# SHACL과 ShEx를 활용한 국가서지LOD의 품질 평가*

## - 온톨로지 설계 검증과 실데이터 검증의 이중 층위 분석 -

## A Quality Assessment of the National Bibliographic LOD Using SHACL and ShEx: A Dual-Layer Analysis of Ontology Design and Instance Data Validation

박 진 호 (Jin Ho Park)**


### 초 록

본 연구의 목적은 RDF 검증을 위한 W3C 표준 언어인 SHACL과 ShEx를 활용하여 국립중앙도서관 국가서지 LOD의 품질을 온톨로지 설계 층위와 실데이터 층위의 양면에서 진단하는 것이다. 표본 추출이 아닌 전수조사 방식을 채택하여 2026년 4월 1일자로 약 9억 1,500만 트리플 전체를 분석하였다. 분석 결과, 설계 층위에서는 전체 속성의 41.4%가 정의역을 명시하지 않는 명세 불완전성과 BIBFRAME 어휘 미반영을 확인하였다. 실데이터 층위에서는 SHACL 검증 결과 99.73%의 높은 적합률을 보였으나 위반의 99.6%가 서지(온라인자료)의 표제 누락에 있음을 확인했다. ShEx 검증 결과 연역 형상과 귀납 형상의 평균 미예상 속성 비율이 98.91%에 달해 설계 명세가 실데이터의 표현 풍부성을 포괄하지 못함도 확인하였다. 특히 설계에 부재하던 BIBFRAME 어휘가 실데이터에서는 광범위하게 사용되는 설계-실데이터 비대칭을 확인하였다.

### ABSTRACT

This study aims to diagnose the quality of the National Library of Korea's national bibliographic Linked Open Data (LOD) at both the ontology design layer and the instance data layer, using SHACL and ShEx, the W3C standard languages for RDF validation. Adopting a complete enumeration approach rather than sampling, the study analyzed the entire dataset of approximately 915 million triples published as of April 1, 2026. The analysis revealed, at the design layer, an incompleteness of specification in which 41.4% of all properties lacked a defined domain, along with the non-incorporation of BIBFRAME vocabulary. At the instance data layer, SHACL validation showed a high conformance rate of 99.73%, yet 99.6% of violations were concentrated in missing titles within the serial (online resources) dataset; ShEx validation showed that the average proportion of unexpected properties between the deductive and inductive shapes reached 98.91%, indicating that the design specification fails to encompass the representational richness of the instance data. In particular, a design-instance asymmetry was identified in which BIBFRAME vocabulary, absent from the design, was extensively used in the instance data.


키워드: 링크드 오픈 데이터, 국가서지, SHACL, ShEx, 데이터 품질 평가
Linked Open Data, National Bibliography, SHACL, ShEx, Data Quality Assessment

---

* 본 연구는 한성대학교 학술연구비 지원과제임.
** 한성대학교 지식정보문화트랙 조교수(jhp@hansung.ac.kr / ISNI 0000 0004 7641 0372)
논문접수일자 : 2026년 5월 21일 논문심사일자 : 2026년 5월 24일 게재확정일자 : 2026년 6월 4일
한국비블리아학회지, 37(2): 321-347, 2026. http://dx.doi.org/10.14699/kbiblia.2026.37.2.321

# 1. 서 론

도서관계는 1960년대 말 미국의회도서관(LC)이 개발한 MARC 형식을 활용해 반세기에 걸쳐 서지 데이터를 생산 및 공유해 왔다. 그러나 웹 기반의 경계 없는 데이터 개방·연계 환경은 도서관계 내부 호환에 집중한 MARC의 구조적 한계를 노출시켰다. LC는 2011년부터 링크드 데이터 환경에 적합한 서지 언어인 BIBFRAME 개발에 착수하여 2012년과 2016년에 각각 1.0과 2.0을 공개하였다. 미국·독일·프랑스 등 주요 국가도서관도 BIBFRAME의 실제 적용을 활발히 추진하고 있다.

국립중앙도서관 또한 2012년부터 국가서지 LOD(Linked Open Data)를 구축하여 약 2,700만 건의 서지·전거·도서관 정보를 RDF로 공개하고 있으며, owl:sameAs 및 skos:closeMatch를 통해 LC, BnF, DNB, Wikidata 등과의 인터링킹을 제공하고 있다. 링크드 데이터(Linked Data)의 가치는 데이터의 공개 자체가 아니라 발행 데이터가 표준에 부합하고 구조적 일관성과 신뢰성 있는 연결을 제공할 때 실현 가능하다. 더욱이 도서관의 기간 데이터가 BIBFRAME 기반 환경으로 전환하는 시점에서 현행 LOD의 품질을 객관적으로 진단하는 작업은 정책적·기술적 의사결정에 중요한 토대로 활용할 수 있다.

이러한 문제의식에서 본 연구는 RDF 검증을 위한 W3C 표준 언어인 SHACL과 ShEx를 상호 보완적으로 활용하여 국가서지 LOD의 품질을 온톨로지 설계 층위와 실데이터 층위의 양면에서 진단한다. 본 연구가 설정한 연구 문제는 두 가지이다. 첫째, 국가서지 LOD 온톨로지의 설계 수준에서 어떠한 구조적 품질 이슈가 존재하는가? 둘째, 온톨로지 설계 명세에 따라 작성한 SHACL 및 ShEx 형상에 비추어 실제 발행된 트리플 데이터는 어느 수준의 적합성을 보이는가?

# 2. 선행연구

본 연구 관련 선행연구는 도서관 LOD 구축 및 활용, LOD 품질 평가 프레임워크, SHACL 및 ShEx 기반 RDF 검증 세 가지 영역으로 볼 수 있다.

## 2.1 도서관 LOD 구축 및 활용

도서관 영역에서 링크드 데이터의 적용은 MARC 기반의 폐쇄적 서지 환경의 한계를 극복하기 위한 패러다임 전환의 일환이다. 국제적으로는 미국의회도서관의 BIBFRAME 이니셔티브가 가장 대표적인 사례로, MARC를 대체할 차세대 서지 프레임워크로서 링크드 데이터 원칙에 기반한 구조를 제시하고 있다. Gaitanou et al.(2024)는 도서관 LOD 연구를 체계적으로 고찰하여 BIBFRAME 적용, IFLA LRM과의 정렬, 전거 데이터의 링크드 데이터화, 상호운용성 문제 등을 연구 주제로 제시하였다. 최근에는 OCLC가 2024년을 링크드 데이터 가속화의 원년으로 평가하며 WorldCat Entities URI 4억 건을 서지 레코드에 부여하는 대규모 구현을 추진하였고, BIBFRAME Interoperability Group(BIG)을 통해 표준 적용 프로파일과 SHACL 기반 검증 도구의 통합 개발을 진행 중이다(BIBFRAME Interoperability Group,

2024). 변건우와 김혜진(2026)은 Web of Science 게재 LD 연구 1,517편을 대상으로 주경로 분석을 수행한 결과, LD 연구가 인프라 구축에서 품질 고도화를 거쳐 지능형 도메인 활용으로 진화해 왔으며 최근에는 데이터 거버넌스와 링크드 에듀케이션 영역으로 확산되고 있음을 보고하였는데, 이는 본 연구가 다루는 품질 평가 및 거버넌스 논의가 국제적 연구 흐름의 중심 의제임을 시사한다.

박옥남(2012)은 기록물 전거통제 기반 링크드 데이터 구축에 관한 초기 연구를 수행하였으며, 박지영(2012)은 링크드 데이터 방식을 통한 서지 정보의 확장 가능성을 탐색하였다. 조명대(2010)는 도서관에서의 링크드 데이터 활용 방안을 종합적으로 논의하였다. 이유진 외(2009)는 시맨틱 디지털도서관 서비스를 위한 서지 온톨로지 구축 연구를 수행하면서 MARC, DC, MODS, MarcOnt 등의 메타데이터 모델과 FRBR 모델을 분석한 바 있다. 서지 기술 형식의 차세대 표준에 관해서는 박옥남과 오정선(2014)이 BIBFRAME 1.0의 4개 핵심 개체(Work, Instance, Authority, Annotation) 구조와 MARC-BIBFRAME 변환 매핑을 분석하여 다양한 자원 유형에 대한 BIBFRAME의 유연성을 확인하는 한편 복잡한 어휘 체계의 일관된 사용을 위한 상세 지침 마련 선행의 필요성을 지적하였다. 노지현(2019)은 편목 패러다임의 전환 필요성을 강조하여 국내 기관의 데이터 사일로 현상을 지적하면서 RDA 및 BIBFRAME의 수용, 전거 통제 강화, 링크를 통한 데이터 보강(enrichment) 전략의 필요성을 제언하였다. 최근 이미화 외(2026)는 국내 도서관의 BIBFRAME 도입 결정을 배경으로 LC와 스웨덴 국립도서관(KB)의 교육 사례를 분석하여 시맨틱 웹 기초・온톨로지 매핑・입력기 실습으로 구성된 10개 유닛의 교육 모듈을 개발함으로써 BIBFRAME 전환의 인적 기반 구축 방안을 구체화하였다. 이러한 연구의 축적은 국내 도서관계가 BIBFRAME 전환을 본격적인 정책적・실무적 의제로 다루기 시작하였음을 보여 주는데, 본 연구가 진단한 국가서지 LOD의 BIBFRAME 설계-실데이터 비대칭은 바로 이러한 전환의 출발선에서 해소해야 할 선결 과제임을 보여준다. 전거 데이터의 링크드 데이터화에 관한 국내 연구도 지속적으로 이어져 왔다. 이성숙 외(2017)는 국립중앙도서관 LOD에서 인물 정보가 저자 역할로는 객체로 관리되는 반면 주제 역할로는 단순 개념으로만 취급해 상호 연계가 어려운 문제를 지적하고 FOAF Person 개체로의 통합 식별 및 ISNI・VIAF와 인터링킹 범위 확장을 제안하였다. 박지영(2024)의 법령 기반 분류체계 연구도 같은 맥락에서 전거 LOD의 품질과 외부 연계성 확보를 다루었다. 국제 동향과 관련해서는 이성숙(2022)이 미국・유럽 등 21개 주요 도서관의 LD 구축 현황을 발행 데이터세트・재사용 어휘집・인터링킹의 세 측면에서 전수 조사하여 해외 기관들이 전거 LOD를 기반으로 검색 가시성을 높이고 고유 특화 데이터세트를 적극 발행하고 있음을 확인하였으며, 국내 도서관에도 BIBFRAME 전환과 개체 중심 기술 강화를 제언하였다. 본 연구에서 전거 데이터셋에 대한 품질을 검증하는 것도 이러한 연구와 연속성 상에 있으며 실증 연구로서 가치가 있다고 볼 수 있다. 대용량 서지 LD 구축의 실무적 과제와 운영 차원의 연구도 이어져 왔다. 이문호와 최성필(2017)은

MEDLINE급 대용량 서지 데이터의 LD 변환에서 발생하는 자원 소모 문제를 다루어 이중 일괄 등록 방법을 통해 트리플 저장소 접근 횟수를 약 1만분의 1 수준으로 감축하는 알고리즘을 제안하였는데, 이는 약 9억 1,500만 트리플 규모의 국가서지 LOD를 전수조사 방식으로 분석한 본 연구의 방법론적 정당성을 뒷받침하는 선행 사례이다. 박진호와 곽승진(2021)은 프랑스 국립도서관(BnF) 등의 LD 서비스 사례를 벤치마킹하여 국립중앙도서관 근대문학자료의 LD 기반 서비스 방안을 설계하면서 기계 처리 가능한 고품질 원천 데이터 생성의 중요성을 강조하였다. 이 사례는 본 연구의 분석 대상 기관에서 LD 서비스의 고도화가 실제로 추진되어 왔음을 보여 주는 동시에, 본 연구가 진단한 품질 이슈가 향후 서비스 확장을 위해 우선적으로 해소해야 할 토대 과제임을 시사한다. 그러나 이러한 국내 선행연구의 흐름은 주로 LOD의 구축 방안, 온톨로지 설계 모형 제안, 정책적 가이드라인 제시, 서비스 개편 방안 등에 집중하고 있으며, 이미 구축한 LOD 데이터의 품질을 정량적·체계적으로 평가하는 연구는 극히 제한적이다. 특히 국립중앙도서관의 국가서지 LOD가 2012년부터 운영되어 왔음에도 그 데이터의 품질 수준에 대한 학술적 진단 연구는 본 연구자가 검토한 범위 내에서 발견하지 못했다. 이는 국가 단위 서지 인프라의 신뢰성 평가가 아직 학술적 공백으로 남아 있음을 시사한다.

### 2.2 LOD 품질 평가 프레임워크

LOD 데이터의 품질 평가는 시맨틱 웹 연구의 초기부터 핵심 과제로 다루어져 왔다. 가장 영향력 있는 이론적 기여는 Zaveri et al.(2016)가 Semantic Web 저널에 발표한 체계적 문헌 고찰로, 18개 품질 차원과 69개 세부 품질 지표를 제안하여 LOD 품질 평가의 표준적 분류체계를 제공한 바 있다. 이 분류체계는 후속 연구에서 광범위하게 인용되며 LOD 품질 평가의 기준 역할을 수행하고 있다. Debattista et al.(2016)는 Luzzu 프레임워크를 통해 자동화된 LOD 품질 평가 방법론을 제시하였으며, Sejdiu et al.(2019)는 Apache Spark 기반 분산 처리 프레임워크인 DistQualityAssessment를 제안하여 대용량 LOD 데이터에 대한 확장 가능한 품질 평가의 가능성을 입증하였다. Zhang et al.(2023)는 의료 정보학 영역에서 RDF 자원 품질 평가를 위한 자동화 접근법을 제시하면서 URI 품질 차원의 중요성을 강조하였다. 이러한 일련의 연구는 LOD 품질 평가가 추상적 논의 단계를 넘어 자동화·확장성·재현성을 갖춘 실증 분석의 영역으로 진입하였음을 보여 준다. 도서관 영역에 특화된 품질 평가 연구로는 Király(2024a)의 QA Catalogue 도구가 주목할 만하다. 이 도구는 MARC21, UNIMARC, PICA 등 전통적 서지 포맷에 대한 품질 평가를 자동화하는 프레임워크로서, Shacl4Bib과 결합하여 도서관 메타데이터 품질 평가의 통합적 도구 생태계를 형성하고 있다. 한편 Cortés et al.(2025)는 Zaveri et al.(2016)의 69개 품질 지표 전체에 대해 SHACL 형상을 정의하고 자동 평가 프로토타입을 구현함으로써, 기존의 추상적 품질 차원 분류와 실제 검증 도구 사이의 간극을 메우는 중요한 방법론적 진전을 보여 주었다.

### 2.3 SHACL 및 ShEx 기반 RDF 검증

SHACL(Shapes Constraint Language)과 ShEx(Shape Expressions)은 RDF 데이터의 구조적 적합성을 검증하기 위한 W3C의 양대 표준 언어로서, 각각 다른 설계 철학을 바탕으로 발전해 왔다. ShEx는 2014년 W3C 커뮤니티 그룹을 중심으로 개발, RDF 그래프의 구조를 기술하는 스키마 언어로서 설계되었다. 반면 SHACL은 2017년 W3C 권고안으로 채택된 제약 기반 검증 언어이다. 두 언어의 핵심적 차이는 검증 모델의 철학에 있다. ShEx는 RDF 그래프 구조를 기술하고 그에 대해 검증하는 것을 목표로 하는 스키마 기반 접근으로, 노드가 정의된 형상(shape)에 부합하는지를 판정한다. 반면 SHACL은 RDF 그래프가 일련의 제약 조건을 만족하는지를 검증하는 제약 기반 접근으로, 위반 사항의 구체적 보고에 강점을 갖는다. 두 언어의 형식적 공통점과 차이에 관해서는 Ahmetaj et al.(2025)가 SHACL, ShEx, PG-Schema의 공통 기반을 형식적으로 비교하여 정리한 바 있다. 본 연구의 맥락에서 두 언어의 차이를 정리하면 〈표 1〉과 같다.

ShEx 기반 검증 연구의 대표적 성과는 Wikidata 공동체에서의 적용이다. Thornton et al.(2019)는 ESWC 2019에서 ShEx를 활용하여 Wikidata와 FHIR(Fast Healthcare Interoperability Resources)의 RDF 데이터 모델을 공유하고 검증한 사례를 보고하였다. 도구적 측면에서는 Fernandez-Álvarez et al.(2022)가 Knowledge-Based Systems에 발표한 sheXer 라이브러리가 중요한 진전을 이루었는데, 이는 RDF 데이터로부터 ShEx 형상을 자동으로 추출하는 알고리즘으로 본 연구에서도 핵심적으로 활용한다. 도서관 영역에 특화한 ShEx 적용으로는 Candela et al.(2023)가 Semantic Web 저널에 발표한 도서관 LOD 품질 평가 프레임워크가 가장 직접적인 선행 사례이다. SHACL 기반 검증 연구는 보다 광범위하고 다양한 도메인에서 이루어지고 있다. Spahiu et al.(2016)는 ABSTAT 도구로부터 추출한 의미적 프로파일을 통해 SHACL 제약을 학습하는 방법론을 제안하였다. Duan et al.(2024)는 ESWC 2024에서 다양한 제약 언어를 통합하는 SCOOP 프레임워크를 발표하여 SHACL의 확장 가능성을 제시하였다. 도서관 데이터에 특화된 SHACL 적용은 Király(2024b)의 Shacl4Bib 연구가 가장 대표적이다. 이 연구는 SHACL의 부분집합을 구현하여 비-RDF 기반 포맷(XML, CSV, JSON, MARC21, UNIMARC, PICA)까지 SHACL

〈표 1〉 SHACL과 ShEx의 주요 특성 비교

| 구분 | SHACL | ShEx |
|---|---|---|
| 표준화 | W3C 권고안(2017) | W3C 커뮤니티 그룹 최종 보고서(2017) |
| 설계 철학 | 제약 기반(constraint-based) | 스키마 기반(schema-based) |
| 검증 초점 | 제약 위반의 구체적 보고 | 형상 부합 여부의 판정 |
| 구문 형식 | RDF Turtle 구문 | ShExC 등 전용 컴팩트 구문 |
| 주요 활용 영역 | 기업 데이터 거버넌스, BIBFRAME 검증 | Wikidata, FHIR/RDF 등 |
| 형상 자동 추출 | 부분 지원 | sheXer 등 도구로 잘 지원됨 |

유사 검증을 확장하였으며, 독일디지털도서관(DDB)과 같은 실제 기관에서의 적용 사례를 제시하였다. 또한 BIBFRAME Interoperability Group은 2024년부터 BIBFRAME 데이터 교환의 일관성을 확보하기 위해 DCTap과 SHACL을 결합한 검증 프로파일을 개발하고 있어, SHACL이 차세대 서지 환경의 핵심 검증 도구로 자리잡아 가고 있음을 보여 준다. SHACL의 LOD 품질 평가 적합성에 관한 가장 최근의 이론적 기여는 앞서 언급한 Cortés et al.(2025)의 연구이다. 이 연구는 Zaveri et al.(2016)의 69개 품질 지표 전체에 대해 SHACL 형상을 정의하고 자동 평가 프로토타입을 구현하여 그 적용 가능성을 입증하였으며, SHACL을 통해 LOD 품질 차원 전반을 측정할 수 있는 가능성을 종합적으로 제시했다는 점에서 본 연구의 직접적인 방법론적 토대를 이룬다.

SHACL과 ShEx를 단일 품질 평가 체계 안에서 상호보완적으로 운용하기 위해서는 두 언어가 각각 어떤 품질 층위를 담당하는지 명확히 해야한다. Zaveri et al.(2016)의 LOD 품질 차원 분류체계에 따르면, SHACL은 사전 정의한 제약 조건에 대한 개별 노드의 준수 여부를 판정함으로써 적합성(conformance), 완전성(completeness), 일관성(consistency) 차원을 검증한다. 이는 "현재 데이터가 규정을 준수하는가"라는 데이터 생산 통제 문제를 다룬다. 반면 ShEx는 연역 형상과 귀납 형상의 비교를 통해 설계 명세가 실데이터의 표현 양상을 얼마나 포괄하는지를 기술함으로써 표현 풍부성(representational richness)과 설계-실데이터 정합성을 검증한다. 이는 "현재 온톨로지 설계가 실데이터를 충분히 기술하는가"라는 온톨로지 거버넌스의 문제를 다룬다. 두 도구의 결과는 동일 품질 평가 체계 안에서 서로 다른 층위를 독립적으로 진단하는 것으로, 단순 비교나 중복이 아닌 층위별 분업 관계를 형성한다.

이상의 검토를 통해 다음과 같은 연구 공백을 확인할 수 있다. 첫째, 국제적으로는 SHACL과 ShEx를 활용한 도서관 LOD 품질 평가 연구를 빠르게 이루어지고 있지만, 국내에서는 이러한 표준 검증 언어를 활용한 체계적 품질 평가 연구를 찾아보기 어렵다. 둘째, 국립중앙도서관의 국가서지 LOD는 2012년부터 운영되어 약 2,700만 건의 서지 데이터를 공개하고 있음에도, 그 데이터의 품질 수준에 대한 학술적 진단은 이루어진 바가 없다. 셋째, 선행연구의 대부분은 온톨로지 설계 또는 실데이터 검증 중 한쪽에만 초점을 맞추고 있어, 양 층위를 통합적으로 분석한 사례가 드물다. 넷째, Cortés et al.(2025)의 SHACL-Zaveri 매핑 프레임워크나 Fernandez-Álvarez et al.(2022)의 sheXer 자동 추출 도구와 같은 최신 방법론적 자원을 비영어권 국가서지 데이터에 적용한 실증 연구는 부재한 상태이다. 본 연구는 SHACL과 ShEx의 상호 보완적 활용을 통해 국가서지 LOD의 품질을 온톨로지 설계 층위와 실데이터 층위에서 동시에 진단한다는 점에서 의의를 갖는다. 특히 표본 추출이 아닌 전수조사 방식의 분석으로 국가 단위 서지 데이터 품질 진단의 결락 없는 실증 결과를 산출하고자 한다는 점도 기존 연구와 차별점이다.

## 3. 연구방법

본 연구는 4단계로 수행한다. 1단계는 국가

서지 LOD 온톨로지 파일을 정량적으로 분석하여 설계 층위의 품질 이슈를 식별한다. 2단계는 그 결과에 기반하여 SHACL과 ShEx 형상을 설계한다. 3단계에서 트리플 데이터에 대해 검증을 실행하며, 4단계에서 결과를 Zaveri et al.(2016)의 LOD 품질 차원 관점에서 해석한다. 전체 연구절차는 〈표 2〉와 같다.

분석 대상 자료는 두 종류이다. 첫째는 국립중앙도서관이 공개한 국가서지 LOD 온톨로지 파일(nlk_ontology.rdf)이며, 둘째는 국가서지 LOD에서 제공하는 벌크 데이터셋이다. 본 연구는 표본 추출에 따른 대표성 한계를 피하기 위해 단행본, 연속간행물, 학술기사, 온라인자료, 저자명·주제명 전거, 도서관 정보 등 모든 데이터셋을 전수조사한다. 1단계 온톨로지 분석은 Python rdflib로 수행하며, 기본 통계, 클래스 상속 구조, 외부 표준 어휘 활용, 정의역·치역 명시 여부, 네임스페이스 선언-사용 불일치를 점검한다.

2단계에서 SHACL 형상은 W3C SHACL Core 권고안에 따라 Turtle 구문으로, ShEx 형상은 ShExC 형식으로 작성한다. ShEx는 연역적 접근(설계 명세 기반 이상 형상)과 귀납적 접근(실데이터로부터 자동 추출)을 병행하여 두 형상의 비교를 통해 설계 의도와 실제 데이터 구조의 괴리를 정량화한다. 카디널리티 제약은 KORMARC 필수 필드, RDA 핵심 요소, BIBFRAME 2.0 권장 요소를 참조하여 설정하며 그 근거를 본문에 명시한다.

3단계에서 SHACL과 ShEx 검증을 실행한다. 이 둘은 설계 목적이 다르므로 각기 다른 지표를 산출한다. SHACL은 W3C 권고 제약 기반 검증 언어로서 사전에 정의한 형상에 개별 노드가 부합하는지를 판정하고 위반 사항을 구체적으로 알 수 있는 강점을 갖는다(Ahmetaj et al., 2025). 이에 따라 SHACL 검증은 pySHACL로 RDFS 추론을 활성화하여 수행하며, 클래스별 적합 노드 비율, 속성별 위반 빈도, 위반 유형별 분포(필수 속성 누락·형식 패턴 위반)를 산출한다. 이 지표들은 각각 데이터셋 전반의 구조적 준수 수준, 어떤 속성에 기재 오류가 집중 발생하는지, 품질 문제가 값의 오류에서 비롯하

〈표 2〉 연구절차

| 단계 | 활동 | 분석대상 | 활용도구 | 결과물 |
|---|---|---|---|---|
| 1단계 | 온톨로지 설계 품질 분석 | 국가서지 LOD 온톨로지 파일(RDF/XML) | Python rdflib | 클래스·속성 구조 기술, 설계 품질 이슈 목록 |
| 2단계 | SHACL 및 ShEx 형상 설계 | 1단계 분석 결과 + 실데이터 표본 | SHACL Turtle, ShExC, sheXer | 클래스별 SHACL 형상, ShEx 이상 형상·귀납 형상 |
| 3단계 | SHACL 검증 실행 | 국가서지 LOD 트리플 데이터 표본 | pySHACL | 클래스별 적합 노드 비율, 속성별 위반 빈도, 위반 유형별 분포(필수 속성 누락·형식 패턴 위반) |
| | ShEx 검증 실행 | 국가서지 LOD 트리플 데이터 전체 | rdflib 기반 ShEx 귀납 추출 | 속성 일치율, 카디널리티 일치율, 미예상 속성 비율 |
| 4단계 | LOD 품질 차원 기반 해석 | 1·3단계 산출물 | Zaveri 외(2016) 품질 차원 프레임워크 | 차원별 품질 진단 결과 및 개선 과제 |

는지 아니면 핵심 요소의 부재에서 비롯하는지를 구분하여 진단할 수 있게 한다. 반면 ShEx는 스키마 기반 접근으로 RDF 그래프 전체의 구조적 패턴을 기술하고, 설계 명세(연역 형상)와 실데이터에서 자동 추출한 귀납 형상을 비교함으로써 설계 의도와 실제 데이터 구조 사이의 괴리를 정량화하는 데 강점을 갖는다(Fernandez-Álvarez et al., 2022). 귀납 ShEx 형상 추출은 rdflib 기반으로 클래스별 속성 출현율과 카디널리티를 집계하며 채택 임계값은 0.1로 설정한다. ShEx 산출 지표는 속성 일치율(연역 형상 속성 중 귀납 형상에도 출현하는 비율), 카디널리티 일치율(연역과 귀납 형상 간 카디널리티 제약의 부합도), 미예상 속성 비율(연역 형상에 규정되지 않았으나 귀납 형상에 출현하는 속성의 비율)로 설정한다. 이 세 지표는 각각 설계의 포괄성, 설계 의도와 실제 사용의 부합도, 온톨로지 명세 밖에서 비공식적으로 사용되는 어휘의 규모를 보여 주며, SHACL이 포착하기 어려운 설계-실데이터 표현 격차를 독립적으로 측정하는 역할을 수행한다. 두 도구를 병용하는 이유는 Zaveri et al.(2016)의 품질 차원 분류체계에서 각 도구가 담당하는 층위가 상이하기 때문이다. SHACL은 적합성・완전성・일관성 차원에서 개별 노드의 제약 위반을 정밀하게 보고하지만 온톨로지 설계 명세 자체의 완전성은 측정하지 못한다. ShEx는 설계-실데이터 정합성과 표현 풍부성 차원에서 온톨로지 거버넌스 문제를 드러내지만 개별 위반의 원인을 특정하기 어렵다. 두 도구가 동일한 구조적 사실을 독립적으로 가리킬 경우 방법론적 삼각검증(triangulation) 효과가 발생하여 진단의 신뢰성을 높인다.

4단계에서 1단계와 3단계 결과를 통합하여 설계 결함이 실데이터 품질로 전이되는 구조적 해석을 도출한다. 재현성 확보를 위해 형상과 분석 스크립트는 GitHub 저장소에 공개하고, 데이터셋 버전과 형상 설계의 주요 판단 근거를 명시한다. 분석에 사용한 Python 환경, 스크립트별 기능, 데이터셋 버전 등 재현 환경의 상세 정보는 [부록 1]에 수록하였다.

## 4. 국가서지 LOD 온톨로지 설계 분석

본 장에서는 국가서지 LOD 온톨로지 파일을 정량적・구조적으로 분석하여 첫 번째 연구 문제에 답한다. 모든 분석은 온톨로지 파일에 정의된 어휘 체계 수준으로 한정하며, 실데이터의 어휘 활용 양상에 대한 판단은 6장에서 다룬다.

'nlk_ontology.rdf'를 rdflib로 파싱한 결과, 본 온톨로지는 29개의 자체 클래스와 157개의 속성(ObjectProperty 33개, DatatypeProperty 124개)으로 구성된다. DatatypeProperty가 ObjectProperty의 약 3.8배인 분포는 표제・발행지・분류기호・식별자 등 다양한 속성값을 리터럴로 표현하는 데 비중을 둔 설계임을 보여 준다. 클래스는 자료 유형(Book, NonBook, ElectronicBook 등), 행위자(Author, Library 등), 코드성 분류(DataType, PublicationFrequencyType 등)의 세 개 군으로 분류된다. 대부분의 nlon 클래스는 BIBO・FOAF 상위 클래스를 상속하는 단순한 다층 계층을 형성하며, nlon:Sound가 bibo:Document와 nlon:OnlineMaterial

을 동시 상속하는 다중 상속 1건을 확인할 수 있다. 모든 클래스와 속성에 vs:term_status가 stable로 부여되어(총 181개 항목) 운영 안정화 단계임을 명시하고 있다.

외부 어휘 재사용 평가는 단순 출현 빈도가 아니라 어떤 어휘가 어떤 도메인에 매핑되었는지의 의미적 적합성을 기준으로 이루어져야 한다. 본 온톨로지는 서지 자원 표현에 BIBO, 행위자 표현에 FOAF, 지식 조직 체계에 SKOS, 메타데이터 기술에 Dublin Core, 조직 정보에 W3C Organization, 지리 정보에 WGS84 geo를 채택하여 도메인별 국제 표준에 부합하는 어휘 선택을 확인할 수 있다. 다만 두 가지 한계를 확인할 수 있다. 첫째, BIBFRAME 어휘가 RDF 헤더에 네임스페이스로 선언되어 있으나 어휘 정의 수준에서는 활용하지 않고 있다. 둘째, 자료 형태・발행 빈도・발행 상태 등 코드성 분류 클래스가 국제적으로 skos:Concept 통제 어휘로 표현되는 것이 보편적임에도 자체 클래스로 정의되어, 외부 어휘와의 매핑 가능성을 제약할 수 있다.

속성의 정의역(rdfs:domain)과 치역(rdfs:range)은 추론 엔진의 작동과 형상 설계의 기초가 되는 핵심 정보이다. 157개 속성 전체를 점검한 결과는 〈표 3〉과 같다.

전체 157개 속성 중 41.4%가 정의역을, 26.1%가 치역을 명시하지 않으며, 특히 ObjectProperty는 66.7%가 정의역을 명시하지 않고 있다. nlon:classificationNumber, nlon:keyword, nlon:medium, nlon:isni, nlon:uniformTitle 등 서지 기술의 핵심 속성이 정의역 없이 정의하고 있어 자동 검증 도구로 적용 범위를 판정하기 어렵다. 본 연구의 형상 설계에서는 이러한 속성에 대해 실데이터 출현 패턴을 분석하여 정의역을 보완적으로 추정한다.

## 5. SHACL 및 ShEx 형상 설계

5장은 4장의 분석 결과를 토대로 검증 형상을 설계한다. 형상 설계는 네 가지 원칙을 따른다. 첫째, 검증 단위는 클래스별 NodeShape으로 하며, nlon 자체 클래스(자료 유형 11, 행위자 5, 코드성 분류 7)와 외부 표준 클래스 6개를 포함한 29개 클래스를 대상으로 한다. 단, 본 연구의 연역 형상은 KORMARC 필수 필드, RDA 핵심 요소, BIBFRAME 2.0 권장 요소를 기준으로 한 최소 필수 요소 중심으로 설계했음을 명시한다. 이는 4장에서 확인한 바와 같이 국가서지 LOD 온톨로지가 전체 속성의 41.4%에 대해 정의역을 명시하지 않는 명세 불완전성을 안고 있어, KORMARC 반복 필드・RDA Core Element 전체・BIBFRAME Application Profile・controlled vocabulary・

〈표 3〉 국가서지 LOD 온톨로지의 속성 명세 완전성

| 속성유형 | 전체개수 | domain 미정의 | range 미정의 |
|---|---|---|---|
| DatatypeProperty | 124 | 43 (34.7%) | 28 (22.6%) |
| ObjectProperty | 33 | 22 (66.7%) | 13 (39.4%) |
| 합계 | 157 | 65 (41.4%) | 41 (26.1%) |

identifier 패턴 등을 포괄하는 정교한 연역 형상을 온톨로지 명세만을 근거로 설계하기 어렵다는 현실적 제약을 반영한 것이다. 따라서 본 연구의 ShEx 비교 결과, 특히 미예상 속성 비율은 이러한 설계 범위의 제한을 전제로 해석해야 한다.

외부 클래스를 포함하는 것은 실데이터에서 일부 인스턴스가 외부 표준 클래스로 직접 선언될 가능성을 고려한 것이며, 검증 단계에서 RDFS 추론을 활성화하여 검증 누락을 최소화한다. 둘째, 카디널리티 제약은 KORMARC 필수 필드, RDA 핵심 요소, BIBFRAME 2.0 권장 요소를 종합 참조하여 설정하고 근거를 본문에 명시한다. 셋째, 정의역 미명시 속성은 실데이터 출현 패턴을 분석하여 한 속성이 특정 클래스에 80% 이상 집중 출현할 때 해당 클래스를 정의역으로 귀납 추정한다. 넷째, 데이터 유형 검증은 W3C XSD 표준에 따라 일관된 매핑 규칙을 적용하되, 식별자 검증은 ISBN · ISSN 기본 자릿수와 ISNI 16자리 형식 확인 등 기본 형식 검증으로 한정한다.

SHACL 형상은 각 클래스에 NodeShape을 정의하고 sh:targetClass로 검증 대상을 지정한 뒤 sh:property로 속성 제약을 명세하는 방식으로 작성하였다. 자료 유형 클래스 중 가장 빈도가 높은 nlon:Book의 형상은 표제(dc:title)를 KORMARC 245 필드이자 RDA 핵심 요소로서 필수(sh:minCount 1)로 설정하고, 발행연도 · 저자 · ISBN · 분류기호는 권장 속성으로 두되 존재 시 데이터 유형과 형식 패턴을 엄격히 검증하도록 설계하였다. 이는 고서나 특수자료에서 발행연도 · ISBN이 부재할 수 있는 도서관 실무를 반영한 것이다. 행위자 클래스는 FOAF 어휘를 기반으로 foaf:name을 RDA 우선접근점 핵심 요소로서 필수 설정하였다. 코드성 분류 클래스는 일반 NodeShape과 SKOS 호환성 평가 형상을 병행하여 SKOS 대체 가능성을 정량 측정할 수 있도록 설계하였다. 대표 형상 코드는 [부록 4]에 수록하였다.

ShEx 형상은 ShExC 형식으로 SHACL 형상과 일대일 대응되도록 작성하였다. ShExC의 카디널리티 표기는 정확히 1회(기호 없음), 1회 이상(+), 0회 또는 1회(?), 0회 이상(*)의 네 가지 기호로 표현하며, 형상 간 참조가 SHACL의 sh:class보다 직관적이다. 연역적 형상이 설계 명세에 기반한 이상적 형상을 표현하는 데 비해, 귀납적 형상은 실데이터 출현 패턴으로부터 자동 추출한다. 본 연구는 rdflib 기반으로 클래스별 속성 출현율과 카디널리티를 전수 집계하여 귀납 형상을 추출하며, 채택 임계값 0.1을 적용한다. 귀납 ShEx 형상의 대표 사례와 각 속성별 출현율은 [부록 5]에 수록하였다. 두 형상의 비교에서는 속성 일치율(연역 형상 속성 중 귀납 형상에도 출현하는 비율), 카디널리티 일치율, 미예상 속성 비율(연역 형상에 없으나 귀납 형상에 출현하는 비율)의 세 지표를 산출한다. 이 세 지표는 각각 설계의 포괄성, 설계 의도와 실제 사용의 부합도, 비공식적 사용 패턴의 규모를 보여 준다.

본 장의 형상 설계는 온톨로지 어휘 정의에 기반한 연역적 설계이므로 실제 발행 데이터의 클래스 구성과 차이가 있을 수 있다. 연역 형상이 실데이터의 모든 클래스를 포괄하는지, 그리고 설계 형상과 실데이터 사이의 정합성 문제는 6장에서 데이터 규모를 진단하고 검증 대상을 재확정하는 절차를 통해 다룬다.

## 6. 국가서지 LOD 실데이터 검증 결과

### 6.1 검증 데이터셋의 규모와 구성

본 연구가 분석한 데이터셋은 2026년 4월 1일 기준 벌크 데이터로, 도서관 정보, 서지(오프라인・온라인), 전거(개인명・단체명・주제명)의 여섯 개 데이터셋이다. 데이터셋 형식은 JSON-LD 이다. 전체 규모는 〈표 4〉와 같다.

전체 규모는 약 48.5GB, 3,146만여 건의 레코드, 약 9억 1,500만 트리플이다. 서지(온라인)가 전체 트리플의 약 66.8%, 두 서지 데이터셋이 합산 약 93.9%를 차지하여 국가서지 LOD가 서지 자원 중심 구성임을 확인할 수 있다. 출현 클래스는 중복 제외 31종이며, 외부 표준 어휘 클래스 인스턴스가 약 820만 건으로 무시할 수 없는 비중을 차지하여 5장에서 외부 표준 클래스를 형상 대상에 포함한 결정이 실증적으로 타당했음을 보여 준다. 한편 전거(단체명) 데이터셋에는 단체명 전거 클래스가 아닌 nlon:Author 클래스 인스턴스가 117,419건을 포함하고 있어 데이터셋 명칭과 실제 클래스 구성이 일치하지 않는 현상을 확인하였다. 또한 nlon:ComplexDocument (3건)・nlon:Software(17건)와 같이 극소량만 출현하는 클래스도 확인하였다.

### 6.2 실데이터 기반 형상 검증 대상의 확정

6.1절의 진단 결과 5장의 연역 형상과 실데이터 클래스 구성 사이에 두 유형의 괴리를 확인하였다. 첫째, 설계하였으나 미출현한 클래스가 13종이다. 코드성 분류 6종, 행위자 3종, 자료 유형 1종, 외부 표준 3종이 이에 해당하며, 특히 SKOS 호환성 형상을 설계했던 코드성 분류 클래스가 단 한 건도 인스턴스화되지 않아 코드성 분류가 독립 인스턴스가 아닌 속성값으로만 존재함을 확인하였다. 이 13종은 검증 대상에서 제외하며 미출현 사실 자체를 이후에 분석하여 제시한다. 둘째, 실데이터에 출현하나 5장에서 설계하지 않은 클래스가 15종이다. 특히 자료 유형 최상위 클래스인 nlon:OnlineMaterial (2,181만여 건)과 nlon:OfflineMaterial(714만여 건)이 누락되어 있는데, 이들을 제외하고 검증했다면 전체 인스턴스의 상당 부분이 누락되었을 것이다. 본 연구는 이 15종에 대한 형상을 보완(자료 유형 상위 클래스 기본 NodeShape 추가, BIBO 외부 클래스 7종 형상 추가, 미수

〈표 4〉 국가서지 LOD 검증 데이터셋의 규모

| 데이터셋 | 파일 | 용량(MB) | 레코드 | 트리플 |
|---|---|---|---|---|
| 도서관 정보 | 1 | 8.06 | 14,054 | 174,469 |
| 서지(오프라인) | 36 | 12,271.59 | 7,144,183 | 247,710,228 |
| 서지(온라인) | 110 | 34,632.48 | 21,810,815 | 611,056,066 |
| 전거(개인명) | 10 | 1,941.80 | 1,814,804 | 39,718,380 |
| 전거(단체명) | 1 | 253.11 | 165,652 | 5,075,436 |
| 전거(주제명) | 3 | 574.72 | 515,412 | 11,188,832 |
| 합계 | 161 | 49,681.76 | 31,464,920 | 914,923,411 |

록 nlon 자료 유형 5종 형상 추가)하여 최종 검증 대상을 실데이터 출현 31종 클래스 전체로 확정하였다.

### 6.3 실데이터 속성 활용의 개관

실데이터에 출현하는 속성은 총 240종으로 온톨로지 정의(157개)보다 많다. 형상 검증 결과 해석에 직접 관련되는 세 가지 특성을 확인할 수 있다. 첫째, 의미적으로 동일한 속성이 복수 표기로 분산된다. 저작자 속성은 dc:elements/creator(8,650,322건), dcterms:creator(3,873,174건), dc:terms/creator(3,527,448건) 등 다섯 가지 표기로 분산되어 통합 시 약 2,174만 건이며, 언어 속성도 최소 여섯 가지 표기로 약 1,265만 건이 출현한다. 이는 표상 일관성 측면의 핵심 품질 이슈이다. 둘째, BIBFRAME 어휘가 실데이터에서 사용된다. id.loc.gov 계열 bibframe:language(5,016,134건), bibframe:language 축약형(1,527,821건) 등 총 약 670만 건이 사용되어, 설계에는 부재하나 실데이터에는 존재하는 어휘로서 이중 층위 분석 방법론의 유효성을 입증한다. 셋째, 인터링킹 속성도 sameAs(3,516,750건), owl:sameAs(4,172건), closeMatch(4,577건)로 표기가 분산되어 있다. 이에 따라 6.4절 이하 검증에서는 의미적으로 동일한 속성의 복수 표기를 통합 처리하는 절차를 적용하였다.

### 6.4 SHACL 검증 결과

pySHACL로 RDFS 추론을 활성화하여 161개 파일을 파일 단위로 순차 검증하였다. 검증 대상은 형상이 설계된 28종 클래스, 검증 인스턴스 총수는 약 6,712만 건이다. 전체 적합률은 99.73%로, 약 6,694만 건이 적합하고 약 18.3만 건이 위반하여 전반적으로 높은 구조적 적합성을 보였다. 클래스별 적합률은 〈표 5〉와 같다.

28종 중 17종이 적합률 100.00%를 기록하였으며, 특히 전거(개인명)의 nlon:Author는 161만여 건 전체가 위반 없이 적합하여 개인명 전거의 필수 요소 기재가 충실함을 보여 준다. 적합률 99% 미만은 nlon:Software(94.12%), nlon:ElectronicBook(97.25%), bibo:Website(98.49%)이며, 위반 절대 수는 nlon:OnlineMaterial(약 9.1만 건)과 nlon:ElectronicBook(약 7.9만 건)이 가장 크다. 두 클래스 모두 온라인 자료 계열이라는 점을 주목할 필요가 있다. 위반 유형 분석 결과 전체 위반의 99.66%가 필수 속성 누락,

〈표 5〉 SHACL 검증의 주요 클래스별 적합률(전체 28종 클래스 결과 [부록 2] 참조)

| 클래스 | 인스턴스 | 위반 노드 | 적합률(%) |
|---|---|---|---|
| nlon:OnlineMaterial | 21,810,815 | 90,663 | 99.58 |
| nlon:ElectronicJournal | 16,901,548 | 332 | 100.00 |
| nlon:Book | 6,845,116 | 632 | 99.99 |
| nlon:ElectronicBook | 2,895,364 | 79,487 | 97.25 |
| nlon:Author | 1,618,009 | 0 | 100.00 |
| bibo:Website | 698,310 | 10,579 | 98.49 |
| nlon:Software | 17 | 1 | 94.12 |

0.34%가 형식 패턴 위반으로, 품질 문제가 값의 오류가 아니라 핵심 요소의 부재에서 비롯함을 보여준다. 속성별로는 표제 누락이 181,326건으로 절대 다수이며(dc:title과 dcterms:title 통합 경로 검증), KDC 형식 위반 622건, 주제명 레이블 누락 8건이 뒤를 잇는다.

데이터셋별 분석은 더욱 결정적이다. 전체 위반의 99.6%(181,290건)가 서지(온라인자료)에 집중되었고, 서지(오프라인자료)는 약 2,143만 건 중 위반이 658건에 불과하며, 전거 데이터(개인명·단체명)와 도서관정보는 위반이 전무하였다. 이 뚜렷한 데이터셋 간 품질 격차는 온라인 자료의 메타데이터 생산 과정이 전거나 오프라인 자료에 비해 핵심 요소 기재 통제가 취약함을 시사한다.

### 6.5 ShEx 검증 결과: 연역 형상과 귀납 형상의 비교

연역 형상과 귀납 형상의 비교 종합 지표는 비교 형상 28종에 대해 평균 속성 일치율 8.57%, 평균 카디널리티 일치율 1.79%, 평균 미예상 속성 비율 98.91%로 산출되었다. 표면적으로 두 형상이 거의 일치하지 않으나, 각 지표는 구조적 사실을 드러낸다.

첫째, 속성 일치율이 낮은 결정적 원인은 표제 속성의 표기 불일치이다. 5장 연역 형상은 표제를 dc:title로 규정했으나 실데이터에서 dc:title은 어느 클래스에서도 채택 임계값 이상 출현하지 않았고, 실제 표제 속성은 dcterms:title로 거의 모든 자료 유형 클래스에서 100%에 가깝게 출현하였다. 이는 6.4절 SHACL 검증에서 표제를 dc:title과 dcterms:title 통합 경로로 검증한 것이 타당했음을 ShEx 분석이 독립적으로 입증한 것으로, 두 검증 방법이 동일한 구조적 사실(Dublin Core 두 네임스페이스 혼용, terms 계열이 실질 표준)을 일관되게 가리킨다. 그러나 속성 일치율 저하의 원인은 표제 속성의 표기 불일치만으로 설명하기 어렵다. 저자 속성의 경우 nlon:Book의 연역 형상은 dc:creator를 규정하였으나 귀납 형상에서는 dcterms:creator(출현율 20.24%)와 dc:elements/creator(출현율 37.2%)가 분산 출현하며, dc:creator 단일 표기는 임계값 이상으로 집계되지 않았다. nlon:OfflineMaterial과 nlon:OnlineMaterial에서도 동일한 양상이 확인되어 저자 속성의 표기 분산이 표제 속성과 마찬가지로 속성 일치율 저하에 구조적으로 기여하고 있음을 알 수 있다. 식별자 및 분류 속성에서도 유사한 현상이 나타난다. nlon:Book의 연역 형상에 규정된 nlon:isbn은 귀납 형상에서 dcterms:identifier(출현율 28.23%)로 대체되는 경향을 보이며, nlon:kdc는 귀납 형상에서 30.32%로 출현하여 연역 형상 속성 중 유일하게 임계값을 초과한 속성이다. 이는 nlon:Book이 28종 클래스 중 속성 일치율 40%로 가장 높은 값을 기록하는 배경이기도 하다. 반면 nlon:ElectronicBook을 포함한 온라인 자료 계열 클래스는 연역 형상에 규정된 속성이 1개에 불과하고 그마저도 표기 불일치로 귀납 형상에서 확인되지 않아 속성 일치율 0%를 기록하였다.

둘째, 카디널리티 일치율 1.79%는 JSON-LD 직렬화에서 단일 값이 배열로 직렬화되어 복수로 해석되는 특성과, 연역 형상(규범적 관점)과 귀납 형상(기술적 관점)의 카디널리티 의미론 차이에서 비롯된 것으로 단순한 품질 결함

으로 환원할 수 없다. 카디널리티 불일치의 원인은 두 가지 층위에서 작동한다. 첫 번째 층위는 JSON-LD 직렬화의 구조적 특성으로, JSON-LD에서는 단일 값도 배열로 직렬화될 수 있어 rdflib가 이를 파싱할 때 복수 값으로 해석하는 경우가 발생한다. 이에 따라 연역 형상에서 1회 출현(sh:maxCount 1)으로 규정한 속성이 귀납 형상에서는 복수 출현으로 집계되는 현상이 광범위하게 나타난다. nlon:Author의 foaf:name이 연역 형상에서는 필수 단수 속성으로 규정되었으나 귀납 형상에서 복수 출현으로 집계되는 것이 대표적 사례이다. 두 번째 층위는 연역 형상과 귀납 형상의 카디널리티 의미론 차이이다. 연역 형상은 규범적 관점에서 이상적으로 기재되어야 할 최소·최대 출현 횟수를 규정하는 반면, 귀납 형상은 실데이터에서 실제로 관찰된 출현 패턴을 기술적으로 반영한다. 예컨대 nlon:Concept의 skos:prefLabel은 연역 형상에서 필수 단수 속성으로 규정되었으나, 실데이터에서는 동일 개념에 대해 한국어와 영어 레이블이 병기되어 복수로 출현하는 사례가 존재한다. 이는 다국어 레이블 기술이라는 실무적 필요가 단수 규정의 연역 형상과 충돌하는 현상으로, 카디널리티 불일치가 데이터 오류가 아닌 기술 정책의 차이에서 비롯될 수 있음을 보여 준다. 따라서 카디널리티 일치율 1.79%는 데이터 품질의 전반적 저하를 의미하기보다, 직렬화 방식과 규범적-기술적 카디널리티 의미론의 간극이 수치로 표현된 결과로 해석하는 것이 타당하다. 이 지표를 실질적 품질 판단에 활용하려면 직렬화 방식을 통제한 별도의 검증 절차가 선행되어야 한다.

셋째, 평균 미예상 속성 비율 98.91%는 본 연구의 가장 중요한 발견 중 하나이다. 연역 형상이 클래스당 1~5개 속성을 규정한 반면 귀납 형상은 클래스당 훨씬 많은 속성을 포함하였다(〈표 6〉 참조).

이 격차는 데이터의 결함이 아니라 온톨로지 설계 명세가 실데이터의 표현 풍부성을 포괄하지 못함을 보여 주는 구조적 사실이다. 거의 모든 자료 유형 클래스의 미예상 속성에 bibframe:extent, bibframe:language, bibframe:place가 공통 등장하여, BIBFRAME 사용이 특정 영역에 국한되지 않고 자료 유형 전반에 광범위하게 분포함이 클래스 단위에서 정밀하게 입증된다. 한편 nlon:Author와 nlon:Concept은 연역 형상 속성이 모두 귀납 형상에서 확인되어 속성 일치율 100%를 기록하였으나, nlon:Author 명칭이 foaf:name·skos:prefLabel·rdfs:label로 중복 표현되고 nlon:Concept의 rdfs:label(100%)과 skos:prefLabel(27.62%)이 불균형하여 어휘 혼용을 확인하였다. 이상

〈표 6〉 주요 클래스의 연역 형상 대비 귀납 형상 속성 수(전체 28종 비교 결과 [부록 3] 참조)

| 클래스 | 인스턴스 | 연역 | 귀납 | 미예상 |
|---|---|---|---|---|
| nlon:OnlineMaterial | 21,810,815 | 1 | 36 | 36 |
| nlon:OfflineMaterial | 7,144,183 | 1 | 43 | 43 |
| nlon:Book | 6,845,116 | 5 | 43 | 41 |
| nlon:Author | 1,618,009 | 2 | 15 | 13 |
| nlon:Concept | 515,412 | 1 | 8 | 7 |

의 분석을 종합하면, 세 지표는 국가서지 LOD의 설계-실데이터 관계가 세 가지 구조적 층위에서 균열이 발생하고 있음을 일관되게 가리킨다. 속성 일치율 저하는 온톨로지 설계가 어휘 표기 규범을 충분히 반영하지 못한 결과로, 표제・저자・식별자 등 핵심 속성 전반에 걸친 네임스페이스 분산이 외부 응용의 재사용 비용을 증가시키는 직접적 요인이다. 카디널리티 일치율 저하는 부분적으로 직렬화 방식과 의미론적 차이에서 비롯되므로 즉각적인 데이터 수정 대상이라기보다 검증 방법론의 정교화와 기술 정책의 명문화가 필요한 영역이다. 미예상 속성 비율 98.91%는 실데이터가 온톨로지 설계의 경계를 대규모로 벗어나 있음을 보여 주며, 그 중심에 BIBFRAME 어휘의 비공식적 도입이 있다. SHACL 검증(6.4절)이 위반의 99.6%가 온라인 자료 표제 누락에 집중됨을 보여 주었다면, ShEx 검증은 그 이면에서 설계 명세와 실데이터 표현 전체가 광범위하게 유리되어 있음을 드러낸다는 점에서 두 검증 결과는 상호 보완적으로 국가서지 LOD의 품질 지형을 입체적으로 조명한다.

## 7. 종합 해석: 설계 층위와 실데이터 층위의 통합

본 장은 4장의 설계 분석과 6장의 실데이터 검증을 통합하여 두 층위의 현상이 어떻게 상호 연결되는지를 규명한다. 첫째, 설계 명세와 실데이터의 표현 격차이다. 4장에서 확인한 명세 불완전성(정의역 미명시 41.4%)이 6.5절의 평균 미예상 속성 비율 98.91%로 귀결된다. 4장에서 정의역 미명시로 지적된 속성(nlon:classificationNumberOfNLK, nlon:keyword 등)이 실데이터에서 대규모로 출현한다는 사실은 설계 단계의 명세 공백이 실데이터 단계의 무통제 사용으로 직접 이어졌음을 보여 준다. 즉 설계 층위의 결함과 실데이터 층위의 현상은 인과적으로 연결되어 있다.

둘째, BIBFRAME 어휘의 설계-실데이터 비대칭은 본 연구의 이중 층위 방법론이 가장 선명하게 가치를 발휘하는 지점이다. 설계 층위에서 BIBFRAME은 선언만 되었을 뿐 어휘 체계에 통합되지 않았으나, 실데이터에서는 약 670만 건 사용되며 거의 모든 자료 유형 클래스에 분포한다. 온톨로지만 분석했다면 BIBFRAME 미사용, 실데이터만 분석했다면 BIBFRAME 활용으로 결론지었을 것이나, 두 층위를 함께 분석할 때 비로소 BIBFRAME이 설계의 승인 없이 실데이터에 도입되어 있다는 정확한 진단이 가능하다. 이는 데이터 생산 현장이 차세대 표준의 필요성을 인식・대응하고 있으나 그것이 온톨로지 설계에 반영하지 않은, 데이터 생산과 온톨로지 관리의 분리된 거버넌스 구조를 시사하며, BIBFRAME 전환 준비 시점에서 우선 해소해야할 과제이다.

셋째, 속성 표기 분산은 표상 일관성과 직접 관련된다. 동일 의미 속성이 복수 표기로 분산되면 외부 응용의 재사용 비용이 증가한다. 본 연구에서 SHACL 검증(통합 경로)과 ShEx 분석(dcterms:title이 실질 표준임 확인)이 동일한 사실을 독립적으로 가리킨 점은 두 검증 도구의 상호 보완적 활용이 표상 일관성 진단에 실효적임을 입증하며, 단일 도구 의존 시 표기 분산을 단순 누락으로 오판할 위험이 있었음을

시사한다. 넷째, 데이터셋 간 품질 격차의 구조적 원인이다. 전거 데이터의 높은 품질은 전거 통제 작업의 성격상 핵심 요소 기재가 엄격히 관리되기 때문이며, 온라인 자료의 표제 누락 집중은 그 규모와 이질성으로 인한 기재 통제의 느슨함에서 비롯된다. 온톨로지 설계 단계에서 자료 유형별 필수 요소를 차등 명세하지 않은 점이 이를 통제하지 못한 배경으로 작용하였다.

이상의 통합 해석을 바탕으로 품질 개선 시사점을 제시한다. 단기적으로는 동일 의미 속성 표기를 dc:terms 계열로 표준화하는 것이 실데이터의 실질 표준에 부합하므로 적용 부담이 적고 외부 재사용성을 즉각 개선한다. 중기적으로는 실데이터에서 광범위하게 사용하는 속성을 온톨로지 명세에 반영하고 정의역 누락을 보완하여 설계-실데이터 정합성을 확보해야 한다. 장기적으로는 데이터 생산과 온톨로지 관리를 연동하는 거버넌스 체계를 정비하여 BIBFRAME 전환을 준비하고, 품질 통제가 취약한 온라인 자료에 대해서는 핵심 요소 기재를 강제하는 검증 절차를 데이터 생산 단계에 통합하는 자료 유형별 차등 품질 관리가 필요하다. 본 연구의 핵심 발견들은 모두 단일 검증 도구나 단일 분석 층위로는 포착할 수 없는 것으로, SHACL · ShEx 이중 적용과 설계-실데이터 이중 층위 분석이라는 방법론은 여타 도메인의 대규모 링크드 데이터 품질 진단에도 적용 가능한 분석 틀로서 일반화 가능성을 지닌다.

## 8. 결론 및 제언

본 연구는 SHACL과 ShEx를 활용하여 국가서지 LOD의 품질을 설계 층위와 실데이터 층위의 양면에서 전수조사로 진단하였다. 첫 번째 연구 문제와 관련하여, 본 온톨로지는 29개 클래스와 157개 속성으로 이루졌으며, 전체 속성의 41.4%가 정의역을 명시하지 않는 명세 불완전성과 BIBFRAME 미반영, 코드성 분류의 SKOS 미활용이라는 설계 이슈를 안고 있었다. 두 번째 연구 문제와 관련하여, 국가서지 LOD는 SHACL 검증에서 99.73%의 높은 적합률을 보였으나 위반의 99.6%가 서지(온라인자료)의 표제 누락에 집중되었고, ShEx 검증에서 평균 미예상 속성 비율이 98.91%에 달하여 설계 명세가 실데이터의 표현 풍부성을 포괄하지 못함이 드러났다. 특히 설계에 부재하던 BIBFRAME 어휘가 실데이터에서 광범위하게 사용되는 비대칭을 확인하였으며, 이는 설계 단계의 명세 공백이 실데이터 단계의 무통제 사용으로 인과적으로 이어졌고 데이터 생산과 온톨로지 관리가 분리되어 운영됨을 시사한다.

본 연구는 국내 최초로 SHACL과 ShEx를 활용하여 국가 단위 서지 LOD의 품질을 전수조사로 진단하였다. 둘째, 설계 층위와 실데이터 층위를 통합하는 이중 층위 분석 틀을 제안하고 그 유효성을 입증하였다. 셋째, 두 검증 도구가 표제 표기 불일치라는 동일 사실을 독립적으로 가리킨 점에서 SHACL · ShEx 병용의 방법론적 삼각검증 효과를 실증하였다. 실무적으로는 속성 표기 표준화, 설계-실데이터 정합성 확보, BIBFRAME 전환 거버넌스 정비, 자료 유형별 차등 품질 관리를 제언하며, 본 연구의 형상과 분석 스크립트는 GitHub 저장소에 공개하여 지속적 품질 모니터링의 기반이 된다.

본 연구는 다음 한계를 지닌다. 첫째, 연역 형상이 KORMARC 필수 필드, RDA 핵심 요소, BIBFRAME 2.0 권장 요소 중심의 제한적 제약으로 설계되어 있어 미예상 속성 비율이 구조적으로 과대 산출될 가능성이 있으며, 온톨로지 전체 어휘 체계를 포괄하는 포괄적 연역 형상을 적용할 경우 해당 수치는 달라질 수 있다. 이는 4장에서 확인된 온톨로지 명세 불완전성에서 비롯된 현실적 제약이기도 하나, KORMARC 반복 필드・RDA Core Element 전체・BIBFRAME Application Profile・controlled vocabulary・identifier 패턴 등 다층적 기준을 적용한 형상 설계 수준별 민감도 분석 및 비교 실험이 후속 연구로 수행된다면 본 연구 결과의 해석 범위와 설득력을 한층 강화할 수 있을 것이다. 둘째, ShEx 귀납 형상의 카디널리티 산정이 JSON-LD 직렬화 특성의 영향을 받아, 카디널리티 일치율 지표는 데이터 품질보다 직렬화와 추론 알고리즘의 상호작용을 반영하는 측면이 있다. 셋째, 특정 시점의 데이터를 분석한 단면 연구로 품질의 시계열적 변화는 다루지 못하였다. 넷째, 단일 사례에 한정되어 분석 틀의 일반화 가능성은 논리적으로 논증했을 뿐 경험적으로 검증되지 않았다. 이에 따라 의미 정확성 심층 검증, 설계-실데이터 정합성 개선 실행 연구, BIBFRAME 2.0・IFLA LRM 매핑 연구, 이중 층위 분석 틀의 타 도메인 일반화 검증을 후속 연구로 제언한다. 본 연구가 드러낸 설계-실데이터의 괴리는 국가 단위 서지 인프라가 표준의 진화와 운영의 안정성 사이에서 직면하는 보편적 긴장의 구체적 발현이며, 본 연구의 진단과 제언이 국가서지 LOD의 품질 고도화와 차세대 서지 환경으로의 전환에 실증적 토대를 제공할 수 있기를 기대한다.

## 참 고 문 헌

노지현 (2019). 편목의 관점에서 본 링크드 데이터: 현황과 과제. 한국도서관・정보학회지, 50(3), 71-95. http://dx.doi.org/10.16981/kliss.50.201909.71

박옥남 (2012). 기록물 전거통제 기반 Linked Data 구축에 대한 연구. 한국비블리아학회지, 23(2), 5-25.

박옥남, 오정선 (2014). 링크드 데이터 환경에서의 서지기술형식 BIBFRAME과 그 활용에 대한 고찰. 한국비블리아학회지, 25(4), 235-263. http://dx.doi.org/10.14699/kbiblia.2014.25.4.235

박지영 (2012). 링크드 데이터 방식을 통한 서지 정보의 확장에 관한 연구. 정보관리학회지, 29(1), 231-251. http://doi.org/10.3743/KOSIM.2012.29.1.231

박지영 (2024). 법령 기반 분류체계의 유형 분석을 통한 BRM 기반 기록분류 개선 방안 연구. 한국기록관리학회지, 24(2), 139-163. http://doi.org/10.14404/JKSARM.2024.24.2.139

박진호, 곽승진 (2021). 링크드 데이터 기반 근대문학자료의 서비스 방안 연구. 한국문헌정보학회지, 55(2), 5-24. http://dx.doi.org/10.4275/KSLIS.2021.55.2.005

변건우, 김혜진 (2026). 주경로 분석을 이용한 링크드 데이터 연구의 지적 확산 궤적 분석. 한국도서관·정보학회지, 57(1), 387-411. http://dx.doi.org/10.16981/kliss.57.1.202603.387

이문호, 최성필 (2017). 극대용량 서지 링크드 데이터 구축의 효율성을 위한 트리플 저장소 접근 최소화에 관한 연구. 한국도서관·정보학회지, 48(3), 233-257. http://dx.doi.org/10.16981/kliss.48.3.201709.233

이미화, 송민선, 박진호 (2026). BIBFRAME 구축을 위한 교육 프로그램 개발에 관한 연구. 한국도서관·정보학회지, 57(1), 1-19. http://dx.doi.org/10.16981/kliss.57.1.202603.1

이성숙 (2022). 해외 도서관 링크드 데이터 구축의 최근 동향 연구: 발행 데이터세트, 재사용 어휘집, 인터링킹 외부 데이터세트를 중심으로. 한국문헌정보학회지, 56(4), 5-28. http://dx.doi.org/10.4275/KSLIS.2022.56.4.005

이성숙, 박지영, 이혜원 (2017). 링크드 데이터에서 인물 정보의 식별 및 연계 범위 확장에 관한 연구: 국립중앙도서관 링크드 데이터를 중심으로. 정보관리학회지, 34(3), 7-21. http://dx.doi.org/10.3743/KOSIM.2017.34.3.007

이유진, 양성권, 송민아, 김홍기 (2009). 시맨틱 디지털도서관 서비스를 위한 서지 온톨로지 구축. 정보관리학회지, 26(1), 215-230. http://doi.org/10.3743/KOSIM.2009.26.1.215

조명대 (2010). 도서관에서의 Linked Data 활용방안에 관한 연구. 한국문헌정보학회지, 44(1), 181-198. http://doi.org/10.4275/KSLIS.2010.44.1.181

Ahmetaj, S., Boneva, I., Hidders, J., Hose, K., Jakubowski, M., Labra Gayo, J. E., Pieris, A., Prud'hommeaux, E., Record, J., Sherkhonov, E., & Tomaszuk, D. (2025). Common foundations for SHACL, ShEx, and PG-schema. In Proceedings of the ACM on Web Conference 2025, 8-21. https://doi.org/10.48550/arXiv.2502.01295

BIBFRAME Interoperability Group (2024). BIG 2024 work plan: tabular application profiles, DCTap/SHACL validation, and data exchange test project. Program for Cooperative Cataloging, Library of Congress. Available: https://bf-interop.github.io/DCTap/

Candela, G., Escobar, P., Sáez, M. D., & Marco-Such, M. (2023). A Shape Expression approach for assessing the quality of Linked Open Data in libraries. Semantic Web, 14(2), 159-179. https://doi.org/10.3233/SW-210441

Cortés, C., Ehrlinger, L., Etcheverry, L., & Naumann, F. (2025). Is SHACL suitable for data quality assessment? arXiv preprint arXiv:2507.22305. https://doi.org/10.48550/arXiv.2507.22305

Debattista, J., Auer, S., & Lange, C. (2016). Luzzu—a methodology and framework for linked data quality assessment. Journal of Data and Information Quality (JDIQ), 8(1), 1-32. https://doi.org/10.1145/2992786

Duan, X., Chaves-Fraga, D., Derom, O., & Dimou, A. (2024). SCOOP all the Constraints' Flavours

for your Knowledge Graph. In European Semantic Web Conference, 217-234. https://doi.org/10.1007/978-3-031-60635-9_13

Fernandez-Álvarez, D., Labra-Gayo, J. E., & Gayo-Avello, D. (2022). Automatic extraction of shapes using sheXer. Knowledge-Based Systems, 238, 107975. https://doi.org/10.1016/j.knosys.2021.107975

Gaitanou, P., Andreou, I., Sicilia, M. A., & Garoufallou, E. (2024). Linked data for libraries: creating a global knowledge space, a systematic literature review. Journal of Information Science, 50(1), 204-244. https://doi.org/10.1177/01655515221084645

Király, P. (2024a). QA Catalogue: a quality assessment tool for library catalogues. GWDG Nachrichten, 2024(4-5), 19-24.

Király, P. (2024b). Shacl4Bib: Custom validation of library data, arXiv, arXiv:2405.09177. https://doi.org/10.48550/arXiv.2405.09177

Sejdiu, G., Rula, A., Lehmann, J., & Jabeen, H. (2019). A scalable framework for quality assessment of RDF datasets. In International Semantic Web Conference, 261-276. https://doi.org/10.1007/978-3-030-30796-7_17

Spahiu, B., Porrini, R., Palmonari, M., Rula, A., & Maurino, A. (2016). ABSTAT: ontology-driven linked data summaries with pattern minimalization. In European Semantic Web Conference, 381-395.

Thornton, K., Solbrig, H., Stupp, G. S., Labra Gayo, J. E., Mietchen, D., Prud'Hommeaux, E., & Waagmeester, A. (2019). Using shape expressions (ShEx) to share RDF data models and to guide curation with rigorous validation. In European Semantic Web Conference, 606-620. https://doi.org/10.1007/978-3-030-21348-0_39

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: a survey. Semantic Web, 7(1), 63-93. https://doi.org/10.3233/SW-150175

Zhang, S., Benis, N., & Cornet, R. (2023). Automated approach for quality assessment of RDF resources. BMC Medical Informatics and Decision Making, 23(Suppl 1), 90. https://doi.org/10.1186/s12911-023-02182-8

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

Byun, Keonwoo & Kim, Hye-Jin (2026). Intellectual diffusion trajectories of linked data research:

a main path analysis. Journal of Korean Library and Information Science Society, 57(1), 387-411. http://dx.doi.org/10.16981/kliss.57.1.202603.387

Cho, Myungdae (2010). A study on applications for linked data in libraries. Journal of the Korean Society for Library and Information Science, 44(1), 181-198. http://doi.org/10.4275/KSLIS.2010.44.1.181

Lee, Mihwa, Song, Min Sun, & Park, Jin Ho (2026). A study on the development of a training program for BIBFRAME implementation. Journal of Korean Library and Information Science Society, 57(1), 1-19. http://dx.doi.org/10.16981/kliss.57.1.202603.1

Lee, Moon-Ho & Choi, Sung-Pil (2017). Research on minimizing access to RDF triple store for efficiency in constructing massive bibliographic linked data. Journal of Korean Library and Information Science Society, 48(3), 233-257. http://dx.doi.org/10.16981/kliss.48.3.201709.233

Lee, Sung-Sook (2022). A study on recent trends in building linked data for overseas libraries: focusing on published datasets, reused vocabulary, and interlinked external datasets. Journal of the Korean Society for Library and Information Science, 56(4), 5-28. http://dx.doi.org/10.4275/KSLIS.2022.56.4.005

Lee, Sungsook, Park, Ziyoung, & Lee, Hyewon (2017). Expanding the scope of identifying and linking of personal information in linked data: focusing on the linked data of National Library of Korea. Journal of the Korean Society for Information Management, 34(3), 7-21. http://dx.doi.org/10.3743/KOSIM.2017.34.3.007

Lee, You-Jin, Yang, Sungkwon, Song, Mina, & Kim, Hong-Gee (2009). Implementing bibliographic metadata model for social semantic digital libraries. Journal of the Korean Society for Information Management, 26(1), 215-230. http://doi.org/10.3743/KOSIM.2009.26.1.215

Park, Jin Ho & Kwak, Seung Jin (2021). A study on the service method of modern literature based on linked data. Journal of the Korean Society for Library and Information Science, 55(2), 5-24. http://dx.doi.org/10.4275/KSLIS.2021.55.2.005

Park, Ok Nam & Oh, Jung Sun (2014). Deployment of BIBFRAME as a new bibliographic framework in linked data. Journal of the Korean Biblia Society for Library and Information Science, 25(4), 235-263. http://dx.doi.org/10.14699/kbiblia.2014.25.4.235

Park, Ok Nam (2012). The design and development of linked data from authority data in National Archives of Korea, Journal of the Korean Biblia Society for Library and Information Science, 23(2), 5-25.

Park, Ziyoung (2012). Extending bibliographic information using linked data. Journal of the

Korean Society for Information Management, 29(1), 231-251.
http://doi.org/10.3743/KOSIM.2012.29.1.231

Park, Ziyoung (2024). Improving the records classification system based on the Business Reference Model (BRM) through an analysis of legislative classification system types. Journal of Korean Society of Archives and Records Management, 24(2), 139-163.
http://doi.org/10.14404/JKSARM.2024.24.2.139

Rho, Jee-Hyun (2019). The current state and challenges of linked data in library cataloging. Journal of Korean Library and Information Science Society, 50(3), 71-95.
http://dx.doi.org/10.16981/kliss.50.201909.71

# [부록 1] 분석 도구 및 재현 환경

본 연구의 모든 분석은 Python 3.13 환경에서 수행되었으며, 분석에 사용한 스크립트는 GitHub 저장소(https://github.com/jhphansung/shacl-shex-national-bibliographic-lod)에 공개하였다. 각 스크립트의 기능은 다음과 같다.

| 스크립트 | 기능 |
|---|---|
| 01_데이터규모진단.py | 6개 데이터셋의 파일 수·용량·레코드 수·클래스별 인스턴스 수·속성별 출현 빈도를 전수 집계 (논문 6.1절) |
| 02_SHACL검증.py | 클래스별 NodeShape 기반 SHACL 전수 검증. 적합률·위반 유형·속성별 위반 빈도 산출 (논문 6.4절) |
| 03_ShEx검증.py | rdflib 기반 귀납 ShEx 형상 추출 및 연역 형상과의 비교. 속성 일치율·카디널리티 일치율·미예상 속성 비율 산출 (논문 6.5절) |

분석 환경의 재현을 위해, 검증 대상 데이터는 2026년 4월 1일자 국립중앙도서관 국가서지 LOD 벌크 데이터(JSON-LD 형식)를 사용하였음을 명시한다. ShEx 귀납 형상 추출 시 속성 채택 임계값은 0.1을 적용하였다.

# [부록 2] SHACL 검증 결과: 클래스별 적합률

실데이터에 출현한 28종 클래스에 대한 SHACL 검증 결과는 다음과 같다. 인스턴스 수 기준 내림차순으로 정렬하였다.

| 클래스 | 전체 인스턴스 수 | 위반 노드 수 | 적합률(%) |
|---|---|---|---|
| nlon:OnlineMaterial | 21,810,815 | 90,663 | 99.58 |
| nlon:ElectronicJournal | 16,901,548 | 332 | 100.00 |
| nlon:OfflineMaterial | 7,144,183 | 634 | 99.99 |
| nlon:Book | 6,845,116 | 632 | 99.99 |
| bibo:Book | 3,588,544 | 236 | 99.99 |
| nlon:ElectronicBook | 2,895,364 | 79,487 | 97.25 |
| bibo:Thesis | 1,869,680 | 154 | 99.99 |
| nlon:Author | 1,618,009 | 0 | 100.00 |
| bibo:Article | 1,200,240 | 7 | 100.00 |
| bibo:Website | 698,310 | 10,579 | 98.49 |
| nlon:Sound | 626,033 | 30 | 100.00 |
| nlon:Concept | 515,412 | 8 | 100.00 |
| nlon:NonBook | 299,070 | 2 | 100.00 |
| bibo:Image | 246,414 | 195 | 99.92 |
| nlon:VideoDocument | 227,144 | 0 | 100.00 |
| nlon:DigitalizedScore | 104,007 | 2 | 100.00 |
| nlon:Score | 103,492 | 0 | 100.00 |
| nlon:OldBook | 101,452 | 0 | 100.00 |
| bibo:AudioDocument | 91,992 | 2 | 100.00 |
| bibo:Periodical | 85,200 | 235 | 99.72 |
| bibo:AudioVisualDocument | 71,662 | 0 | 100.00 |
| nlon:Library | 28,108 | 0 | 100.00 |
| nlon:ElectronicDocument | 21,143 | 0 | 100.00 |
| nlon:AlternativeMaterial | 15,765 | 0 | 100.00 |
| nlon:Map | 10,677 | 0 | 100.00 |
| nlon:DigitalizedMap | 3,382 | 1 | 99.97 |
| nlon:Software | 17 | 1 | 94.12 |
| nlon:ComplexDocument | 3 | 0 | 100.00 |

## [부록 3] ShEx 검증 결과: 연역 형상과 귀납 형상의 비교

5장의 연역 형상과 실데이터로부터 추출한 귀납 형상의 비교 결과는 다음과 같다. 실데이터에 출현한 클래스를 인스턴스 수 기준 내림차순으로 정렬하였다.

| 형상 | 인스턴스 수 | 연역 속성 | 귀납 속성 | 속성일치(%) | 미예상(%) |
|---|---|---|---|---|---|
| nlon:OnlineMaterial | 21,810,815 | 1 | 36 | 0.0 | 100.0 |
| nlon:ElectronicJournal | 16,901,548 | 1 | 36 | 0.0 | 100.0 |
| nlon:OfflineMaterial | 7,144,183 | 1 | 43 | 0.0 | 100.0 |
| nlon:Book | 6,845,116 | 5 | 43 | 40.0 | 95.35 |
| bibo:Book | 3,588,544 | 1 | 38 | 0.0 | 100.0 |
| nlon:ElectronicBook | 2,895,364 | 1 | 39 | 0.0 | 100.0 |
| bibo:Thesis | 1,869,680 | 1 | 33 | 0.0 | 100.0 |
| nlon:Author | 1,618,009 | 2 | 15 | 100.0 | 86.67 |
| bibo:Article | 1,200,240 | 1 | 21 | 0.0 | 100.0 |
| bibo:Website | 698,310 | 1 | 24 | 0.0 | 100.0 |
| nlon:Sound | 626,033 | 1 | 31 | 0.0 | 100.0 |
| nlon:Concept | 515,412 | 1 | 8 | 100.0 | 87.5 |
| nlon:NonBook | 299,070 | 1 | 34 | 0.0 | 100.0 |
| bibo:Image | 246,414 | 1 | 33 | 0.0 | 100.0 |
| nlon:VideoDocument | 227,144 | 1 | 30 | 0.0 | 100.0 |
| nlon:DigitalizedScore | 104,007 | 1 | 19 | 0.0 | 100.0 |
| nlon:Score | 103,492 | 1 | 26 | 0.0 | 100.0 |
| nlon:OldBook | 101,452 | 1 | 26 | 0.0 | 100.0 |
| bibo:AudioDocument | 91,992 | 1 | 31 | 0.0 | 100.0 |
| bibo:Periodical | 85,200 | 1 | 33 | 0.0 | 100.0 |
| bibo:AudioVisualDocument | 71,662 | 1 | 36 | 0.0 | 100.0 |
| nlon:Library | 28,108 | 1 | 13 | 0.0 | 100.0 |
| nlon:ElectronicDocument | 21,143 | 1 | 42 | 0.0 | 100.0 |
| nlon:AlternativeMaterial | 15,765 | 1 | 20 | 0.0 | 100.0 |
| nlon:Map | 10,677 | 1 | 34 | 0.0 | 100.0 |
| nlon:DigitalizedMap | 3,382 | 1 | 24 | 0.0 | 100.0 |
| nlon:Software | 17 | 1 | 29 | 0.0 | 100.0 |
| nlon:ComplexDocument | 3 | 1 | 29 | 0.0 | 100.0 |

속성일치(%)는 연역 형상의 속성 중 귀납 형상에도 출현한 속성의 비율이며, 미예상(%)은 귀납 형상의 속성 중 연역 형상에 규정되지 않은 속성의 비율이다. foaf:Person과 foaf:Organization은 실데이터에서 nlon:Author 등 하위 어휘로 표현되어 직접 인스턴스가 출현하지 않았다.

# [부록 4] SHACL 형상 중 일부

본 연구에서 설계한 SHACL 형상의 대표 사례를 제시한다. 전체 형상은 GitHub 저장소에 shacl_shapes.ttl 파일로 공개된다. 동일 의미 속성의 표기 분산을 sh:alternativePath로 통합 처리한 점이 특징이다.

```
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix nlon: <http://lod.nl.go.kr/ontology/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix ex: <http://example.org/nlk-shapes/> .

# 표제 필수 제약 (표기 분산 통합)
ex:TitleConstraint a sh:PropertyShape ;
    sh:path [ sh:alternativePath (
        dc:title dcterms:title ) ] ;
    sh:minCount 1 ;
    sh:name "표제" ;
    sh:message "표제(title)가 누락되었습니다." .

# 도서 형상
ex:BookShape a sh:NodeShape ;
    sh:targetClass nlon:Book ;
    sh:property ex:TitleConstraint ;
    sh:property [
        sh:path nlon:kdc ;
        sh:datatype xsd:string ;
        sh:pattern "^[0-9]{3}(\\.[0-9]+)?$" ;
        sh:severity sh:Warning ;
        sh:name "한국십진분류" ; ] .

# 저자 형상
ex:AuthorShape a sh:NodeShape ;
    sh:targetClass nlon:Author ;
    sh:property [
        sh:path foaf:name ;
        sh:minCount 1 ;
        sh:name "성명" ; ] .
```

## [부록 5] 귀납 ShEx 형상 중 일부

실데이터로부터 추출한 귀납 ShEx 형상의 대표 사례를 제시한다. 각 속성 옆의 출현율은 해당 클래스 인스턴스 중 그 속성이 출현한 비율이다. 전체 형상은 GitHub 저장소에 shex_inductive_shapes.shex 파일로 공개한다.

```
# 국가서지 LOD 귀납 ShEx 형상 (실데이터 전수 집계 기반)
# 생성일: 2026-05-18
# 속성 채택 임계값: 0.1

# nlon:OnlineMaterial (인스턴스 21,810,815건)
<nlon:OnlineMaterial> {
    nlon:datePublished + # 출현율 100.0%
    rdfs:label * # 출현율 99.95%
    nlon:genre * # 출현율 99.87%
    dcterms:title * # 출현율 99.58%
    dcterms:accessRights * # 출현율 99.55%
    nlon:audienceNote * # 출현율 97.79%
    dc:publisher * # 출현율 96.99%
    nlon:typeOfResource ? # 출현율 95.98%
    nlon:medium * # 출현율 93.53%
    nlon:antecedentSource ? # 출현율 93.25%
    dcterms:issued * # 출현율 93.12%
    nlon:publicationPlace * # 출현율 90.26%
    dc:format * # 출현율 88.31%
    bibframe:extent * # 출현율 87.71%
    nlon:holdingInstitution ? # 출현율 82.05%
    nlon:titleOfHostItem * # 출현율 77.8%
    nlon:relatedParts * # 출현율 63.98%
    nlon:newsPosition * # 출현율 60.11%
    dcterms:description * # 출현율 59.33%
    nlon:keyword * # 출현율 56.68%
    nlon:uci * # 출현율 56.65%
    dcterms:hasFormat * # 출현율 50.31%
```

```
    nlon:reproductionNote * # 출현율 49.69%
    nlon:otherNumber ? # 출현율 43.16%
    dc:creator * # 출현율 37.2%
    nlon:kdc * # 출현율 30.32%
    dcterms:identifier * # 출현율 28.23%
    nlon:kdcn * # 출현율 21.39%
    dcterms:creator * # 출현율 20.24%
    dcterms:subject * # 출현율 18.54%
    nlon:restriction * # 출현율 18.15%
    dcterms:abstract * # 출현율 17.59%
    dcterms:alternative * # 출현율 14.92%
    nlon:languageNote * # 출현율 12.58%
    dcterms:isPartOf * # 출현율 12.48%
    nlon:remainderOfTitle * # 출현율 10.85%
}

# nlon:ElectronicJournal (인스턴스 16,901,548건)

(이하 생략. 전체 31종 클래스의 귀납 형상은 GitHub 저장소 참조.)
```