이때 적용 순서의 경우, 무엇을 우선하더라도 중간 단계에서의 차이 외 최종 식별 결과에서의 차이는 없으므로 내용유형-언어-기타식별특성의 순서로 진행하였다.

우선 단일 표현형 속성만으로 표현형 클러스터를 식별하였다. 이는 각 속성이 단독으로 적용되었을 때 어느 정도의 표현형 파생을 유발하는지 관찰하기 위한 것이다. 다만 모든 속성을 함께 고려한 것이 아니므로, 서로 다른 표현형이 하나의 클러스터에 혼재할 가능성이 남아 있어 그 결과를 최종적으로 식별된 표현형으로 볼 수는 없다는 점을 유의해야 한다.

56,749개의 저작을 대상으로 내용유형 관련 필드만 적용한 경우 425개의 표현형이 파생되었으며, 저작당 평균 1.01개의 표현형이 도출되었다. 예시로는 저작 '벌거벗은 임금님'이 있으며, 텍스트와 구어의 두 표현형으로 구분되어 각각 7건과 1건의 레코드가 배정되었다. 다음으로 언어 관련 필드만을 적용한 경우 총 1,515개 파생되었으며, 저작당 평균 1.03개의 표현형이 도출되었다. 대표적인 예시로는 '엄마의 속삭임: 결혼이민여성을 위한 다국어 태교 동화'가 있으며, 베트남어, 몽골어, 영어, 타갈로그어 등 7개 언어의 표현형으로 구분되었다. 내용유형과 언어 필드의 적용 결과는 모두 미미한 수준의 파생에 그쳤는데, 이는 분석 대상 서지레코드가 모두 한국문학에 속하여 대부분의 자료가 한국어 텍스트로 구성되어 있기 때문으로 보인다.

반면 기타식별특성 관련 필드만을 적용한 경우 총 46,417개가 파생되었으며, 세 속성 가운데 가장 많은 파생이 나타났다. 이는 문학 자료의 특성상 번역과 편역, 삽화 추가 등 기여자의 개입에 따른 변환이 빈번하고, 개정과 증보 등의 판차 갱신 또한 자주 발생하는 데 따른 것으로 판단된다. 대표적인 예시로는 고전 저작 '구운몽'이 있으며, 단일 저작에서 총 92개의 표현형이 구분되었다.

다음으로는 표현형 세 속성을 모두 반영한 최종 결과를 확인하기 위하여 내용유형, 언어, 기타식별특성에 해당하는 필드를 모두 순차적 적용하여 표현형을 식별하였다(〈표 9〉 참조). 특히 저작 현황 및 표현형 속성의 분포를 살펴본 바와 같이 요목별로 표현형의 식별 결과를 살펴보았다.

〈표 9〉 요목별 표현형 식별 결과

KDC 한국문학	파생 표현형 수				저작당 평균 표현형 수
	내용유형 (336, 리더/06)	언어 (041, 008/35-37)	기타식별특성 (245, 250, 700, 710)	파생 표현형 수 합계	
810 (한국문학-일반)	33	205	2,875	3,113	1.64
811 (시)	97	263	10,634	10,994	1.71
812 (희곡)	4	7	421	432	1.73
813 (소설)	206	870	21,964	23,040	1.94
814 (수필)	31	42	4,309	4,382	1.90
815 (연설, 웅변)	0	0	15	15	1.79
816 (일기, 서한, 기행)	11	37	1,115	1,163	1.98
817 (풍자, 해학)	1	2	263	266	1.71
818 (르포르타주 및 기타)	42	83	3,502	3,627	1.74
819 (기타 문학)	0	0	2	2	2.00
계	425	1,509	45,100	47,034	1.83

내용유형을 기준으로 한 표현형 식별 결과는 앞서 단일 속성을 적용한 결과와 동일하다. 이 단계에서는 텍스트 외에 오디오북(구어)이나 음악 등으로 실현된 자료가 별개의 표현형으로 분리되어 총 425개의 파생이 나타났으며, 이는 전체 저작의 0.75%에 해당하는 미미한 수준이다. 요목별로는 소설(813)이 206개, 시(811)가 97개로 전체 파생의 71.29%를 차지하였으나, 두 요목이 전체 저작의 70.18%를 차지하고 있음을 감안하면 소설과 시라는 문학장르 고유의 특성에 기인한 것으로 보기는 어렵다. 저작 수 대비 비율로 환산할 경우 일기, 서한, 기행(816)이 0.93%, 소설(813)과 르포르타주 및 기타(818)가 각 0.84%로 상대적으로 높고 나머지 요목은 모두 0.7% 미만에 머물러, 요목 간 편차는 크지 않은 것으로 확인되었다. 336 필드의 기입률이 0.02%에 불과하였음에도 425개의 파생이 확인된 것은 리더/06을 보완적으로 활용한 전략이 기능하였음을 보여줌과 동시에 한국문학 데이터에 한해서는 내용유형이 표현형 분화의 주된 기준으로 작동하지는 않는다는 점 또한 드러낸다.

다음으로 언어 관련 필드를 추가로 적용한 결과, 언어의 차이 및 번역 여부에 따라 1,509개의 표현형이 추가로 분리되었다. 저작 수 대비 파생은 한국문학 일반(810)에서 4.20%로 가장 높았고, 소설(813) 3.56%, 일기, 서한, 기행(816) 3.11%가 뒤를 이었다. 810의 결과는 041 기입률

이 7.10%로 높게 나타났던 〈표 8〉의 결과에 부합하며, 해당 요목에 다국어 자료와 외국어 번역본이 다수 포함되어 있음을 시사한다.

다만 희곡(812)과 소설(813)에서 예측과 상반된 양상이 확인되었는데, 희곡(812)의 경우 041 기입률이 8.90%로 전 요목 중 가장 높았음에도 언어 단계의 파생은 7개, 저작 대비 1.19%에 그쳤으며, 반대로 소설(813)은 041 기입률이 2.40%로 낮은 편이었음에도 저작 대비 3.56%의 파생이 발생하였다. 이러한 결과는 필드의 기입률이 해당 속성의 활용 가능성을 보여주는 지표일 뿐, 실제 표현형 파생과 직결되지는 않음을 보여준다.

마지막으로 판사항 및 기여자 정보와 같은 기타식별특성을 기준으로 표현형을 추가 식별하였으며, 이 단계에서 45,100개의 파생이 관찰되었다. 이는 전체 47,034개의 표현형 파생 대비 95.89%에 해당하는 규모로, 단일 속성 적용에서와 마찬가지로 기타식별특성이 표현형 분화를 주도하는 속성임이 재확인되었다. 저작 수 대비 파생은 일기, 서한, 기행(816)에서 93.86%로 가장 높았고 소설(813) 89.95%, 수필(814) 88.19%로 뒤를 이었으며, 한국문학 일반(810)은 58.96%로 가장 낮았다.

이상의 세 단계를 거쳐 56,749개의 저작으로부터 47,034개의 표현형이 추가로 분화되었으며, 최종적으로 205,951건의 서지레코드가 103,783개의 표현형 클러스터로 분할되었다. 요목별 최종 결과를 '저작당 평균 표현형 수'로 살펴보면 일기, 서한, 기행(816)이 1.98개로 가장 높고 소설(813) 1.94개, 수필(814) 1.90개 순으로 나타나며, 한국문학 일반(810)이 1.64개로 가장 낮았다. 이는 앞선 요목별 속성 분포에서의 기대에 부합하지 않는 결과이며, 판사항의 기입 여부와 표현형 분화가 직결되지 않는 데서 비롯된 것으로 보인다. 250 필드에 기록된 값 가운데에는 인쇄 차수와 같이 내용의 변화를 수반하지 않는 표시가 포함되어 있으며, 이는 구현형 차원의 차이에 해당하여 표현형 파생으로 이어지지 않는다.

요목별 표현형 파생의 규모는 개별 필드의 기입률보다 저작 집중도와 더 밀접하게 연동되는 경향을 보였다(〈표 6〉, 〈표 9〉 참조). 저작당 서지레코드 수가 많은 일기, 서한, 기행(816, 4.42건)과 소설(813, 4.11건)이 저작당 표현형 수에서도 각각 1위와 2위를 차지하였는데, 이는 하나의 저작에 묶인 서지레코드가 다양할수록 그 내부에서 서로 다른 속성값이 발견될 여지가 커지기 때문으로 해석된다.

한편 풍자, 해학(817)은 700 기입률이 61.83%로 가장 낮아 표현형 식별이 어려울 것으로 예상되었으나, 기타식별특성 단계의 파생률은 71.66%로 전체 평균 79.47%와 큰 괴리를 보이지는 않았다. 이는 기타식별특성을 단일 필드가 아닌 245, 250, 700, 710의 조합으로 구성한 데 따른 결과로, 특정 필드가 결측이더라도 나머지 필드가 이를 부분적으로 보완할 수 있음을 보여준다.

5. 논의 및 결론

본 연구는 전통적인 KCR4 체제에 맞춰 구축된 KORMARC 서지레코드로부터 KCR5가 요구하는 핵심 개체인 표현형을 식별하기 위한 방법을 설계하고, 이를 국립중앙도서관의 한국

문학 서지레코드에 적용하여 검증하였다. 서지데이터의 FRBR화는 이용자에게 지적 또는 예술적 창작물의 다양한 변형을 직관적으로 제공하기 위해 필수적이나, 기존 데이터베이스를 단번에 새로운 규칙으로 재목록화하는 것은 불가능에 가깝다. 이에 기존 데이터에 산재한 다양한 필드와 자연어를 분석하고 정제하여 점진적으로 표현형을 식별하는 과정을 진행하였다.

필드 현황 분석 결과, RDA 수용 과정에서 신설된 표현형 관련 필드(336, 377, 381 등)는 기입률이 극히 저조하여 단독 활용이 불가능한 것으로 확인된 반면, 표제와 책임표시사항(245), 부출표목(700, 710), 판사항(250), 언어부호(041) 등 KCR4 체제에서 관행적으로 사용되어 온 범용 필드에는 표현형을 구분할 수 있는 단서가 자연어 형태로 포함되어 있었다. 이를 정규표현식과 통제어휘 변환을 통해 정제한 뒤 적용한 결과, 56,749개 저작에 속한 205,951건의 서지레코드가 103,783개의 표현형 클러스터로 식별되었다.

표현형을 식별하는 과정에서는 표현형의 속성 중 내용유형, 언어, 기타식별특성을 중심으로 살펴보았다. 그 결과 전체 파생의 95.89%는 기타식별특성 단계에서 발생하였는데, 이는 한국문학이 국내 창작물의 텍스트 자료 위주로 구성되어 내용유형과 언어에 의한 변별력이 제한적인 반면 개정과 기여자 개입은 활발하다는 자료 특성을 반영한 결과이다.

이러한 경향은 문학장르에 따라 상이하였다. 저작당 표현형 수는 일기, 서한, 기행(816) 1.98개, 소설(813) 1.94개로 높았던 반면, 기타식별특성 관련 필드의 기입률이 전 요목 중 가장 높았던 한국문학 일반(810)은 1.64개로 오히려 가장 낮았다. 이는 필드 기입률은 해당 속성의 활용 가능성을 보여줄 뿐, 동일 저작에 속한 레코드 사이에서 표현형이 분화되기 위해서는 실제 입력값의 차이가 중요하다는 점을 보여준다.

이는 곧 표현형 식별에는 대상 자료별 주제와 형식 등 특성에 따라 적절한 값이 포함된 KORMARC 필드를 선정하는 것이 중요함을 시사한다. 만약 본 연구의 표현형 식별 과정을 영미문학(KDC 84X)이나 서양음악(KDC 67X)과 같은 타 주제 분야의 데이터에 확장 적용한다면, 언어부호(041), 악보/음원(336) 등에 따른 표현형 수가 대거 식별될 것이다. 따라서, 성공적인 표현형 식별을 달성하기 위해서는 사전 데이터 분석을 통해 대상 서지레코드를 면밀히 파악하고 이에 최적화된 필드 및 식별 순서를 유연하게 채택하는 접근이 필요하다.

본 연구의 결과는 향후 KCR5 환경에서 KORMARC 서지레코드를 작성할 때 고려해야 할 필드와 그 기술 방안에 대한 시사점도 제시한다. '내용유형'과 관련하여, 입력된 레코드가 현저히 적었던 336 필드는 '리더/06'로 대체할 수 있었지만, 이를 위한 매핑 과정과 정규화 과정을 수반하므로, 불필요한 과정을 줄이기 위해서는 336 필드를 필수적으로 기술할 수 있도록 해야 한다.

다음으로 '언어' 관련 필드 중 레코드에서 기술하는 개체와 관련된 언어부호를 입력하는 377 필드는 입력 값이 전무한 상태였다. 이를 활용하기 위해서는 저작의 언어와 표현형의 언어가 다르거나 동일 저작의 다른 표현형과 구별할 필요가 있는 경우에 한하여 필수 기술 필드로 규정하는 방안을 검토할 수 있다.

'기타식별특성'에 관련된 381 필드 또한 입력

된 값이 전무하여 활용할 수 없었으므로 통제어휘 또는 부호화된 값을 통해 필수적으로 기술하는 방안을 고려할 만하다. 이 외에도 기여자 역할의 통제가 요구된다. 표현형 파생은 대부분 기타식별특성에서 발생하였음에도 그 근거가 되는 245▼e와 700▼e에는 역할어가 자연어로 혼재되어 있으므로, KCR5의 관계표시어를 700▼e에 적용하거나 관계코드를 700▼4에 함께 기록한다면 별도의 정제 없이도 역할 판별이 가능할 것으로 기대된다.

하지만 연구 결과에는 한계점 또한 존재한다. 우선 표현형을 구분하는 핵심 요소인 역할어가 목록자에 따라 상이한 자연어로 입력되어 있어 정제 과정만으로는 완전한 정규화에 이르지 못하였다는 한계가 있으며, 표현형 수준의 고유 식별자가 부재한 상황에서 문자열 매칭에 의존한 결과 띄어쓰기나 오탈자의 차이가 식별 결과에 영향을 미쳤을 가능성이 있다. 따라서, 표현형 식별 결과의 정확성과 타당성을 보장할 수 없다는 한계가 존재한다. 또한, 분석 범위가 단일 기관의 특정 주제에 속하는 서지레코드로 한정되어 있어, 본 연구에서 확인된 속성별 기여도를 다른 주제 분야나 목록 관행이 상이한 기관의 데이터에 그대로 일반화하기는 어렵다.

이와 같은 한계를 바탕으로 후속 연구에서는 번역과 음악, 시청각자료 등 언어와 내용유형에 의한 파생이 활발한 주제 분야로 대상을 확장하여, 자료 특성에 따라 속성별 기여도가 어떻게 달라지는지를 분석할 필요가 있다. 또한, 표현형 식별 결과의 정확성을 검증하기 위하여, 목록 전문가의 판정을 거친 정답 데이터를 일정 규모로 구축하고 표현형 식별 결과와 비교하여 정확성을 정량적으로 검증하는 과정의 추가를 고려해야한다. 아울러 자연어처리 기술과 거대언어모델(LLM)을 접목하여 기여자 및 판차 정보를 보다 정교하게 파싱하고 오기입과 누락을 자동으로 보정하는 방안을 시도해볼 수 있을 것이다.

참 고 문 헌

국립중앙도서관 (2023). KORMARC 통합서지용. 출처: https://www.nl.go.kr

김정현 (2025). 한국목록규칙 제5판과 제4판의 서지기술 비교 분석. 한국도서관·정보학회지, 56(3), 1-18. https://doi.org/10.16981/kliss.56.3.202509.1

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

나상오, 강우진, 이종욱 (2025). 한국문학 분야 KORMARC 서지레코드를 활용한 저작 식별 개선 방안 연구. 한국도서관·정보학회지, 56(2), 109-132. https://doi.org/10.16981/kliss.56.2.202506.109

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 (2026). KCR5 구현형 기술규칙의 KORMARC 적용 방안에 관한 연구. 한국문

헌정보학회지, 60(2), 125-147. https://doi.org/10.4275/KSLIS.2026.60.2.125

이미화 (2019). BIBFRAME에서 LRM 표현형 및 대표표현형 속성 적용시 고려사항. 한국비블리아학회지, 30(2), 33-50. https://doi.org/10.14699/kbiblia.2019.30.2.033

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, 이은주, 노지현 (2026). KORMARC에서의 KCR5 저작·표현형 기술에 관한 연구. 한국문헌정보학회지, 60(1), 55-75. https://doi.org/10.4275/KSLIS.2026.60.1.055

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.

한국도서관협회 (2003). 한국목록규칙 (제4판). 서울: 한국도서관협회.

한국도서관협회 (2025). 한국목록규칙 (제5판). 서울: 한국도서관협회.

Hickey, T. B. & Toves, J. (2005, April). FRBR Work-set algorithm. Available: https://www.oclc.org/content/dam/research/activities/frbralgorithm/2005-04.pdf

Hickey, T. B. & Toves, J. (2009). FRBR Work-set algorithm version 2.0. Available: http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf

Kim, M., Chen, M., & Montgomery, D. (2021). Moving toward BIBFRAME and a linked data environment. In S. S. Hines eds. Advances in Library Administration and Organization (Vol. 42). Leeds: Emerald Publishing Limited, 131-154.

Kroeger, A. (2013). The road to BIBFRAME: The evolution of the idea of bibliographic transition into a post-MARC future. Cataloging & Classification Quarterly, 51(8), 873-890. https://doi.org/10.1080/01639374.2013.823584

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

• 국문 참고자료의 영어 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, Jeong-Hyen (2025). A comparative analysis of bibliographic description in KCR5 and KCR4. Journal of Korean Library and Information Science Society, 56(3), 1-18. https://doi.org/10.16981/kliss.56.3.202509.1

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

Korean Library Association (2003). Korean Cataloguing Rules (4th ed.). Seoul: Korean Library Association.

Korean Library Association (2025). Korean Cataloguing Rules (5th ed.). Seoul: Korean Library Association.

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 (2019). Considerations for BIBFRAME acceptance of expression and representative expression attributes in LRM. Journal of the Korean Biblia Society for Library and Information Science, 30(2), 33-50. https://doi.org/10.14699/kbiblia.2019.30.2.033

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, Mihwa, Lee, Eun-Ju, & Rho, Jee-Hyun (2026). A study on the description of work and expression of KCR5 in KORMARC. Journal of the Korean Society for Library and Information Science, 60(1), 55-75. https://doi.org/10.4275/KSLIS.2026.60.1.055

Na, Sangoh, Kang, Woojin, & Lee, Jongwook (2025). A study on improving work identification using KORMARC bibliographic records in the field of Korean literature. Journal of Korean Library and Information Science Society, 56(2), 109-132. https://doi.org/10.16981/kliss.56.2.202506.109

National Library of Korea (2023). KORMARC for integrated bibliographic data. https://www.nl.go.kr/

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 (2026). A study on the application of KCR5 manifestation description rules to KORMARC. Journal of the Korean Society for Library and Information Science, 60(2), 125-147, https://doi.org/10.4275/KSLIS.2026.60.2.125