국내 유통 전자출판물의 납본 및 수집을 위한 데이터 요구사항 및 품질 검증 연구*

A Study on Data Requirements and Quality Verification for Legal Deposit and Acquisition Tasks of Domestic Electronic Publications

김 규 환 (Gyuhwan Kim)**
김 수 정 (Soojung Kim)***
정 대 근 (Daekeun Jeong)****

초 록

본 연구는 국내 유통 전자출판물의 납본 및 수집을 위한 데이터의 속성과 속성값의 표준화 방안과 정책 및 제도적 고려사항을 제시하고자 하였다. 연구 결과, 필수 및 선택 속성은 총 21개가 도출되었으며, 이는 국립중앙도서관 납본 및 수집 업무 담당자들의 설문조사 및 FGI 결과를 바탕으로 선정되었다. 데이터 품질 검증 과정에서 추가적으로 필요한 속성이 발견되어, 전자책, 오디오북, 웹툰, 웹소설 등 자료유형별로 필수 및 선택 속성을 구체화하였다. 속성값의 표준화는 ISO 8601 규칙에 따른 날짜 및 시간의 표기, 파일 형식과 성인 여부 등 제한된 범위의 속성값의 명확한 지정, 제목과 관련된 정보의 상세한 기술 등을 포함하였다. 정책 및 제도적 고려사항은 표준화된 메타데이터 요구사항의 확립, 지속적인 데이터 품질 관리 및 모니터링 체계의 구축의 필요성을 제시하였다.

ABSTRACT

This study aimed to propose considerations for attributes and their standardization strategies during the data collection process for electronic publications by domestic distributors for the National Library of Korea. The research identified a total of 21 essential and optional attributes based on a survey and a Focused Group Interview (FGI) with the staff responsible for legal deposit and acquisition tasks at the National Library of Korea. Additional attributes were found necessary during the data quality verification process, leading to the specification of essential and optional attributes for various types of materials, including eBooks, audiobooks, webtoons, and web novels. The standardization of attribute values, essential for enhancing the identifiability and management efficiency of electronic publications, included adherence to ISO 8601 rules for dates and times, clear designation of limited-range attribute values such as file format and adult content, and detailed description of information related to titles. Furthermore, the study highlighted the need for establishing standardized metadata requirements and continuous data quality management and monitoring systems.

키워드: 전자출판물, 납본, 수집, 데이터 요구사항, 데이터 품질 검증, 국립중앙도서관
Electronic Publications, Legal Deposit, Acquisition, Data Requirements, Data Quality Verification, National Library of Korea

---

* 본 연구는 2023년도 국립중앙도서관 '국내 전자출판물 통계조사 기초 연구'의 일부분을 수정・보완한 것임.
** 인천대학교 문헌정보학과 부교수(gyuhwan@inu.ac.kr) (제1저자)
*** 전북대학교 문헌정보학과 교수, 문화융복합아카이빙연구소 연구원(kimsoojung@jbnu.ac.kr) (공동저자)
**** 광주대학교 문헌정보학과 조교수(dkjeong@gwangju.ac.kr) (교신저자)
논문접수일자 : 2024년 2월 19일 논문심사일자 : 2024년 2월 21일 게재확정일자 : 2024년 3월 6일
한국비블리아학회지, 35(1): 127-148, 2024. http://dx.doi.org/10.14699/kbiblia.2024.35.1.127

1. 서 론

1.1 연구 배경과 목적

전자출판물은 전자적 매체를 통해 정보를 전달하는 출판 형태로, 이용자가 컴퓨터와 같은 정보 처리 장치를 사용하여 내용을 읽거나, 보거나, 들을 수 있게 하는 간행물을 의미한다(출판문화산업 진흥법, 제2조(정의) 제4호). 한국출판문화산업진흥원은 매체 특성을 기준으로 전자출판물을 전자책, 앱북, 오디오북, 웹소설, 웹툰으로 구분하고 있다(한국출판문화산업진흥원, 2021). 전자책은 전통적인 종이책의 내용을 디지털 형식으로 변환한 것으로, 다양한 기기를 통해 접근할 수 있다. 앱북은 애플리케이션 형태의 전자책으로, 이용자 경험을 풍부하게 하는 다양한 멀티미디어 요소를 포함한다. 오디오북은 텍스트 기반 콘텐츠를 오디오 형태로 제공하는데, 이는 독자에게 새로운 청취 경험을 제공한다. 웹소설은 온라인 플랫폼을 통해 연재되는 소설이며, 웹툰은 디지털 만화로, 멀티미디어 요소와 함께 스크롤 형식의 독특한 읽기 경험을 제공한다. 이러한 다양한 형태의 전자출판물은 디지털 시대의 이용자들에게 새로운 형식과 경험을 제공해 주고 있다.

출판업계는 디지털 기술의 보편화에 따라 종이 형식의 전통 출판에서 디지털 형식의 출판으로 변모하고 있다. 이런 변화는 종이책 판매의 감소와 함께 전자책 매출의 지속적인 상승으로 명확히 드러나고 있다. 2019년의 전자책 시장 매출은 약 4천억 원으로 추정되며, 전자책 ISBN 신청 건수는 종이책을 넘어서고 있다. 특히, 웹소설과 웹툰 산업의 급성장이 이러한 전환을 더욱 가속화시키고 있다. 2020년 대비 2021년 웹툰 시장의 증가 추세와 함께 오디오북과 챗봇 등 새로운 디지털 콘텐츠 형태의 등장이 관측되고 있다. 또한, 전자책, 웹소설, 웹툰, 오디오북 등 다양한 전자출판물의 유통량이 꾸준히 증가하고 있는 추세이다(김규환, 정대근, 김수정, 2023).

이러한 상황에서 국립중앙도서관은 도서관법과 내부 지침에 따라 전자출판물을 지속적으로 납본 및 수집해 오고 있다. 국제표준도서번호(이하 ISBN)를 받은 전자책, 웹소설, 웹툰, 오디오북의 전자출판물은 납본의 대상이며, ISBN을 미발급받았으나 보존가치가 높은 전자책, 웹소설, 웹툰, 오디오북의 전자출판물은 자체수집의 대상이 된다. 최근 3년간 국립중앙도서관의 전자출판물 납본 및 수집 현황에 대한 통계조사 결과를 보면, 납본 방식으로 입수된 전자출판물은 41.74%, 자체 수집 방식으로 입수된 전자출판물은 57.07%인 것으로 나타났다. 납본과 자체수집별 전자출판물의 유형을 보면 자체수집된 자료유형은 대부분이 웹툰이며 납본된 자료유형은 주로 전자책(웹소설 완결본 포함)인 것으로 조사되었다. 한편, 최근 3년간 ISBN을 발급받은 전자출판물의 납본율은 9.5%로 매우 저조한 것으로 나타났다(국립중앙도서관, 2023). 이 통계조사를 통해 전자출판물의 납본과 수집 현황을 파악할 수 있다. 다만 이 통계조사는 국립중앙도서관의 내부 데이터베이스를 기반으로 하였기에 국내 전자출판물 시장에서 실제 유통되고 있는 전자출판물 현황을 충분히 반영하지 못한 한계점이 있다.

이에 국립중앙도서관에서는 국내에서 실제

유통되는 전자출판물에 대한 데이터의 수집과 분석에 대한 필요성을 인식하고 있다. 하지만 실제 출판 시장에서 유통되는 전자출판물 데이터의 수집 및 분석은 현실적으로 많은 어려움을 가지고 있다. 특히 유통사별로 전자출판물 데이터 관리 방식이 서로 달라서, 국립중앙도서관이 여러 유통사들의 데이터를 수집하여 분석하는 것 자체가 큰 과제가 되고 있다. 이에 본 연구는 국립중앙도서관 납본 및 수집 업무담당자 관점에서 유통사 전자출판물 데이터 수집 시 요구사항을 파악하고 실제 유통사가 관리하고 있는 데이터의 품질을 검증하려 한다. 이를 통해 국립중앙도서관이 국내 유통 전자출판물 데이터를 수집 및 통합하는 초기단계에 속성 및 속성값의 표준화 방안과 정책 및 제도적 고려사항을 제안하고자 한다. 본 연구에서 설정한 연구문제는 다음과 같다.

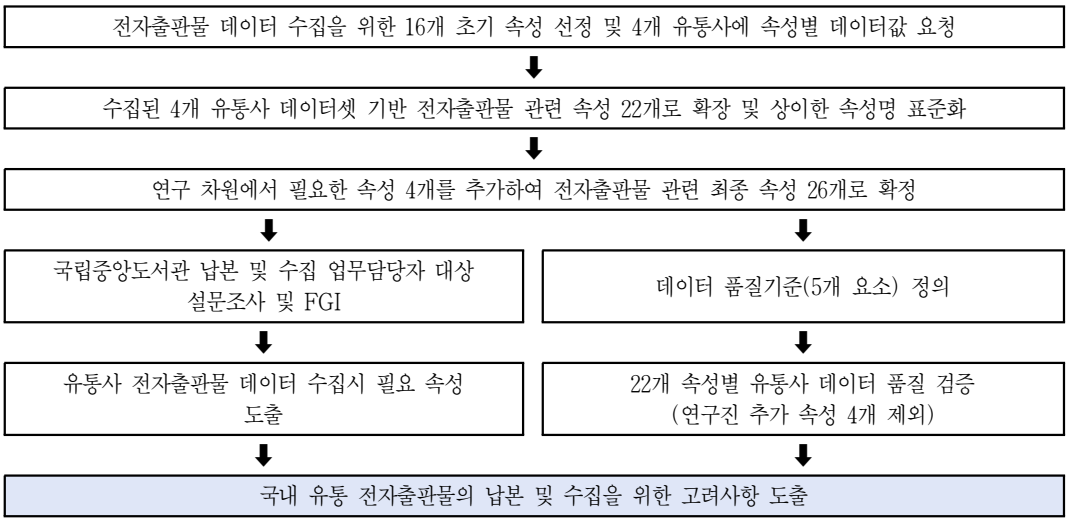
- 연구문제1: 국립중앙도서관이 유통사 전자출판물 데이터 수집시 필요한 속성은 무엇인가?
- 연구문제2: 국내 유통사가 관리하고 있는 전자출판물 데이터의 속성값 품질은 어떠한가?

1.2 연구내용과 방법

본 연구의 연구내용과 방법은 다음과 같다. 첫째, 국립중앙도서관 업무담당자의 업무적 필요성을 토대로 전자출판물 데이터 수집을 위한 16개 초기 속성을 선정하고, 국내 전자출판물 유통사 중 4개를 선정하여 16개 속성에 대한 데이터값 작성을 요청하였다. 둘째, 4개 유통사에서 보내온 데이터셋에 초기 선정한 16개 속성보다 더 많은 속성들이 포함되어 있어 전자출판물 관련 속성을 16개에서 22개로 확장하였고 유통사별로 관리하고 있는 전자출판물의 속성명이 상이하여 속성명에 대한 표준화 작업을 진행하였다. 또한 국립중앙도서관 납본 및 수집 업무담당자 의견수렴을 위해 필요하다고 판단한 4개의 속성(유통사, 연재시작일, 연재종료일, 연재주기)들을 추가하여 최종 26개의 전자출판물 속성을 확정하였다. 셋째, 최종 확정된 26개 속성을 토대로 국립중앙도서관 납본 및 수집 업무담당자들에게 설문조사 및 FGI를 진행하였으며, 이를 통해 전자출판물 데이터 분석에 필요한 속성들을 도출하였다. 넷째, 4개 유통사에서 입수한 데이터셋을 대상으로 연구진에서 추가한 4개 속성을 제외한 22개 속성에 대해서, 5개 요소(완전성, 유일성, 유효성, 일관성, 정확성) 측면에서 데이터 품질 검증을 실시하였다. 이상의 연구 절차를 통해 국내 유통사 전자출판물 데이터 수집시 고려해야 할 사항을 도출하여 제안하였다. 연구내용과 방법을 도식화하면 〈그림 1〉과 같다.

2. 관련 선행연구

본 연구와 관련한 국내·외 선행연구들은 국내 전자출판 시장의 동향, 전자출판물에 대한 납본 및 수집 관련 연구, 전자출판물을 포함한 전자자원 수집 및 관리를 위한 메타데이터 요소 표준화를 다룬 연구들을 중심으로 살펴보고자 한다. 먼저, 국내 전자출판 시장의 동향과 관련한 연구의 경우에는 손애경(2012)이 한국



〈그림 1〉 연구내용과 방법

의 전자출판에 대한 기술 지원 정책의 현황을 검토하여, 문화기술(CT)이 이 분야에서 중요한 역할을 하고 있음을 강조하였다. 미디어 융합으로 인한 전자출판의 패러다임 변화를 논의하고, 새로운 출판산업 가치사슬에 적합한 문화기술 요소를 제안하였다. 연구는 또한 스마트폰과 태블릿 PC를 바탕으로 한 새로운 글쓰기 형태, 제작, 유통 및 독서 유형의 변화에 따른 트렌드 분석을 통해, 새로운 출판산업 가치사슬 구조에 적합한 문화기술 요소를 도출하였다. 또한 소셜미디어 및 개인출판 기술 동향을 분석하여 출판시장과의 연계성 및 관련 정책 지원 범위를 알아보고, 나아가 미래 클라우드 출판산업의 중장기 발전 방안을 모색하였다. 공병훈과 조정미(2021)는 해방 이후 한국 출판 기술의 발전 및 진화 단계를 분석하였다. 여기에는 한글 활자 개발, 인쇄소 설립 및 기술 도입 등 출판 인프라 구축 단계부터 시작하여, 디지털 기술을 출판 과정 전반에 활용하는 단계, 인터넷 및 디지털 기반 유통 환경에서의 인터넷 서점 및 전자책 서점의 활성화, 모바일 환경에서 플랫폼 기반 출판이 본격화되는 시기 등 다양한 단계를 포함하였다. 연구 결과, 한국 출판 생태계는 폐쇄적 가치사슬 체제에서 개방적 가치 네트워크 체제로 패러다임이 이동하는 질적 변화가 진행 중에 있었다. 단계적 진화에서 기술 발달과 혁신 과정이 핵심 동인으로 작용했으며, 비즈니스 모델과 시장 변화를 창출시켜 복잡도 높은 지형을 형성하는 요인으로 작용하며, 다양한 출판 행위자들은 변화되는 지형에서 지속 가능한 생태계 적응 전략을 통해 적합도를 높이는 최적의 방법을 찾아내고 있었다. 한국출판문화산업진흥원(2021)은 국내 전자출판 산업이 가파르게 성장하는 것에 비해 산업 진흥 정책 수립을 위한 통계자료나 연구 자료가 부족한 실정을 보완하고 빠르게 변화하는 전자출판 콘텐츠의 양상을 반영하지 못하는 현실을 개선하기 위해서 전자출판 산업에 대한

전반적인 현황을 파악하여 시장 활성화를 위한 전략 수립 방향을 제시하였다. 연구 결과, 미디어 기술의 발전과 더불어 코로나19라는 예기치 못한 글로벌 팬데믹이 겹치면서 전 세계적으로 온라인 중심으로 소비의 구조가 옮겨가고 있으며 출판시장 또한 이러한 흐름에 있는 것으로 분석되었다. 그리고 향후에도 국내 전자출판 시장은 오디오북 시장의 성장 잠재력, 폭발적인 웹소설과 웹툰 시장의 성장이 동반되어 향후 전자출판 시장은 빠른 속도로 성장할 것으로 전망하였다.

다음으로 전자출판물의 납본 및 수집과 관련한 연구의 경우에는 Gooding과 Terras(2020)가 "전자 법정 납본: 미래의 도서관 컬렉션을 형성하기"라는 주제로, 법정 납본 도서관이 최근 디지털화된 자료와 디지털 태생 자료를 수집하는 데 필요한 새로운 규정에 대해 다루었다. 법정 납본 도서관은 국가 및 학술 기관으로, 문화적 기록을 체계적으로 보존하는 역할을 한다. 이 연구는 전자 법적 납본이 도서관, 그리고 현재와 미래의 이용자들에게 미치는 영향에 대해 국제적인 전문가들의 의견을 모았으며 디지털 보존과 접근 문제에 대한 복잡한 이슈들을 다루었다. 다양한 국가의 주요 도서관과 연구자, 실무자들과의 협력을 통해 수집된 실제 사례 연구를 기반으로 하였고 도서관 이용자, 연구자, 출판사 등 다양한 이해관계자들의 관점을 고려하였다. 특히, 도서관과 출판사간의 협력을 통해 웹사이트부터 전자책까지 국립 전자출판물 컬렉션을 체계적으로 수집하고 보존하는 공동 노력을 제시하였다. 국내에서는 디지털 자료의 확산에 따라 디지털 자료의 납본 추진을 위한 납본 제도 및 납본시스템 개선에 관한 연구가 2000년대 초반부터 진행되었다(곽승진 외, 2008; 곽승진 외, 2013; 서혜란, 2003; 윤희윤, 2003; 이숙현, 2003; 장보성, 남영준, 2010; 최재황, 곽승진, 김정택, 2009; 한혜영, 2003). 윤희윤(2003)은 납본 법령이 전자출판 및 유통 환경을 반영하지 못한다는 문제를 지적하고, 이를 개선하기 위한 모형을 제안하였다. 서혜란(2003)과 이숙현(2003)은 디지털자료의 수집과 보존을 위한 납본 제도의 확대와 온라인 전자출판물을 납본 대상으로 규정하는 것을 제안하였다. 한혜영(2003)은 대부분 국가의 납본제도가 인쇄물과 오프라인 전자출판물 중심임을 지적하고 온라인 전자출판물 수집과 보존을 위한 납본제도와 시스템 구축의 필요성을 제시하였다. 이를 위해서 국내외 납본제도 변화를 분석하고 전자출판물 납본시스템 모델과 구축에 필요한 기술 및 관리 측면을 제시하였다. 장보성과 남영준(2010)은 전자책의 보존에 중점을 두고 납본제도 개선 방안을 제안하였다. 최재황, 곽승진, 김정택(2009)과 곽승진 외(2013)는 온라인 디지털자료의 납본체계와 이용에 관한 가이드라인을 제안하였다. 또한 곽승진 외(2008)는 디지털자료의 납본 보상 비용 기준을 제시하였다.

근래에는 김규환, 정대근, 김수정(2023)이 국립중앙도서관 납본 및 수집 데이터베이스를 기반으로 최근 3년간 전자출판물 납본 및 수집 현황을 분석하였다. 분석 결과, ISBN을 발급받은 전자출판물의 납본율은 9.8%로 매우 저조한 것으로 조사되었다. 이에 개선방안으로 납본 의무에 대한 인식 제고 노력과 다양한 인센티브 제공과 함께 제재 조치 강화, 납본 현황에 대한 정보 공개 및 공유를 통해 발행처들의 자발적 참

여를 유도할 것을 제안하였다. 또한 ISBN 발급 및 납본 데이터의 정확성 확보를 위한 기술적 조치가 필요하다고 하였다.

전자출판물을 포함한 전자자원 수집 및 관리를 위한 메타데이터 요소 표준화를 다룬 연구의 경우에는 하진희, 김성혁, 임순범(2003)이 전자책 서비스업체들이 제공하는 불충분한 메타데이터로 인해 도서관이 전자책 정보를 자동으로 카탈로그에 추가하는 데 어려움을 겪고 있음을 지적하였다. 이 연구는 다양한 메타데이터 표준(예: KS X 6100, 더블린코어, MARC, TEI Header) 간의 호환성과 상호운용성을 확보하기 위해 전자책 라이브러리를 위한 메타데이터를 개발하였다. 연구 결과, 공통의 메타데이터 요소들을 핵심 기술 요소로 정의하고 전자책의 고유한 특성을 나타내는 메타데이터 요소를 상세 및 추가 기술 요소로 정의하였다. 남영준과 장보성(2006)은 전자자원의 양적 급증으로 인해 전통적인 수서정책이나 관리시스템만으로는 효과적인 관리가 어려워진 도서관의 문제를 다루었다. 이 연구에서는 전자자원의 효율적 관리와 활용을 위한 필수 메타데이터 요소를 제안하고 이를 활용한 효율적인 관리방안을 제안하였다. 김정명과 박찬수(2022)는 한국과 일본의 출판물 메타데이터 활용과 운영 방식을 비교・분석하였다. 한국에서는 출판유통통합전산망을 통해 출판 산업 통계와 도서 대출 통계 등을 제공하는 반면, 일본은 출판 예정 일정을 포함한 종이책, 전자책, 오디오북에 대한 정보를 제공하고 있다. 이 연구에서는 한국의 출판유통시스템에서 ISBN과의 실질적 연계, 출판물 메타데이터 등록비의 유료화, 출판유통통합전산망의 공공・민간 운영에 대해서 집중 논의하였다. 이들 연구들은 전자출판물과 관련된 메타데이터 관리의 다양한 측면을 탐구하였고 전자책 라이브러리와 전자자원 관리에 필요한 메타데이터 요소들을 분석하였다.

이상과 같이 국내 전자출판 시장 동향, 전자출판물의 납본 및 수집, 전자출판물의 데이터 속성을 다룬 국내외 선행연구들을 살펴보았다. 국내 선행연구 중에서는 김규환, 정대근, 김수정(2023)이 국내 전자출판물의 납본 및 수집 데이터 분석을 진행한 바가 있다. 그러나 국립중앙도서관 납본 및 수집 업무담당자들의 데이터 분석 요구사항을 파악하고 국내 유통사 전자출판물 데이터의 품질을 검증한 연구는 아직 수행되지 않았다. 이에 본 연구에서는 이를 보완하기 위한 연구를 진행하고자 한다.

3. 연구 설계

3.1 전자출판물 관련 초기 속성 선정 및 유통사 데이터 요청

4개 유통사에 전자출판물 데이터셋을 수집하기 위해 초기 요청 속성을 선정하였다. 선정 작업은 국립중앙도서관의 전자출판물 납본 및 수집 업무담당자의 업무적 필요성을 토대로 진행하였으며, 초기 속성은 〈표 1〉과 같이 16개가 선정되었다. 16개 속성들을 엑셀파일에 정리하여 4개 유통사 전자출판물담당자에게 이메일로 전송하였고, 해당 유통사가 관리하고 있는 전자출판물의 유형(전자책, 웹소설, 웹툰, 오디오북 등)별로 16개 속성에 해당하는 데이터 값을 채워서 보내줄 것을 요청하였다.

〈표 1〉 국립중앙도서관 업무담당자가 선정한 초기 분석 속성(16개)

- 구분(전자책, 웹소설, 웹툰, 오디오북)	- 소장가(정가)
- 제목	- 대여가(대여만 해당)
- 저자	- 연재정보/파일개수
- 낭독자(오디오북만 해당)	- 총재생시간(오디오북만 해당)
- 출판사	- 파일형태
- 출간일	- ISBN
- 유통사등록일	- ECN
- 현재판매여부(Y, N)	- 분야/분류

3.2 4개 유통사 데이터셋 기반 전자출판물 관련 속성 확장 및 속성명 표준화

4개 유통사에 전자출판물 관련 16개 속성에 대한 데이터값을 요청하였으나, 해당 유통사들은 16개 속성과 함께 추가적인 전자출판물 관련 속성들에 대한 데이터도 작성하여 보내왔다. 4개 유통사가 보내온 전자출판물 데이터셋을 살펴보면, 전자출판물 유형에 따라 새로운 속성들이 추가적으로 포함되어 있었고, 유통사별로 동일한 속성명이 상이하게 표기된 경우도 있었다. 또한 특정 유통사의 경우 전자출판물 관리를 위해 부여한 자체 식별번호(sale_cmdtid)가 활용되고 있었다. 구체적으로 살펴보면, 가장 차별성이 두드러지는 데이터셋은 D 유통사의 것으로, 웹툰 자료의 특성이 반영된 속성들이 존재하고(예, 최초회차등록일, 최신회차등록일, 성인여부), '저자' 속성을 '글작가', '그림작가', '원작가' 속성으로 세분화하였다는 특징이 있다. 반면에 오디오북을 유통하는 A, B, C 유통사는 모두 '낭독자' 속성을 제공하고 있고, 연재 중인 자료를 유통하는 C, D 유통사는 '파일개수'와 '완결여부' 속성을 제공하고 있었다. 이러한 결과는 유통사 전자출판물 데이터 수집시 전자출판물 유형에 상관없이 공통적으로 필요한 속성뿐만 아니라 자료유형별 고유 속성을 모두 포함하여 수집할 필요성이 있음을 시사한다. 이에 본 연구에서는 4개 유통사 데이터를 토대로 전자출판물 속성명 표준화 작업을 진행하였다(〈표 2〉 참조). 표준화 작업내용은 유통사들에서 이미 다수 사용하고 있는 속성명을 우선적으로 고려하되, 속성명을 단순화하거나(예, 소장가(정가) → 정가; 연재정보/파일개수 → 파일개수), 속성명의 의미가 불명확한 경우 좀 더 명확하게 수정하였고(예, 구분 → 자료유형; 파일형태 → 파일형식), 표준식별기호를 하나로 통합하였다(예, ISBN, ECN → ISBN/ECN). 그리고 A 유통사에서만 활용되는 sale-cmdtid는 미사용하였다.

3.3 전자출판물 관련 최종 속성 확정

본 연구에서는 '유통사', '연재시작일', '연재종료일', '연재주기'의 4개 속성을 추가하였다. '유통사'의 경우에는 납본 및 수집 업무시 '유통사'가 중요한 속성일 수 있기 때문이며, '연재시작일', '연재종료일', '연재주기'는 웹툰과 웹소설의 납본 및 수집 시기를 예측하기 위해서 중요한 속성으로 판단하였기 때문이다. 본 연구에서 확정한 전자출판물 관련 최종 26개 속성은 〈표 3〉과 같다.

〈표 2〉 유통사 데이터의 속성명 표준화

A 유통사	B 유통사	C 유통사	D 유통사	표준화
sale_cmdtid	-	-	-	미사용
구분	구분	구분	-	자료유형
제목	제목	제목	제목	제목
-	부제	-	-	부제
저자	저자	저자	글작가	저자
-	-	-	그림작가	그림작가
-	-	-	원작가	원작가
낭독자	낭독자	낭독자	-	낭독자
전자책 정가	소장가(정가)	소장가(정가)	-	정가
회차상품정가	대여가	대여가	-	대여가
출판사	출판사	출판사	-	출판사
ISBN	ISBN	ISBN	연재 ISBN	ISBN/ECN
-	ECN	ECN	-	
판매상태	현재판매여부	현재판매여부	서비스 상태	판매상태
총재생시간	총재생시간	총재생시간	-	총재생시간
분야/분류	분야/분류	분야/분류	속성장르	주제
파일형태	파일형태	파일형태	-	파일형식
등록일	유통사등록일	유통사등록일	-	유통사등록일
-	-	-	최초회차등록일	최초회차등록일
-	-	-	최신회차등록일	최신회차등록일
출간일	출간일	출간일	발행일	발행일
연재정보/ 파일개수	연재정보/ 파일개수	연재정보/ 파일개수	등록회차수	파일개수
완결여부	완결여부	완결여부	완결	완결여부
-	-	-	성인여부	성인여부

〈표 3〉 최종 확정된 전자출판물 관련 26개 속성 리스트

4개 유통사 데이터셋 기반 22개 속성			연구진에서 추가한 4개 속성
- 자료유형 - 제목 - 부제 - 저자 - 그림작가 - 원작가 - 낭독자 - 정가	- 대여가 - 출판사 - ISBN/ECN - 판매상태 - 총재생시간 - 주제 - 파일형식 - 유통사등록일	- 최초회차등록일 - 최신회차등록일 - 발행일 - 파일개수 - 완결여부 - 성인여부	- 유통사 - 연재시작일 - 연재종료일 - 연재주기

4. 전자출판물 납본 및 수집 업무시 필요 속성

4.1 설문조사 및 FGI

국립중앙도서관 납본 및 수집 업무담당자 5명을 대상으로 설문조사를 실시하였으며, 26개 분석 속성 중 국립중앙도서관 납본 및 수집 데이터 분석시 필요한 전자출판물 속성을 조사하였다. 또한, 전자출판물 데이터 분석시 필요한 속성에 대한 심층적 의견을 듣기 위해 중간관리자 및 업무담당자 등 6명을 대상으로 FGI를 진행하였다. 설문조사 및 FGI 대상과 조사 내용은 〈표 4〉와 같다.

4.2 조사결과

4.2.1 설문조사 결과

국립중앙도서관 납본 및 수집 업무담당자들을 대상으로 납본 및 수집 통계조사시 필요한 전자출판물 속성이 무엇인지 조사하였다. 조사는 전자책, 오디오북, 웹툰, 웹소설의 전자출판물 유형별로 26개 속성을 제시하고 각각의 속성의 필요여부를 체크하게 하였고 최종적으로 응답률을 산출하였다. 응답률이 60%일 경우에는 필요한 속성으로 간주하였다. 설문조사 결과는 〈표 5〉와 같다.

모든 전자출판물 유형에서 응답률이 60%인 속성들을 살펴보면, 제목, 부제, 저자, 자료유형, ISBN/ECN, 파일유형, 발행처, 주제, 발행일, 정가로 총 10개 속성으로 조사되었다. 이들 10개 속성들은 국내 유통사로부터 전자출판물 데이터를 수집할 때 우선적으로 고려되어야 할 속성으로 봐야 할 것이다. 각각의 전자출판물 유형별로 살펴보면, 전자책은 11개 속성, 오디오북은 13개 속성, 웹툰은 18개 속성, 웹소설은 15개 속성이 필요한 것으로 나타났다. 전자출판물 유형의 특성에 따라 추가적으로 수집되어야 할 속성들이 존재함을 알 수 있다.

4.2.2 FGI 인터뷰 결과

26개 속성에 대해서 국립중앙도서관 업무담당자를 대상으로 FGI를 진행하였고, 그 결과를 정리하여 제시하면 〈표 6〉과 같다.

FGI에 참여한 중간관리자 및 업무담당자의 주요한 의견을 살펴보면, 제목 관련 속성에서 '제목'은 통계분석의 대상은 아니나 필수 속성이라는데 이견이 없었으며, '부제'는 전자책의

〈표 4〉 설문조사 및 FGI 일자, 대상, 인터뷰 내용

구분	일자	대상	인터뷰 내용
설문조사	2023년 7월 7일~10일	국립중앙도서관 업무담당자 5명	- 모든 전자출판물 유형에서 공통적으로 필요한 속성 리스트 - 전자출판물 유형(전자책, 오디오북, 웹툰, 웹소설)별 필요한 속성 리스트
FGI	2023년 7월 13일	국립중앙도서관 중간관리자 및 설문조사 에 응답한 업무담당자 등 6명	- 전자출판물 데이터 분석시 필요한 속성에 대한 심층 의견 수렴

〈표 5〉 자료유형별 필요 속성에 대한 응답률

필드명	응답결과				필드명	응답결과			
	전자책	오디오북	웹툰	웹소설		전자책	오디오북	웹툰	웹소설
자료유형	100%	80%	80%	80%	주제	100%	100%	100%	100%
제목	80%	80%	80%	80%	파일형식	100%	100%	100%	100%
부제	60%	60%	60%	60%	유통사등록일	0%	0%	0%	0%
저자	80%	60%	80%	80%	최초회차등록일	0%	0%	20%	20%
그림작가	40%	20%	60%	40%	최신회차등록일	0%	0%	0%	0%
원작가	40%	40%	40%	40%	발행일	80%	80%	60%	60%
낭독자	0%	80%	0%	0%	파일개수	20%	60%	100%	80%
정가	60%	80%	60%	60%	완결여부	20%	20%	80%	80%
대여가	20%	20%	40%	40%	성인여부	20%	20%	60%	40%
출판사	100%	100%	100%	100%	유통사	40%	40%	60%	60%
ISBN/ECN	100%	100%	100%	100%	연재시작일	0%	0%	60%	60%
판매상태	60%	40%	60%	40%	연재종료일	0%	0%	60%	60%
총재생시간	0%	80%	0%	0%	연재주기	0%	0%	40%	20%
60% 이상 응답률을 보인 속성 개수						11개	13개	18개	15개

〈표 6〉 FGI 인터뷰 결과

속성명	의견	속성명	의견
자료유형	• 필요	주제	• 전자책, 오디오북 / 웹툰, 웹소설 별도로 구분 필요 • 유통사별로 달라서 체계적으로 정리하기 어려움 - 주제명표목표나 출판연감에서 제시하는 주제 분야 정도 분류 - 현재 분류된 ISBN 등록 속성을 중심으로 주제 분류
제목	• 필요		
부제	• 전자출판물의 경우 권차별로 권차명(부제목), 시리즈명 등이 별도로 있음 - 수험서 등도 권차표제가 별도로 있음 • 부제목, 권차명, 권차표제 별도로 제시할 필요가 있음		
저자	• 필요	파일형식	• 필요
그림작가	• 그림작가 필요(ISBN 발급 속성에 기본값으로 포함되어 있음)	유통사등록일	• 불필요
원작가	• 불필요	최초회차등록일	• 불필요
낭독자	• 오디오북 필요	최신회차등록일	• 불필요
정가	• 필요	발행일	• 발행, 제작 후 30일 이내에 납본율을 확인하기 위해서 필요 - 출간일 이내 납본 비율
대여가	• 정가 없이 대여가만 있는 자료에 대한 고려 필요	파일개수	• 총 자료수는 필요함(현재 오디오북은 종으로만 카운트됨)
출판사	• 필요	완결여부	• 필요
ISBN/ECN	• 필요	성인여부	• 성인 자료의 대량 납본이 많은데 이용 제한을 위해 필요함
판매상태	• 현 상황에서 판매(유통)되고 있는 자료에 대한 요청 - ISBN은 받았지만 판매되고 있지 않은 자료 등 정확한 데이터 확인을 위해 필요	유통사	• 불필요 • 제작처에 대한 정보 필요(오디오북) - 출판사는 종이책, 제작사는 오디오북 제작 - 출판사에 제작정보가 없으면 제작처로 구분 필요
총재생시간	• 오디오북 필요(납본 속성에 기본값으로 포함되어 있음)	연재시작일	• 연재종료일은 완결여부와 연결됨 - 연재기간이 중요함: 연재 종료 시점 예측 가능 - 완결 자료에 대한 납본 요청에 필요
		연재종료일	
		연재주기	• 불필요

경우 시리즈명, 권차명 등이 부제목으로 많이 사용되므로 필요하다고 하였다.

"전자책은 권차별로 제목이 다릅니다. 따라서 권차명, 시리즈명과 같이 권차명이 따로 있으며, 수험서 등은 권차표제도 존재합니다." (B 사서)

저자 관련 속성은 '저자', '그림작가' 등은 이견없이 필요하다고 하였으며, '원작가'는 납본 및 수집 차원에서는 필요하지 않다고 하였다. '낭독자'는 오디오북에만 해당되는 속성으로 정의하였고, 가격 관련 속성은 '정가'는 이견이 없었으나, '대여가'는 정가가 없는 자료에 대해서는 고려해야 할 필요가 있다고 하였다. '주제'는 필요한 속성이나 유통사별로 구분이 매우 달라 체계적인 정리가 어렵다고 하였다.

"분야 및 분류는 유통사별로 달라서 체계적으로 정리하기가 어렵습니다. 따라서 주제명표목표나 출판연감에서 제시하는 주제분야 정도로 분류해야 할 것입니다." (A 사서)

"장르 구분은 전자책, 오디오북과 웹툰, 웹소설 사이에 별도의 구분이 필요하며, 현재 분류된 ISBN을 가지로 목록을 중심으로 주제분류를 실시해야 한다고 생각합니다." (C 사서)

납본과 관련한 속성을 보면, '발행일'은 발행이나 제작 후 얼마나 빨리 납본이 되는지를 확인하기 위해 필요하다고 응답하였고, '연재종료일'은 완결 자료에 대한 납본 요청의 측면에서 '완결여부'와 연결되어 필요한 속성으로 규정하였다. '판매상태'는 현시점에서 판매 혹은 유통되는 자료에 대한 납본 요청을 위해서는 필요한 속성이라고 하였다. '성인여부'는 성인자료가 대량 납본되는 사례가 있어 해당 자료의 납본 제한을 위해 필요한 속성으로 규정하였다. 반면, '유통사'는 납본을 위한 통계자료 작성을 위해 특별하게 필요치 않은 속성이라고 하였으나, '오디오북' 측면에서는 제작처와 관련한 속성이 필요하다고 하였다.

"납본에 있어서는 어느 유통사이냐가 중요하지 않다. 일반적으로 출판사는 종이책, 제작사는 오디오북을 제작하고 있어 오디오북의 측면에서는 제작처는 필요하다고 생각합니다." (B 사서)

'ISBN/ECN'에 대해서는 ISBN이 발급되었으나 판매되지 않는 자료를 확인하는 것이 핵심이며, 유통되지 않는 자료를 제외하고 납본 통계를 작성하는 것이 필요하다고 하였다. 또한 자체 수집 자료에 대한 통계치 포함에 대해 언급하였다.

4.2.3 종합

본 절에서는 설문조사 및 FGI를 통해 국립중앙도서관 납본 및 수집 업무담당자가 필요로 하는 전자출판물의 속성 및 통계분석 시 필요한 요구사항을 파악하였다. 업무담당자가 설문조사와 FGI 모두에서 필요하다고 의견을 제시한 전자출판물 속성은 '자료유형', '제목', '저자', '정가', '출판사', 'ISBN/ECN', '파일형식' 등 총 7개 속성이었다. 반면, '원작가', '유통사등록일', '최초회차등록일', '최신회차등록일', '연재주기' 등 5개 속성은 불필요하다고 응답하였다. 자료유형별로 고려해야 할 속성을 살펴보면 '낭독

자', '총재생시간'은 오디오북에만 해당되는 속성이며, '그림작가'와 '성인여부'는 웹툰의 경우 필수 속성으로 분류하였다. '완결여부'와 '연재종료일'은 웹소설과 웹툰에서 필수 속성으로 분류하였다.

유통사로부터 데이터 수집시 고려해야 할 사항으로는 첫째, '부제'와 관련해서는 전자출판물의 경우 권차별로 권차명, 시리즈명 등이 따로 있고, 수험서 등은 권차표제가 별도로 있어 이러한 속성을 작성할 수 있는 '부제' 속성이 고려되어야 한다는 점이다. 둘째, '대여가'는 설문조사 시 전반적으로 필요치 않다고 응답하였으나, FGI 결과 정가 없이 대여가만 있는 자료에 대한 고려가 필요하다는 의견이 있어 전자출판물 데이터 수집시 염두해 두어야 하며, 셋째, '판매상태'의 경우 설문조사에서는 전자책과 웹툰에서만 일부 필요한 속성이라고 응답하였으나, FGI에서는 ISBN을 발급받았지만 판매(유통)되고 있지 않은 자료 등에 대한 정확한 통계를 위해 고려되어야 할 속성이라고 하였다. 넷째, '주제'와 관련해서는 유통사별로 체계가 매우 달라 어려움은 있지만, 모든 유형에서 100% 필요한 속성이라고 응답해 '주제'로 통일할 수 있는 속성값에 대한 고려가 필요하다는 것을 알 수 있었다. 다섯째, '완결여부'는 설문조사 시 웹툰과 웹소설에서만 필요한 속성으로 응답하였으나, FGI 결과 모든 속성에서 필요한 속성으로 의견이 모아져 이에 대한 고려가 필요하다. 여섯째, '발행일' 속성은 설문조사와 FGI 모두에서 필요 속성으로 응답하였는데, 발행 후 30일 이내에 납본율을 파악하기 위해 필요한 요소로 고려되어야 한다. 일곱째, '파일개수' 속성은 설문조사 시 전자책을 제외한 모든 유형에 필요하다고 응답하였으며, FGI에서는 오디오북은 현재 종으로만 계산되고 있어 총자료 수를 표시할 수 있는 속성이 고려되어야 한다. 여덟째, '유통사' 속성은 설문조사 시 웹툰과, 웹소설에서 일부 필요하다는 의견이 있었으나 FGI 결과 별도로 필요치 않는 속성으로 분류하였다. 다만, 일반적으로 종이책은 출판사가, 오디오북은 제작사가 제작을 담당하고 있어 출판사에서 오디오북에 대한 정보가 없을 경우 제작처를 확인할 필요가 있어 이에 대한 고려가 필요하다고 하였다. 본 연구에서는 이러한 의견을 반영하여 전자출판물 데이터 수집시 필요한 속성 및 속성값을 제시할 필요가 있다.

5. 유통 전자출판물의 데이터 품질

5.1 데이터 품질기준 정의

4개 유통사로부터 입수된 전자출판물 데이터에 대한 품질 검증을 실시하였다. 데이터 품질 검증은 데이터 품질 요소인 완전성, 유일성, 유효성, 일관성 정확성 측면에서 진행하였다. 검증 대상 속성은 총 22개(연구진이 추가한 4개 속성 제외)이며 최근 3년간(2020-2022) 유통사 전자출판물 속성별 데이터값을 검증 대상으로 하였다. 본 연구에서는 한국데이터베이스진흥원에서 제시한 〈표 7〉의 데이터 품질기준 정의에 따라 데이터 품질을 검정하였다.

〈표 7〉 데이터 품질기준 정의

품질기준	정의
완전성	필수항목에 누락이 없어야 한다.
유일성	데이터 항목은 유일해야 하며 중복되어서는 안 된다.
유효성	데이터 항목은 정해진 데이터 유효범위 및 도메인을 충족해야 한다.
일관성	데이터가 지켜야 할 구조, 값, 표현되는 형태가 일관되게 정의되고, 서로 일치해야 한다.
정확성	실세계에 존재하는 객체의 표현 값이 정확히 반영이 되어야 한다는 것을 의미한다.

출처: 한국데이터베이스진흥원 (2009). p. 17.

5.2 데이터 품질검증 결과

5.2.1 완전성

완전성 측면의 데이터 품질을 검정하기 위해 4개 유통사 데이터셋의 누락 데이터를 확인하였다. 누락 데이터의 유형은 속성 자체가 존재하지 않는 경우와 속성은 존재하나 일부 혹은 전체 속성값이 누락된 경우로 구분할 수 있다. 전자의 경우, 속성 자체가 없는 문제를 해결하기 위해서는 전자출판물 데이터와 관련하여 관리해야 할 전체 속성들을 확인하고 이를 포함하는 메타데이터 표준(안)을 마련하는 작업이 필요하다. 후자의 경우, 특히 제목, 저자, ISBN과 같이 기본적인 데이터가 누락되는 사례가 다수 발견되었는데, 예를 들어, A 유통사는 연재자료의 하위자료인 것으로 추정되는 자료(제목에 *부 *화, 외전, 프롤로그, 완결, 엔딩, 오프닝 등이 포함)의 상당 부분에 ISBN을 기재하지 않은 것으로 파악되었다. 따라서 제목, 저자, ISBN 등 유통사가 필수적으로 기입해야 하는 필수 속성을 지정할 필요가 있다.

한편, 자료유형별로는 전자책에 비해 오디오북 및 웹툰에 대한 데이터가 부실하거나 누락되는 경우가 많았는데 A 유통사의 경우 모든 오디오북의 ISBN이 없고, 저자 및 낭독자도 일부만 포함되어 있었다. 따라서 유통사들이 전자책뿐만 아니라 모든 유형의 전자출판물에 대해 데이터를 누락하지 않고 성실하게 입력하도록 안내해야 한다.

5.2.2 유일성

유통사별 데이터셋을 통합하여 데이터 분석을 하고자 할 때 중요한 작업 중 하나는 중복자료들을 제거하고 유일한 전자출판물들을 식별하는 일이다. 본 연구에서 D 유통사는 모든 연재자료마다 연재 ISBN을 기재하고 있었으나 나머지 유통사의 데이터셋에는 ISBN이 없는 자료가 많은 것으로 나타났다. 또한 ISBN은 동일하나 서지사항이 다르거나 반대로 동일제목을 가진 2개 이상의 자료들이 서로 다른 ISBN을 갖는 등 다양한 사례들이 존재하여 ISBN이 식별자로써 온전한 역할을 하지 못 하는 경우가 다수 발견되었다. 이로 인해 ISBN만을 기준으로 유통사들의 데이터셋을 통합하여 중복자료를 삭제하고 유일한 자료를 식별하는 작업이 매우 어려울 것으로 예측된다.

이러한 유일성의 문제는 여러 요인에 의해 발생하는 것으로 파악되었다. 첫째, 연재자료의 대표표제에만 ISBN이 발급되기도 하고 각 하위자료에 개별적으로 ISBN이 발급되기도 하

는 등 연재자료에 대한 비일관적인 ISBN 발급 기준으로 인해 혼란이 발생하였다. 현재 국립중앙도서관에서는 연재형 웹콘텐츠(웹툰・웹소설)에 대해 ISBN을 2024년 12월 31일까지만 부여할 것을 공지하고 있다(국립중앙도서관 ISBN・ISSN 납본 시스템, 2024). 이후에는 연재 종료 후 전자책 출판 시 ISBN이 부여되어 연재자료에 대한 일관적인 ISBN 발급 기준이 정착될 것으로 보인다. 다만, 기존의 자료들은 ISBN이 발급된 단위가 연재자료 전체이든 개별 회차이든 간에 상관없이 ISBN을 기준으로 식별해야 한다. 둘째, ISBN이 없는 경우, 단행본의 권차, 연재자료의 회차 정보 등이 부실하여 자료를 구별하기 어려운 경우가 있으므로 권차/회차 정보를 독립적인 메타데이터 속성으로 정의하여 해당 속성값을 반드시 입력하도록 해야 한다.

5.2.3 유효성

데이터의 유효성을 확보하기 위해서는 정해진 데이터의 유효범위를 충족해야 한다. 본 연구의 4개 데이터셋에서는 날짜와 ISBN 속성이 유효한 값을 갖지 않는 경우들이 소수 발견되었다(예, 발행일 '22020513'). 이러한 문제를 해결하기 위해 데이터의 유효범위에 포함되지 않는 경우는 기계처리로 확인하여 수정 작업의 효율성을 높일 수 있다.

5.2.4 일관성

4개 유통사 간, 그리고 동일 유통사의 데이터셋 내에서도 속성값을 선정하고 표현하는 규칙이 비일관적인 사례가 다수 발견되었다. 첫째, 모든 데이터 속성에 대해 내용규칙이 필요하지만 그 중에서도 특히 속성값의 범위가 한정되어 있지 않은 '제목', '부제', '저자' 속성의 경우 내용을 기술하는 규칙이 매우 비일관적인 것으로 나타났다. 제목/부제는 서지사항의 가장 기본적인 요소로써 정확하게 기술되어야 함에도 불구하고 아래와 같이 제목/부제와 상관없는 정보가 포함되어 있는 경우가 많았다. 예를 들어, 웹소설이나 웹툰의 경우 장르(예, [BL], [GL], [리얼로맨스]), 연재・세트 여부(예, [연재], [세트]), 완결 여부(예, 미완결) 등의 정보가 제목/부제에 포함되어 있기도 하고, 오디오북의 경우 연예인 낭독자의 이름을 제목 앞에 임의대로 추가한 사례도 발견되었다(예, 이보영의 노인과 바다). 이러한 사례들은 이용자의 도서 검색 및 선정에 도움을 주기 위해 각 유통사에서 제목/부제에 부가적인 정보를 추가한 것으로 보이며, 결과적으로 데이터의 일관성을 해치고 원활한 데이터 통합 및 분석을 저해하는 요인이 되므로 반드시 표준화된 내용 규칙을 마련할 필요가 있다.

〈예〉
[BL] R의 법칙(외전)
[연재] 좀비 뽑기로 초월까지 226화
[고화질] 건달군과 안경양 01
어떤 아이가(오디오북)
마지막 잎새(배우 정해인 낭독)
이보영의 노인과 바다

이와 마찬가지로 저자의 속성값에 글작가만 포함하기도 하고 다른 역할의 저자들을 모두 포함하기도 하는 등 저자에 대한 속성값 선정 규칙이 하나의 유통사 데이터셋 내에서도 통일성이 없는 것으로 나타났다. 표기 방식도 저자명만 나열하기도 하고(예, 김재춘, 배지현), 저자 역할어와 함께 기술하기도 하고(예, 유리아

히로커 저/김은수 역), 오디오북의 경우 연기한 배역을 함께 기술하기도 하는 등(예, 박종희, 유경선(아내 역)) 다양한 방식으로 기술되고 있어서 저자 속성값에 대한 선정 및 표현 규칙에 대한 표준화가 필요한 것으로 나타났다.

둘째, 유통사별로 속성값의 범위가 다른 속성으로 '파일형식', '판매상태', '주제' 등이 있다. 파일형식의 경우, 〈표 8〉과 같이 C 유통사는 MP3 자료를 'AUDIO'라는 속성값으로 표현하고, 세트 자료의 파일형식을 입력하는 대신 'SET'로 표시하고 있어서 유통사 간의 통일성이 있는 속성값을 선정하여 사용할 필요가 있는 것으로 나타났다.

'주제' 속성의 경우 유통사별로 분류체계가 매우 상이하여 표준화 작업은 쉽지 않을 것으로 예상된다. A와 C 유통사는 둘 다 계층적 분류체계를 이용하고 있으나 C 유통사가 훨씬 세분화된 분류체계를 가지고 있었다(〈표 9〉 참조).

〈표 8〉 유통사별 파일형식 및 판매상태의 속성값

속성	A 유통사	B 유통사	C 유통사	D 유통사
파일형식	EPUB PDF MP3	EPUB EPUB_KPC EPUB_COMIC PDF PDF_EDU COMIC	EPUB EPUB_KPC SET PDF AUDIO COMIC	-
판매상태	정상, 예약, 판매금지	Y, N	Y, N	게시 중

〈표 9〉 유통사별 주제 속성값: 최상위 카테고리

A 유통사	B 유통사	C 유통사		D 유통사
가정/육아	-	BL	에세이	감성
건강		가격대별 eBook	여행	개그
경제/경영		건강/취미	역사	드라마
과학		경제경영	예술/대중문화	로맨스
소설		고등학교 참고서	외국어	무협/사극
시/에세이		고전	요리/살림	스릴러
역사/문화		과학	유아	스포츠
예술/대중문화		대학교재/전문서적	인문학	액션
외국어		라이트노벨	자기계발	일상
유아(0~7세)		로맨스	종교/역학	판타지
인문		만화	좋은부모	
자기계발		사전/기타	중학교 참고서	
잡지		사회과학	청소년	
정치/사회		소설/시/희곡	초등참고서	
중/고등참고서		수험서/자격증	컴퓨터/모바일	
취미실용/스포츠		오디오북	판타지/무협	
취업/수험서		어린이	19+	
컴퓨터/IT				

또한 A, C 유통사의 분류체계는 주제뿐만 아니라 자료유형(예, 잡지), 독자층(예, 청소년, 유아), 기타(예, 가격대별 ebook) 등 비주제적인 카테고리를 다수 포함하고 있었다. 반면에 D 유통사는 계층적 분류체계 대신에 '개그', '로맨스', '판타지' 등 웹툰 장르명 중 하나를 선택하여 기입하고 있었다. 요약하면, 각 유통사는 주제 및 비주제를 포함한 서로 다른 카테고리를 제공하고 있고 세분화 수준에서 차이가 커서 향후 전자출판물의 주제 속성값을 표준화하여 사용하고자 한다면 통일된 분류체계를 개발하는 작업이 선행되어야 한다.

셋째, 출판사의 경우, 현재 동일 출판사에 대한 여러 이명 아래에 데이터가 흩어져 있어서 통합이 어려운 것으로 나타났는데 이는 아래와 같이 한글명과 영문명을 동시에 사용하거나 ㈜를 사용하는 이명이 많기 때문이다. 따라서 향후 국립중앙도서관에서는 출판사명에 대한 전거파일을 구축하여 유통사들과 공유할 필요가 있다.

〈예〉
문피아, MUNPiA(문피아), 문피아(MUNPia)
BOOKK, BOOKK(부크크), ㈜부크크
㈜조은세상, 조은세상

넷째, 속성값의 표기 형식이 상이한 속성으로는 ISBN, 유통사등록일, 최초회차등록일, 최신회차등록일, 총재생시간 등이 있다(〈표 10〉 참고). 특히 총재생시간의 경우 유통사별로 단위가 명시되어 있지 않아 통합이 어려운 것으로 나타났다. 이와 같이 단순히 표기 형식이 상이하거나 단위가 명시되지 않은 경우에는 공통적인 표기 형식 및 단위를 선정하여 사용함으로써 비교적 손쉽게 표준화하는 것이 가능할 것으로 보인다. 특히 날짜 및 시간의 경우 국제 규칙을 따르는 것이 바람직하다.

5.2.5 정확성

4개 데이터셋에서 데이터 오류들이 발견되어 전체 데이터의 신뢰성을 저하하는 요인으로 파악되었다. 예를 들어, 유통사등록일이 출간일 이후 시점이어야 하나 그렇지 않은 경우들이 있었고, '필드값 없음'이 입력되어야 할 곳에 '0'이 입력되어 있는 경우들도 있었다. 이러한 문제를 해결하기 위해서는 데이터 입력 단계에서 이상치를 기계처리로 확인하여 수정 작업의 효율성을 높일 수 있다.

〈표 10〉 속성값 표기 방식이 상이한 예

속성	표기 방식
ISBN	NNNN-NN-NNNN-NNN-N NNNNNNNNNNNNN
유통사등록일/ 최초회차등록일/ 최신회차등록일/ 발행일	YYYY-MM-DD YYYYMMDD DDMMYY YYYY---- 등
총재생시간	1~21405 사이의 정수 0:00:00

6. 종합 및 결론

이상과 같이 국내 유통사 전자출판물 데이터 수집을 위해 국립중앙도서관 납본 및 수집 업무담당자가 필요로 하는 속성에 대한 요구사항을 파악하고, 실제 유통사 전자출판물 데이터의 속성값의 품질을 검증하였다. 이를 토대로 본 연구는 국내 유통사 전자출판물 데이터 수집시 고려되어야 할 속성과 속성값 표준화 방안과 정책 및 제도적 고려사항을 제안하고자 한다. 먼저 국내 유통사 전자출판물 데이터를 수집할 때 고려되어야 할 속성들과 속성값 표기 방식의 표준화 요소들을 제시하면 다음과 같다(〈표 11〉 참조). 첫째, 국내 전자출판물 데이터 수집시 고려되어야 할 속성은 필수 및 선택 속성을 합하여 총 21개 속성이다. 최종 속성 도출 및 필수/선택/불필요 속성 구분은 국립중앙도서관 납본 및 수집 업무담당자를 대상으로 실시한 설문조사 결과와 FGI 결과를 토대로 1차 선별하였으며, 유통사로부터 수집된 데이터 품질 검증 과정에서 요구되는 속성 및 속성값을 고려하여 추가로 포함하였다. 자료유형별 속성을 보면, 전자책의 경우 필수 속성 8개, 선택 속성 10개로 총 18개 속성이며, 오디오북은

〈표 11〉 국내 유통사 전자출판물 데이터 수집시 속성 및 속성값 고려사항

속성	자료유형별 필수/선택 요소				속성값 표준화
	전자책	오디오북	웹툰	웹소설	
자료유형	○	○	○	○	전자책, 오디오북, 웹소설, 웹툰
제목	○	○	○	○	표제, 관제, 표제관련정보
부제	▲	▲	▲	▲	부제목, 권차표제
저자	○	○	○	○	저자명 + 역할어
그림작가	▲	▲	○	▲	그림작가명
낭독자	×	○	×	×	낭독자명
권차/회차	▲	▲	▲	▲	권/호
정가	▲	▲	▲	▲	원
대여가	▲	▲	▲	▲	원
발행처	○	○	○	○	출판사, 제작사
ISBN	○	○	○	○	NNNNNNNNNNNNN
ECN	▲	▲	▲	▲	ECN-NNNN-NNNN-NNN-NNNNNNNNN
판매상태	○	○	○	○	Y, N
총재생시간	×	○	×	×	[hh]:[mm]:[ss]
주제	▲	▲	▲	▲	주제 구분 표준화 필요
파일형식	○	○	○	○	PDF, ePUB, MP3, MP4 등
발행일	○	○	○	○	YYYYMMDD
파일개수	×	▲	▲	▲	개
완결여부	▲	▲	○	○	Y, N
성인여부	▲	▲	○	○	전체연령가, 12세 연령가, 15세 연령가, 18세 연령가
연재종료일	▲	▲	○	○	YYYYMMDD

○ 필수, ▲ 선택, × 불필요

필수 속성 10개, 선택 속성 11개로 총 21개, 웹툰은 필수 속성 12개, 선택 속성 8개로 총 20개, 웹소설은 필수속성 11개, 선택 속성 8개로 총 19개 속성이다. 최종 속성 및 속성값 도출에 있어 '권차/회차' 속성은 업무담당자 요구조사에서는 드러나지 않았지만 4개 유통사의 데이터셋 품질 검증 과정에서 자료의 식별을 위해 필요한 것으로 확인되어 추가하였다. 'ISBN/ECN' 속성은 속성값의 표준화와 관련하여 별도의 항목으로 구분하는 것이 효과적으로 판단되어 별도 속성으로 분리하였고, ECN의 경우 향후 온라인 콘텐츠 식별체계의 환경변화에 따라 UCI로의 변경도 고려해야 할 것이다. '발행처'의 경우 기존 유통사에서는 출판사를 주로 사용하였으나, 의견수렴 시 오디오북 등 제작사와 관련된 의견도 제시되어 출판사와 제작사를 포괄할 수 있는 '발행처'로 속성명을 확정하였다. 둘째, 전자출판물 속성별 속성값의 표준화는 다음과 같다. 먼저, 날짜 및 시간 관련 속성은 국제 규칙인 ISO 8601에 따라 발행일, 연재종료일, 총재생시간을 표기하였다. 그리고 속성값 내용의 표준화를 위해 '파일형식', '성인여부' 등 속성값의 범위가 제한되는 속성의 경우 선정할 수 있는 속성값을 나열하였다. 속성값을 자유롭게 기술하는 '제목'의 경우 [합본], [BL] 등 제목과 관련 없는 정보는 포함하지 않고 표제, 관제, 표제관련정보 등을 자세히 기술하여 자료를 식별할 수 있도록 하였다. '저자'는 저자명과 역할어를 기입하도록 하고(예, 유리아 히로커 저/김은수 역), 대신 '그림작가'와 '낭독자'는 이미 역할이 분명하므로 저자명만 기입하도록 하였다. '발행처'의 속성값은 이명의 발행처명을 통일된 기관명으로 연결할 수 있도록 전거제어 작업이 먼저 수행되어야 한다. 이와 더불어, '주제'와 관련하여서는 현재 유통사별로 주제구분이 매우 상이하여 표준화된 속성값이 필요하다. 표준화를 위해서는 주제명표목표, 출판연감 주제구분, ISBN 등록시 주제구분 등을 고려할 필요성이 있다.

다음으로 국내 유통사 전자출판물 데이터 수집을 위한 정책 및 제도적 고려사항을 제시하면 다음과 같다. 첫째, 국내 유통사 전자출판물의 데이터 수집과 관련하여, 가장 우선적으로 필요한 것은 표준화된 메타데이터 요구사항의 확립이다. 현재 전자출판물의 메타데이터는 출판사나 유통사마다 형식이 다르며, 이로 인해 국립중앙도서관이 데이터 수집과 관리에 어려움을 겪을 수 있다. 메타데이터의 표준화는 전자출판물의 제목, 저자, 발행일, 발행처, ISBN, 정가 등 기본적인 정보뿐만 아니라, 파일형식, 주제, 파일개수, 판매상태 등 보다 세밀한 정보를 포함해야 한다. 이를 위해 국가적 차원에서 출판 및 유통업계와 협력하여 국제적으로 통용되는 표준을 참고하되, 국내 실정에 맞는 메타데이터 기준을 마련하고 이를 지원할 필요가 있다. 이 과정에서 출판계의 의견을 적극 수렴하여 실질적으로 활용 가능하고 데이터 품질을 보장할 수 있는 메타데이터 표준을 확립할 필요가 있다. 둘째, 전자출판물 데이터의 수집뿐만 아니라, 이들 데이터의 품질을 지속적으로 관리하고 모니터링하는 체계의 구축도 필요하다. 이는 데이터의 완전성, 정확성, 일관성 등을 보장하기 위함이다. 국립중앙도서관은 전자출판물의 메타데이터 및 내용에 대한 정기적인 검증 절차를 마련하고, 이를 통해 수집된 전자출판물의 품질을 체계적으로 관리해야 한다.

또한, 이러한 모니터링 결과를 바탕으로 필요한 경우 메타데이터 기준의 개정이나 보완을 실시할 수 있어야 한다. 이 과정에서 인공지능 기술을 활용하여 대규모 데이터의 품질 관리와 모니터링을 자동화하는 방안도 고려할 수 있다. 이를 통해 국립중앙도서관은 전자출판물 데이터의 품질과 신뢰성을 지속적으로 유지할 수 있을 것이다.

본 연구는 다음의 한계를 가진다. 첫째, 연구 대상을 4개의 전자출판물 유통사로 한정함으로써 국내 전체 전자출판물 시장을 대표하기에는 부족한 측면이 있다. 다양한 규모와 유형의 유통사를 포괄하지 못했기 때문에, 연구 결과를 일반화하는데 한계가 있다. 둘째, 실시간으로 변화하는 전자출판물 데이터의 특성상, 수집된 데이터는 연구 시점에 따라 차이가 있을 수 있다. 따라서 연구 결과의 시의성을 유지하기 위해서는 지속적인 업데이트가 필요하다. 셋째, 유통사별로 데이터 제공 정책의 차이로 인한 수집된 데이터 간에 편차가 존재하였다. 이로 인해 특정 유통사로부터 충분한 데이터를 확보하는데 어려움이 있었다. 넷째, 본 연구에서는 국립중앙도서관의 전자출판물 납본 및 수집 업무담당자로 한정하여 설문조사 및 FGI를 진행하였기 때문에 전자출판물 관련 모든 업무에 연구결과를 일반화하는데 한계점이 있을 수 있다. 그리고 실제 유통사 전자출판물 담당자들을 대상으로 한 설문조사 및 FGI가 함께 진행되지 못하여, 본 연구에서 제안된 국내 유통사 전자출판물 데이터 수집시 필요한 속성 및 속성값(〈표 11〉)의 객관성과 범용성에 한계점이 있을 수 있다. 이상의 연구의 한계점을 극복하기 위해서 향후 연구에서는 다양한 전자출판물 유통사를 포함시켜 연구의 대표성을 강화할 필요가 있다. 이를 통해 국내 전자출판물 유통 현황을 보다 정확하게 파악할 수 있을 것이다. 전자출판물 데이터의 실시간 변화를 반영하기 위해 정기적인 모니터링과 분석이 중요하다. 이를 위해 자동화된 데이터 수집 및 분석 시스템의 개발을 고려할 필요가 있다. 이를 통해 변화하는 전자출판물 유통 동향에 신속히 대응할 수 있는 기반이 마련될 수 있다. 유통사 전자출판물 데이터에 대한 접근성을 개선하기 위해서는 유통사와의 협력을 강화하는 방안을 모색해야 한다. 이를 위해서 데이터 공유에 대한 인센티브 제공, 공동 연구 프로젝트 수행 등을 고려할 필요가 있다. 국립중앙도서관 납본 및 수집 업무담당자뿐만 아니라 전자출판물 관련 다양한 업무 담당자들을 포함하여 연구가 진행될 필요가 있다. 또한 실제 유통사의 전자출판물 업무담당자를 대상으로 한 연구도 함께 진행하여, 실제 국내 전자출판물 유통 시장의 요구사항도 반영할 필요가 있다. 본 연구를 토대로 향후 국내 전자출판물 데이터의 수집 및 분석 작업이 보다 체계적이고 효율적으로 이루어질 수 있기를 기대해 본다.

참 고 문 헌

공병훈, 조정미 (2021). 한국 출판 기술 발전의 진화 단계 연구. 한국출판학연구, 47(6), 5-32. http://doi.org/10.21732/skps.2021.103.5

곽승진, 김정택, 박옥남, 최재황 (2013). 온라인 디지털자료의 납본 가이드라인에 관한 연구. 한국문헌정보학회지, 47(3), 161-179. https://doi.org/10.4275/KSLIS.2013.47.3.161

곽승진, 최재황, 조영주, 류희경 (2008). 디지털자료 납본에 대한 보상 체계 연구. 한국도서관·정보학회지, 39(2), 65-83. http://doi.org/10.16981/kliss.39.2.200806.65

국립중앙도서관 (2023). 국내 전자출판물 통계조사 기초연구.

국립중앙도서관 ISBN·ISSN 납본 시스템 (2024.1.29.) 연재형 웹콘텐츠(웹툰·웹소설) ISBN 부여 종료 안내.
출처: https://www.nl.go.kr/seoji/contents/S20200000000.do?page=1&schM=view&id=10943&schFld=bd_title

김규환, 정대근, 김수정 (2023). 국내 전자출판물의 납본·수집 현황 분석. 한국도서관·정보관리학회지, 54(4), 281-306. http://doi.org/10.16981/kliss.54.4.202312.281

김정명, 박찬수 (2022). 한·일 출판물 메타데이터 활용 운영의 비교 연구. 한국출판학연구, 48(6), 5-29. http://doi.org/10.21732/skps.2022.109.5

남영준, 장보성 (2006). 전자자원 관리용 메타데이터의 요소 분석에 관한 연구. 정보관리학회지, 23(3), 241-264. http://doi.org/10.3743/KOSIM.2006.23.3.241

서혜란 (2003). 디지털자료의 납본과 보존을 위한 각 국가의 노력. 정보관리학회지, 20(1), 373-399. https://doi.org/10.3743/KOSIM.2003.20.1.373

손애경 (2012). 국내전자출판 기술지원정책 현황 연구: 문화기술(CT) 전자출판 지원정책을 중심으로. 전자출판연구, 1, 85-96.

윤희윤 (2003). 한국의 납본제도 개선모형에 관한 연구. 한국문헌정보학회지, 37(4), 24-52. https://doi.org/10.4275/KSLIS.2003.37.4.024

이숙현 (2003). 하이브리드도서관을 지향하는 국립중앙도서관의 전자출판물 수집 정책. 한국비블리아학회지, 14(1), 63-78.

장보성, 남영준 (2010). 온라인 전자책 보존을 위한 납본제도 개선 방안 연구. 한국문헌정보학회지, 44(4), 435-456. http://doi.org/10.4275/KSLIS.2010.44.4.435

최재황, 곽승진, 김정택 (2009). 온라인 디지털자료의 납본체계 및 이용에 관한 연구. 한국도서관·정보학회지, 40(1), 209-232. http://dx.doi.org/10.16981/kliss.40.1.200903.209

출판문화산업 진흥법. 법률 제19599호.

하진희, 임순범, 김성혁 (2003). 전자책 라이브러리를 위한 메타데이터 개발에 관한 연구. 정보관리학회지, 20(3), 1-16. http://doi.org/10.3743/KOSIM.2003.20.3.001

한국데이터베이스진흥원 (2009). 데이터 품질진단 절차 및 기법(Ver 1.0).

한국출판문화산업진흥원 (2021). 전자출판 산업분석 및 활성화를 위한 조사 연구.

한혜영 (2003). 전자출판물의 납본시스템에 관한 연구. 정보관리학회지, 20(3), 51-79. http://doi.org/10.3743/KOSIM.2003.20.3.051

Gooding, P. & Terras, M. (Eds.). (2020). Electronic Legal Deposit: Shaping the Library Collections of the Future. UK: Facet Publishing.

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

Choi, Jae-Hwang, Kwak, Seung-Jin, & Kim, Jeong-Taek (2009). A study on legal deposit process and use of online digital materials. Journal of Korean Library and Information Science Society, 40(1), 209-232. http://dx.doi.org/10.16981/kliss.40.1.200903.209

Ha, Jin-Hyee, Lim, Soon-Bum, & Kim, Sung-Hyuk (2003). A study on the development of metadata for eBook library. Journal of the Korean Society for Information Management, 20(3), 1-16. http://doi.org/10.3743/KOSIM.2003.20.3.001

Han, Hyeyoung (2003). A study on the deposit system for electronic publications. Journal of the Korean Society for Information Management, 20(3), 51-79. http://doi.org/10.3743/KOSIM.2003.20.3.051

Jang, Bo-Seong & Nam, Young-Joon (2010). Research on improvement of the legal deposit system for the preservation of online electronic book. Journal of the Korean Society for Library and Information Science, 44(4), 435-456. http://doi.org/10.4275/KSLIS.2010.44.4.435

Kim, Gyuhwan, Jeong, Daekeun, & Kim, Soojung (2023). Analysis of the status of legal deposit and acquisition of electronic publications in Korea. Journal of Korean Library and Information Science Society, 54(4), 281-306. http://doi.org/10.16981/kliss.54.4.202312.281

Kim, Jungmyoung & Park, Chan Su (2022). A comparative study on the utilization of publications metadata in Korean and Japan. Studies of Korean Publishing Science, 48(6), 5-29. http://doi.org/10.21732/skps.2022.109.5

Kong, Byounghun & Cho, Jungmi (2021). A study on the evolutionary stages of the development of publishing technology in Korea. Studies of Korean Publishing Science, 47(6), 5-32.

http://doi.org/10.21732/skps.2021.103.5

Korea Data Agency (2009). Procedures and Techniques for Data Quality Diagnosis(Ver 1.0).

Kwak, Seung-Jin, Choi, Jae-Hwang, Cho, Young-Joo, & Ryu, Hee-Kyeung (2008). A study on reimbursement for legal deposit of digital products. Journal of Korean Library and Information Science Society, 39(2), 65-83. http://doi.org/10.16981/kliss.39.2.200806.65

Kwak, Seung-Jin, Kim, Jeong-Taek, Park, Ok-Nam, & Choi, Jae-Hwang (2013). A study on legal deposit guidelines for online digital materials. Journal of the Korean Society for Library and Information Science, 47(3), 161-179.
https://doi.org/10.4275/KSLIS.2013.47.3.161

Lee, Sook Hyeun (2003). Acquisition policy on electronic publications by the National Library of Korea toward hybrid library. Journal of the Korean Biblia Society for Library and Information Science, 14(1), 63-78.

Nam, Young-Joon & Jang, Bo-Seong (2006). The study of the elements analysis of metadata for electronic resource management. Journal of the Korean Society for Information Management, 23(3), 241-264. http://doi.org/10.3743/KOSIM.2006.23.3.241

National Library ISBN · ISSN Deposit System (2024.1.29.) Notice on the end of ISBN assignment for serial Web content (Webtoon, Web novel) Available:
https://www.nl.go.kr/seoji/contents/S20200000000.do?page=1&schM=view&id=10943&schFld=bd_title

National Library of Korea (2023). Basic Research on Domestic Electronic Publication Statistics Survey.

Publication Culture Industry Promotion Act.19599.

Publication Industry Promotion Agency of Korea (2021). Study for Analysis and Activation of the Electronic Publishing Industry.

Son, Ae Kyoung (2012). A study of the current state of the domestic digital publishing technology support policies: Focusing on the culture technology digital publishing support policies. Digital Publishing Research, 1, 85-96.

Suh, Hey-Ran (2003). Legal deposit and preservation of digital materials in various countries. Journal of the Korean Society for Information Management, 20(1), 373-399.
https://doi.org/10.3743/KOSIM.2003.20.1.373

Yoon, Hee-Yoon (2003). A study on the reform model of legal deposit system in Korea. Journal of the Korean Society for Library and Information Science, 37(4), 24-52.
https://doi.org/10.4275/KSLIS.2003.37.4.024