OPEN ACCESS CC BY NC ND

Review https://doi.org/10.23287/KJDH.2025.2.1.1

AI 시대, 국가지식문화자원 데이터 허브 국립중앙도서관 *

National Knowledge and Cultural Data Hub: The Role of the National Library of Korea in the Age of AI

김수정(Kim, Sue-jeong)**
0009-0000-9558-9556

목차

1. 서론
2. AI 시대 지식정보 패러다임의 전환
3. 도서관의 대응
4. 해외 사례
4.1 미의회도서관 (Library of Congress)
4.2 영국국립도서관 (British Library)
4.3 네덜란드국립도서관 (Koninklijke Bibliotheek)
5. 국립중앙도서관 데이터 현황
5.1 데이터 유형별 현황
5.2 데이터 서비스 예시
6. 국립중앙도서관 데이터의 특성
7. 결론

초록

본 논문은 인공지능(AI) 시대를 맞아 국가지식문화자원 데이터 허브로서 국립중앙도서관의 역할을 조명하였다. AI 시대 지식정보 패러다임의 전환을 개관하고, 데이터 허브로서 주요 해외 국가도서관의 사례를 분석하였다. 이어서 국립중앙도서관의 데이터 구축 현황과 서비스 유형, 데이터의 특성(신뢰성, 포괄성, 다양성, 개방성, 확장성, AI 활용성 등)을 제시하고, 데이터 허브로서의 기능 강화를 위한 네 가지 전략 방향을 제안하였다. 본 논문은 국립중앙도서관의 데이터 기반 기능을 재조명함으로써, 국가 차원의 AI 학습 데이터 생태계 구축을 위한 기초 자료로 기여하기를 기대한다.

주제어: 국립중앙도서관, 국가지식문화자원, AI 학습데이터, 데이터 허브

Abstract

This study examines the role of the National Library of Korea (NLK) as a key institution for managing national knowledge and cultural resources and as a central data hub in the era of artificial intelligence (AI). It begins by reviewing the paradigm shift in knowledge and information systems driven by AI technologies and investigates the responses of leading national libraries worldwide that have adopted data hub functions. The paper then outlines the NLK's current status in terms of data

*본 논문은 2024 디지털인문학대회(KADH2024)에서 발표한 내용을 보완 정리한 것입니다.

**단독저자: 국립중앙도서관 사서, (문헌정보학 박사), goldcristal@korea.kr

construction, service provision, and the characteristics of its data resources, including reliability, comprehensiveness, diversity, openness, scalability, and applicability to AI training. Based on this analysis, the study proposes four strategic directions to enhance the NLK's capacity as a national data hub. This research aims to provide foundational insight for building a national-level AI training data ecosystem and to reposition the NLK as a core infrastructure for AI-driven knowledge innovation.
Keywords: National Library of Korea, National Knowledge and Cultural Resources, AI Training Data, Data Hub.

1. 서론

인공지능(AI) 시대는 지식의 축적 방식뿐만 아니라 생성과 활용 방식 전반에 걸쳐 구조적 전환을 이끌고 있다. 이에 따라 국가 차원의 데이터 전략 수립과 데이터 허브 기관의 역할이 점점 더 중요해지고 있다. 본 논문은 이러한 시대적 흐름 속에서, 국립중앙도서관이 국가 지식문화자원의 관리 주체이자 데이터 허브로서 수행해야 할 전략적 역할을 조명하고자 한다. 특히 '소유에서 접근, 접근에서 생성으로' 전환되는 지식정보 패러다임을 개관하고, 데이터 허브로서 주요 해외 국가도서관의 사례를 분석하였다. 아울러 도서관법에 따라 국가대표도서관으로서 국립중앙도서관이 수행해 온 데이터 허브 기능과 현황을 분석하였다. 마지막으로 데이터의 신뢰성, 다양성, 개방성, 확장성, AI 활용 가능성 등의 특성을 바탕으로, 국립중앙도서관이 데이터 허브로서의 기능 강화를 위한 네 가지 전략 방향을 제시하였다. 본 논문은 국립중앙도서관 데이터에 대한 이해를 심화하고, 국가적 차원의 AI 학습데이터 생태계 구축에 기여할 수 있는 정책 기반을 제공하고자 한다.

2. AI 시대 지식정보 패러다임의 전환

Sagan(2014)[1]은 약 100조 비트의 정보가 우리 뇌 안에 있고, 그 정보를 모두 영어로 기술한다면 대략 2,000만권의 책 더미가 쌓일 것이라고 했다. 인간의 생존에 필요한 정보 양이 인간의 두뇌에 보관할 수 없을 정도로 늘어나자, 육체 바깥에 필요한 정보를 저장해 두었으며, 이를 우리는 도서관이라 부른다 하였다. Moller(2023)[2]는 기원전 300년경 알렉산드리아도서관은 이 세상에 존재하는 모든 문서의 사본을 한 장소에 모아놓음으로써 인간의 모든 지식을 아우를 수 있다는 개념을 처음 만들어냈다라고 했다. 이후 수천 년이 흘러 1998년 설립된 구글은 "세상의 모든 정보를 조직하고 접근 가능하게 만드는 것"을 목표로 수십억 개 웹페이지를 인덱싱하였다. 폭발적인 데이터 증가는 정보를 소유에서 접근으로 무게 중심 축을 이동시켰고, 2018년 이후 본격 성장한 ChatGPT와 같은 LLM(대형 언어 모델)은 방대한 데이터를 학습하고 유용한 정보를 생성하는데 이르렀다. 지식의 생산과 확산 속도는 우리 사회 혁명적 변화를 가져왔고, 고서, 근현대자료 등 도서관 소장자료가 디지털화되면서 데이터 처리 분석, 교육 및 학습 혁신 등 새로운 르네상스를 예고하고 있다.

표 1. 알렉산드리아 도서관에서 LLM 까지

주요 기관/시스템		목적	특징	자료 규모
기원전 3세기경	알렉산드리아 도서관	인간 지식의 집대성	모든 문서의 사본을 한곳에 수집하여 보존	약 50,000권 ~ 700,000권[3]
1998년 설립	구글 (Google)	전 세계 정보의 조직화 및 접근 가능성 확대	웹페이지 인덱싱, 검색 기반 정보 접근, Google Books 통한 대규모 디지털화	수십억 개 웹페이지, Google Books[4]
2018년 이후	LLM (GPT, BERT 등)	방대한 데이터 학습을 통한 정보 생성	자연어 처리 기반 지식 생성, 문맥 이해 및 생성 능력 강화	GPT-3: 약 3,000억 토큰 학습[5]

3. 도서관의 대응

일찍이 도서관계는 소장 장서 데이터화를 통해 사람과 기계 모두에게 데이터 장벽을 낮추고, 정보 접근성을 높이려는 노력을 해왔고, 2017년 산타바바라성명[6] 작성을 주도하였다. 2017년 산타바바라 성명은 디지털 인문학 시대를 맞아 장서 데이터화의 중요성을 강조하며, 표준화된 메타데이터를 통해 데이터 상호 운용성을 확보하고, 연구데이터의 개방 접근을 통해 지식 공유 촉진을 주장했다. 2023년 벤쿠버 성명[7]은 인공지능 기술이 발전하고 장서 데이터 활용 방식이 더욱 복잡해짐에 따라, 인공지능과의 공존, 데이터 프라이버시와 윤리, 지속 가능한 데이터 인프라 구축 등 새로운 문제들을 제기하고 해결 방안을 모색했다. 두 성명을 통해 보면 도서관계는 장서 데이터화 등을 추진하며 정보 접근성을 높이려는 노력을 해왔고, 시대적 흐름과 기술 발전에 따라 그 내용이 변화하고 있음을 알 수 있다.

국립중앙도서관 또한 도서관법에 따른 국가대표도서관이자 납본도서관으로서 국가지식문화자원을 망라적으로 수집하고 디지털화 데이터화를 추진하며, 사람과 기계 모두에게 정보 접근성을 높이고자 노력해왔다. Vitiello(2001)[8]는 국가도서관 납본은 국가의 출판물을 보급하고 알리는 국가의 브레인으로 작용하며, 그 국가의 지적패권을 보여 준다 했다. 19세기 유럽에서는 국가도서관들이 예산, 인원, 책의 수, 보유의 질과 관련하여 서로 경쟁하였고, 국가도서관은 승리하는 국가의 표현이었다. 이제 세계 각국의 국가도서관들은 AI 시대를 맞아 AI 도구를 활용하여 디지털 전환을 강화하는 방안을 모색하기 시작했다. 자관의 장서를 디지털화하며 데이터를 확보하고 AI 기술을 지원하는 등 새로운 경쟁에 들어서고 있다. 최근 미의회도서관 및 유럽 국가도서관 13개관을 대상으로 한 조사에 따르면 북유럽 도서관들이 AI 도입에 앞서 있으며, 목록작성, 태깅, 역사적 문서를 디지털 텍스트 문서로 변환하는 실험을 유럽의 다른 지역보다 더 일찍 시작한 것으로 밝혀졌다.[9] 국립중앙도서관 역시 일찍이 1995년부터 소장자료 디지털화를 추진하였다. 도서관 소장자료 뿐 아니라 박물관, 미술관 등 문화예술기관, 학술기관의 자료를 디지털화하고, AI 기술을 활용하여 데이터화를 추진해왔다. 또한 연구 지원 및 데이터 활용을 높이기 위하여 데이터 큐레이션, 데이터셋 개발 등 데이터 허브로서 다양한 노력을 기울이고 있다.

4. 해외 사례

4.1 미의회도서관 (Library of Congress)

미의회도서관은 1억 7천만점이 넘는 세계 최대 규모의 장서를 보유하고 있으며. 조지 워싱턴에서 캘빈 쿨리지에 이르기까지 23명의 대통령 문서를 포함한 수백만 건의 장서를 디지털화하였다. 또한 디지털 컬렉션을 텍스트 데이터로 변환하는 크라우드소싱 프로그램 'By the People'을 운영하며, 전 세계 시민의 참여를 통해 수십만 페이지의 문서를 사람과 기계가 읽을 수 있도록 하여 데이터 활용도를 높이고 있다. 또한 미의회도서관은 다양한 디지털 컬렉션과 광범위한 아카이빙을 통해 고유한 데이터 자원을 축적해왔으며, 이를 API와 오픈 데이터셋 형태로 외부 연구자 및 기관에 개방하여 AI 학습데이터 활용도 지원하고 있다. 특히 LC Labs은 미의회도서관의 디지털 전환을 지원하며, AI 및 머신러닝 기술을 실험하고 데이터 접근성과 활용성을 강화하는 다양한 프로젝트를 수행하고 있다. 미국 의회도서관은 방대한 데이터 자원을 바탕으로 AI와 머신러닝 기술을 실험하고 데이터의 접근성과 활용성을 강화하며 국가지식문화자원의 핵심 데이터 허브로서의 역할을 적극 수행하고 있다.

4.1.1. 미의회도서관 데이터서비스 예시

(1) Chronicling America[10]

1789년부터 1963년까지 발행된 미국의 역사적 신문 수백만 페이지를 디지털화하여 공개하는 프로젝트이다. 각 신문은 고해상도 이미지와 함께 OCR(광학문자인식) 기반의 텍스트 데이터와 메타데이터 형태로 제공하며, 이를 검색, 분석, 연구에 활용할 수 있도록 API 형태로 제공하고 있다.[11]

(2) Sanborn Maps[12]

미국 의회도서관이 제공하는 대표적인 공공 지리정보 데이터 서비스로, 19세기 후반부터 20세기 초반까지 미국 전역의 도시를 상세히 기록한 화재보험용 지도 5만여 장과 고해상도 이미지 44만여 건의 데이터를 제공한다. 데이터는 메타데이터(JSON/CSV), 좌표 정보, 해상도 조절이 가능한 이미지 링크를 포함한 데이터셋으로 제공된다.[13]

표 2. 미의회도서관 데이터서비스 예시

서비스	콘텐츠 범위	디지털화 자료	데이터 형식	제공 방식
Chronicling America	미국 신문 (1789~1963)	신문 이미지 수백만 페이지	고해상도 이미지 + OCR 텍스트 + 메타데이터 (JSON/XML)	API 제공

Sanborn Maps	도시 화재보험 지도(1867~1970)	약 50,000장 지도 + 440,000개 이미지	이미지 링크 + 지도 메타데이터 (JSON, CSV) + 위치정보	데이터셋 다운로드 제공

4.2 영국국립도서관 (British Library)

영국국립도서관은 약 1억 7천만 점 이상의 자료를 보유한 세계적인 연구도서관이다. 2013년 온라인 자료 납본이 시행된 이후 1,150만 개의 전자저널 기사, 86만 권의 전자책, 1.56페타바이트의 영국 웹 도메인 페이지를 수집하였고, 장서의 1억 4천7백만 페이지 이상이 디지털화되어 제공되고 있다. 또한 영국국립도서관은 국가연구데이터 관리와 활용을 주도하고 있으며, 장서의 데이터화와 데이터 연구 혁신을 적극적으로 지원하며 데이터 허브 역할을 확대하고 있다. 특히 BL Labs은 연구자 및 개발자를 위한 실험 공간으로, 대규모 데이터셋 공개, AI 시범 프로젝트 등을 운영하고 있다.

4.2.1. 영국국립도서관 데이터서비스 예시

(1) 19th Century Books – metadata with additional crowdsourced annotations[14]

19세기 출판된 책들의 서지정보와 크라우드소싱으로 추가된 주석이 포함되었다. 연구자들이 텍스트 분석과 같은 데이터 기반 연구에 활용 가능하다.

(2) British Library 19th Century Newspapers[15]

19세기 신문 1천만 건 이상의 기사, 광고, 삽화 OCR 데이터를 제공하고 있으며 데이터셋 형식으로 다운로드 가능하며 NLP/텍스트 마이닝 실험에 활용되고 있다. 영국국립도서관은 신문 디지털화 프로젝트에서 수백만 페이지를 디지털화하는 작업을 수행하였으며, 이를 위해 상업적인 파트너와 협력하였다. 상업적 파트너는 디지털화 비용을 부담하는 대신, 디지털화된 자료를 상업적으로 활용할 권리를 얻는다.[16]

(3) Living With Machines (LWM) project[17]

이 프로젝트는 영국 예술 및 인문학 연구위원회(AHRC, Arts and Humanities Research Council)의 지원을 받아, 영국국립도서관, 앨런 튜링 연구소, 케임브리지 대학교, 이스트앵글리아 대학교, 엑서터 대학교, 퀸메리 런던 대학교, 킹스칼리지 런던 등 다수의 대학이 협력하여 수행한 대규모 디지털 인문학 연구이다. 핵심 목표는 산업혁명이 인간의 삶에 미친 영향을 데이터 기반으로 탐구하는 것으로, 약 50만 페이지에 이르는 신문 자료를 디지털화하고, 역사학자, 데이터 과학자, 언어학자, 도서관 전문가 등 다양한 분야의 연구자들이 협업하여 진행되었다. 본 프로젝트를 통해 총 35편의 연구 출판물, 강연, 논문 등을 발표하였으며, 영국국립도서관은

AI4LAM(Artificial Intelligence for Libraries, Archives and Museums) 전문가 그룹의 공동 의장으로서, 본 프로젝트를 기반으로 관련 AI 기술 개발 및 정책 환경 조성에도 기여하였다.

4.3 네덜란드국립도서관 (Koninklijke Bibliotheek)

네덜란드국립도서관은 약 4,000만 건에 이르는 디지털화된 신문, 책, 잡지 및 전자책과 웹사이트 등 방대한 디지털자원을 보유하고 있다.[18] 네덜란드국립도서관 정책계획 2023-2026(Policy plan 2023 - 2026)에서는 디지털화를 지식과 정보에 대한 접근성을 높이는 가장 강력한 방법 중 하나로 보고 디지털화를 통해 소장자료에 대한 접근성과 포용성을 높일 계획이다. 네덜란드국립도서관 데이터세트는 인공지능 연구에 유용하게 활용되고 있으며, 연구자들에게 다양한 데이터세트를 제공하고, 데이터세트 활용 촉진을 위해 서비스를 개선해 나가고 있다. 또한 네덜란드국립도서관은 2026~2030년에는 디지털화를 통해 종이 유산을 보호하는 것에서 '디지털 기반' 유산을 보호하는 것으로 장서개발의 중심축이 전환될 것이라고 예고했다. 이는 가급적이면 디지털 형태로만 소장하는 방법을 채택한 것으로, 같은 작품을 여러 형태로 복제하는 것보다는 데이터의 다양성 확보에 중점을 두었다고 볼 수 있다. 또한 네덜란드국립도서관은 "Retrotool"이라는 AI 메타데이터 생성 도구를 개발하여 전문사서들이 목록화되지 않은 장서를 정리하는 것을 도우며 고품질 데이터 생산을 지원하는 등 유럽연합 프로젝트와의 협업을 통해 AI 기반 디지털 도서관 모델 구축에 기여하고 있다.[19]. 네덜란드국립도서관은 방대한 데이터, 다양성과 신뢰성을 가진 고품질 데이터를 제공하며 데이터 허브 역할을 수행하고 있다.

4.3.1. 네덜란드국립도서관 데이터서비스 예시

(1) Delpher[20]

과학기관, 도서관 및 문화유산 기관 100만 권 이상의 책, 200만 개 이상의 신문, 1,200만 건의 잡지 페이지, 150만 개의 라디오 뉴스가 디지털화되었다. 이미지, 메타데이터 및 텍스트(OCR, ALTO, XML) 데이터가 제공되며 연구자들의 텍스트 분석 및 데이터 마이닝에 활용 가능하다.

(2) Digital Library for Dutch Literature (DBNL) dataset[21]

데이터 세트에는 500만 페이지 이상의 디지털화된 네덜란드어 및 문학이 포함되어 있고, 전문(PHP), TEI(Text Encoding Initiative) XML 파일, txt 파일 등으로 서비스되고 있다.

앞에 살펴본 바와 같이 전 세계 주요 도서관들은 광범위한 디지털 자원을 확보하고 데이터화하여 연구와 AI 학습을 위한 데이터로 제공함으로써 데이터 허브로 자리잡고 있다.

5. 국립중앙도서관 데이터 현황

국립중앙도서관은 2024년 기준 1,468만권에 이르는 장서와 2,047만 건에 달하는 온라인 자료

를 보유하고 있다. 209만 책, 4억 5천만페이지를 디지털화하였다. 여기에는 박물관, 미술관 등 전국 문화예술기관 133개관 30.7만 책도 포함된다. 뿐만 아니라 빅데이터플랫폼 정보나루(공공도서관 대출정보 등), 학술자원 공유를 위한 OAK 운영, 학술지 저작권 정보관리 등 방대한 데이터를 수집하여 제공하고 있으며, API 및 데이터셋 제공으로 연구 및 AI 학습데이터 활용 등을 지원하고 있다. 또한 AI OCR(광학문자인식) 기술을 활용한 텍스트 데이터를 제공하고, 이용자가 직접 참여하여 OCR 텍스트 데이터를 수정하고 공유할 수 있는 "공유서재" 플랫폼을 운영하고 있다. 국립중앙도서관은 온·오프라인을 아우르는 가장 포괄적인 국가지식문화자원을 보유하고 있으며, 신뢰성 있는 메타데이터를 전문사서가 구축하고 있다. 이를 기반으로 텍스트, 이미지, 음성, 영상 등 다양한 콘텐츠 데이터를 연계하여 제공하며 국가지식문화자원의 데이터 허브로서 역할을 수행하고 있다.

5.1. 데이터 유형별 현황: 메타데이터, 콘텐츠, 이용자/업무 데이터

본 논문에서는 국립중앙도서관이 보유한 데이터를 메타데이터, 콘텐츠 데이터, 이용자/업무 데이터의 세 가지 범주로 구분하여 현황을 기술하였다. 첫째, 메타데이터에는 국가서지, 국가전거(저자명, 주제명), 출판정보(ISBN, ISSN) 등이 포함된다. 국가서지는 국립중앙도서관이 납본제도에 기반하여 도서관법에 따라 작성하는 대표적 메타데이터로, 국가지식문화자원에 대한 신뢰도 높은 지식 지도 역할을 수행한다. 둘째, 콘텐츠 데이터는 디지털 원문, 텍스트 데이터, 목차 등으로 구성되며, 최근 생성형 AI의 학습 데이터로서 그 가치가 부각되고 있다. 디지털 원문과 텍스트 데이터에 대한 수요가 증가함에 따라 메타데이터 역시 LLM의 성능 향상을 위한 주요 자원으로 재조명되고 있다. 셋째, 이용자 및 업무 데이터는 정보나루(도서관 빅데이터 시스템)의 공공도서관 장서정보, 대출정보 등으로 도서관 서비스 분석 및 지능형 추천 시스템 등에 활용 가능하다. 24년 12월 기준, 주요 데이터 규모는 다음과 같다. 메타데이터, 국가서지 2,752만 건, 국가전거 207만 건(저자정보 187만 건, 주제정보 20만건), 콘텐츠 데이터, 원문 209만 책, 4억 5,135만 면, 텍스트데이터 1,771책 6,200만자, 목차 151만건, 이용자/업무데이터 정보나루 공공도서관 1,546개관의 장서데이터 1억 8,653만건, 이용자 데이터 3,818만 건, 대출 데이터 22억 3천만 건이다.

표 3. 도서관 데이터 현황(메타데이터, 콘텐츠데이터, 이용자데이터, 2024 년 12 월 기준)

구분	세부 항목	데이터 내용
메타 데이터	국가서지	총 2,752만 건 (오프라인: 705만 건, 온라인: 2,047만 건)
	국가전거	총 207만 건 (저자명: 187만 건, 주제명: 20만 건)
	LOD(링크드 오픈데이터)	데이터 2,937만 건, 트리플 8억 11만 건
	식별체계	ISBN: 522만 건, ISSN: 2.9만 건, ISNI: 151만 건
	국가자료종합목록	서지: 2,022만 건, 소장 데이터: 8,985만 건 (전국 2,000여 개 도서관)
	한국고문헌종합목록	서지 데이터 51만 건 (국내외 고문헌 소장기관 포함)

	정책정보종합목록	서지: 182만 여건, 콘텐츠: 15만 여건 (정부부처 및 공공기관 자료)
콘텐츠 데이터	원문	209만 책, 4억 5,135만 면(문화예술기관 자료 30.7만 건 포함)
	텍스트, 그림/표 이미지	텍스트 데이터 1,771책, 6,200만자, 그림/표 이미지 13.8만건
	목차, 기사 색인/초록	목차 151만 건 (단행본: 52만 건, 연속간행물: 99만 건), 기사 색인/초록 1.14만건
이용자/업무 데이터	정보나루	전국 도서관 장서 데이터: 1억 8,653만 건, 이용자 데이터: 3,818만 건, 대출 데이터: 22억 3,217만 건
	이용자 활동 데이터	* 이용자 참여 등을 통해 생성된 데이터, 도서관 업무에서 생성된 데이터

5.2. 국립중앙도서관 데이터 서비스 예시

5.2.1. 코리안메모리 [22]

코리안메모리 프로젝트에서는 국립중앙도서관 및 전국 문화예술기관 등 133개 참여기관의 디지털자원을 인물, 사건, 장소 등 주제별로 큐레이션하고 데이터화하며 서비스를 추진하고 있다. 주요 컬렉션으로 “이현세컬렉션[23]”에서는 이현세 만화가의 원화, 구술채록 등 디지털자원에 대한 데이터셋(이미지+메타데이터 등)을 제공하며, “근현대한글잡지 컬렉션[24]”에서는 한글잡지에 대한 메타데이터, 원문 텍스트파일, 이미지 파일 등을 제공하고 있다.

5.2.2. AI OCR 텍스트 데이터 구축

국립중앙도서관은 저작권이 해결된 고서, 근현대자료 1,771책, 6,200만자를(`24년 기준) 텍스트 데이터로 구축하고, 다양한 형식의 데이터셋(텍스트 파일, 이미지, XML, JSON)을 제공하고 있다. 고서의 경우 금속활자본, 목판본, 목활자본 등 다양한 판본을 대상으로 텍스트 데이터를 구축하고 데이터셋을 서비스하고 있다. 근대잡지의 경우는 『청춘』, 『조광』, 『문장』, 『카톨릭청년』 4종(일부 권호에 적용) 잡지에 대한 텍스트 데이터 및 표, 그림, 만화, 광고, 사진 등 이미지 데이터를 구축하여 데이터셋(XML 파일, Text 파일, JSON 파일)을 제공하고 있다.

5.2.3. 공유서재 [25] (2025 년 상반기 오픈 예정)

크라우드소싱 기반 텍스트데이터 구축 서비스로 저작권이 해결된 디지털원문에 대한 AI OCR 텍스트데이터를 이용자가 직접 수정 및 다운로드하여 활용한다, 이용자가 수정한 텍스트데이터를 공유할 수 있다.

5.2.4. 대한민국신문아카이브 [26]

1883년부터 1966년까지 신문 103종의 기사(823만 건)와 색인(1,647만 건)을 구축하여 서비스하고 있다. 소재 99만 건, 인명 25만 건, 단체 20만 건, 장소 348만 건에 대한 검색이 가능하며, 주제 및 광고 200만 건, 사진 6만 건 등 유형별 접근이 가능하다.

5.2.5. 도서관 빅데이터플랫폼 도서관정보나루 [27]

1,530개 국내 공공도서관의 장서, 대출, 회원 데이터 등 도서관 이용데이터를 제공하고 있다. 지역별 비교분석 데이터, 도서별 이용분석 데이터 등의 데이터셋 및 API를 제공하고 있다.

5.2.6. OAK 국가리포지터리 [28]

OAK는 국립중앙도서관이 운영하는 국가리포지터리로 국립중앙도서관을 포함한 국내 63개 기관의 OA자료 136만 건과 OA 학술지 131종, KJCI 국내학술지 3,117종의 저작권정보도 확인할 수 있다.

6. 국립중앙도서관 데이터 특징

국립중앙도서관의 데이터는 다음과 같은 특징을 가진다. 첫째, 신뢰성과 포괄성이다. 국립중앙도서관은 데이터 표준화를 기반으로 전문사서들이 신뢰성 있는 데이터를 완전하고 포괄적 구축 제공하고 있다. 국가지식문화자원의 목록인 국가서지와 국가전거(이름, 표제, 주제 등)를 작성하고 국제표준식별자 ISBN(국제표준도서번호), ISSN(국제표준연속간행물번호), ISNI(국제표준이름식별자) 운영기관으로 국가지식문화자원 공유와 융합의 중심기관이다. 이는 국내에서 가장 완결도 높은 메타데이터 체계로, 국가대표도서관으로서의 법적·제도적 기반 위에서 수행되는 데이터 구축 작업에 기인한다. 또한 메타데이터와 결합된 텍스트데이터, 이미지데이터 등 콘텐츠 데이터는 AI 학습에 강력한 고품질 데이터로 그 중요성은 강조되고 있다. 둘째, 다양성과 확장성이다. 국립중앙도서관은 단행본, 연속간행물, 신문, 고문헌뿐 아니라 영상, 음원, 이미지, 정책자료 등 다양한 유형의 자료를 수집하고, 이를 시대와 주제를 초월해 아우르는 데이터로 구성하고 있다. 콘텐츠 형태도 텍스트, 이미지, 표, OCR 데이터, 메타데이터 등으로 확장되어 융합연구와 활용이 가능하다. 셋째, 접근성과 개방성이다. 국립중앙도서관은 국가지식문화자원 데이터를 사람과 기계 모두가 접근 가능하게 하고, FAIR 원칙(Findability, Accessibility, Interoperability, Reusability) 준수를 위해 노력하며 관리하고 있다. API 제공, 데이터셋 다운로드 등 다양한 방식으로 데이터를 외부에 개방하며, 이용자가 참여하여 디지털원문 OCR 텍스트데이터를 수정하고 공유하는 공유서재 플랫폼 운영 등 이용자 참여형 데이터 구축도 추진하고 있다. 넷째, 협력성과 활용성이다. 국내 도서관, 학술기관, 문화예술기관 및 해외 도서관과 협력하고 데이터를 교류하고 있다. 다섯째, AI 학습에 적합한 고부가가치 데이터라는 점이다. 2020년 이전에 출판된 방대한 장서는 대부분 인간에 의해 생산된 원저작물로 구성되어 있다. 이는 AI 언어모델 학습에 있어 신뢰성 있는 텍스트 자원으로 활용 가능성이 매우 크며, 국립중앙도서관 데이터는 LLM(대형언어모델)의 성능 향상을 위한 학습데이터로도 높은 활용 가치를 가진다.

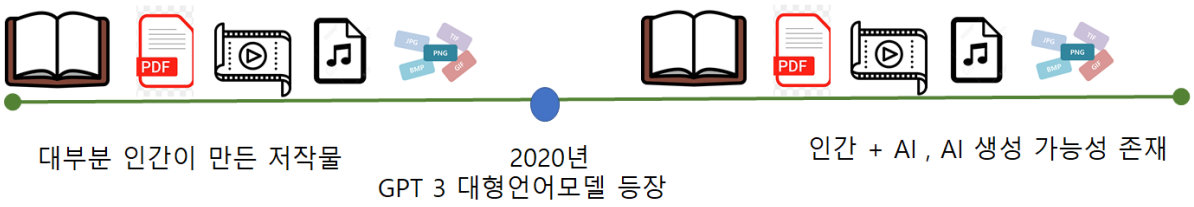

그림 1. AI 시대, 국립중앙도서관 데이터의 전략적 가치

7. 결론

오늘날 사회는 AI 기술로 급격한 전환기에 접어들고 있으며, AI 기술은 산업 전반에 걸쳐 혁신의 핵심 동력이 되고 있다. AI 기술 발전에 있어 신뢰성 있는 대규모 학습데이터의 중요성은 점점 높아지고 있으며, 이는 이제 단순히 기술 기업의 과제를 넘어 국가 차원의 경쟁력 요소로 부상하고 있다. 이러한 맥락에서, 국립중앙도서관은 신뢰성과 전문성을 갖춘 국가 지식문화자원의 데이터 허브로서, AI 시대 국가 경쟁력 강화를 위한 핵심 기반을 제공할 수 있다. 본 논문은 국립중앙도서관이 데이터의 허브로서 향후 지향해야 할 방향성을 다음의 네 가지로 정리하였다. 첫째, 디지털 자원 확대 및 데이터화 강화이다. 국가지식문화자원 특히 디지털자원의 수집을 강화하고, 기존 인쇄자료뿐 아니라 오디오, 영상 등 비도서 자료에 대한 디지털화를 더욱 확대하고, 비정형 텍스트 및 이미지 데이터 구축 등 데이터화를 적극 추진한다. 둘째, 데이터 접근성과 활용성의 제고이다. 다양한 API 제공, 데이터셋 다운로드 서비스 등 사람과 기계 모두에게 데이터 접근성과 활용 편의성을 높인다. 셋째, AI 기술을 활용한 서비스 혁신 및 연구 지원 강화이다. 도서관 업무 전반에 AI 기술을 적용하여 데이터 구축의 효율성을 높이고 서비스를 혁신해 나가며, 데이터 분석 도구 제공 등 데이터 기반 연구 지원을 강화해 나간다. 넷째, 국내외 협력과 글로벌 연계의 확대이다. 국내의 도서관, 학술기관, 문화예술기관뿐 아니라 해외 주요 기관들과의 데이터 공유, 메타데이터 표준 연계 등을 통해 글로벌 협력을 강화해 나간다. 이러한 전략을 지속적으로 추진함으로써, 국립중앙도서관은 AI 시대 국가 데이터 전략의 핵심 축으로서, 데이터 허브로서의 위상을 확립해 나갈 수 있을 것이다.

---

[1] Sagan, Carl. 홍승수 역(2014). <코스모스>. 사이언스북스. 554-557. http://lod.nl.go.kr/resource/KMO200702553

[2] Moller, Violet. 김승진 역(2023). <지식의 지도>. 63. 마농지. http://lod.nl.go.kr/resource/KMO202344235

[3] Vallejo, Irene. 이경민 역(2023). <갈대속의영원>. 반비. 75-76. http://lod.nl.go.kr/resource/KMO202328418

[4] Chalmers, M. K., & Edwards, P. N.(2017). “Producing “one vast index”: Google Book Search as an algorithmic system“. Big Data & Society, 4-2, 1-16. https://doi.org/10.1177/20539517177169.

[5] Brown, Tom B., et al. (2020). Language Models Are Few-Shot Learners. arXiv preprint arXiv:2005.14165v4. 9. https://arxiv.org/abs/2005.14165.

[6] Padilla, Thomas, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. “Santa Barbara Statement on Collections as Data --- Always Already Computational: Collections as Data”. Zenodo, https://doi.org/10.5281/zenodo.3066209.

[7] Padilla, Thomas, Hannah Scates Kettler, Stewart Varner, and Yasmeen Shorish. “Vancouver Statement on Collections as Data”. Zenodo, 13 Sept. 2023, https://doi.org/10.5281/zenodo.8342171.

[8] Vitiello, G.(2001). “National libraries: the concept and the practice, 1700-2000”. Alexandria. 13-3.

139-151. https://doi.org/10.1177/095574900101300302.
[9] Henry, A., Molinari, F., Peralta, A., Scupola, A., & van der Vaart, B.(2024). [Deliverable] D3. 2 Digital Transformation and ICT v. 2.0. D3. 2 Digital Transformation and ICT v2. 0. Horizon Europe Project LibrarIN. 5. https://librarin.eu/resources/public-deliverables/
[10] Library of Congress, "Chronicling America". https://chroniclingamerica.loc.gov/
[11] Library of Congress, "Chronicling America". https://chroniclingamerica.loc.gov/about/api/
[12] Library of Congress, "Sanborn Maps". https://www.loc.gov/collections/sanborn-maps/
[13] Library of Congress, "Sanborn Maps Data Package". https://data.labs.loc.gov/packages/
[14] British Library. "19th Century Books–metadata with additional crowdsourced annotations". https://labs.biblios.tech/item/19th-century-books-metadata-with-additional-crowdsourced-annotations/
[15] British Library. "British Library 19th Century Newspapers". https://www.digitisednewspapers.net/histories/bl/
[16] British Library(2024). The British Library Annual Report and Accounts 2023–2024. 103.
[17] British Library(2024). The British Library Annual Report and Accounts 2023–2024. 14-15.
[18] National Library of the Netherlands. "About us". https://www.kb.nl/en/about-us
[19] Henry , A., Molinari, F., Peralta , A., Scupola, A., & van der Vaart , B.(2024). D3.2 Digital Transformation and ICT v2.0 Horizon Europe Project LibrarIN. 37. https://librarin.eu/resources/public-deliverables/
[20] National Library of the Netherlands. "Delpher". https://www.delpher.nl/
[21] National Library of the Netherlands. "Digital Library for Dutch Literature (DBNL) dataset". https://www.kb.nl/en/research-find/datasets/dbnl-dataset
[22] 국립중앙도서관. "코리안메모리". https://koreanmemory.nl.go.kr/
[23] 국립중앙도서관. "이현세". https://koreanmemory.nl.go.kr/USER/contents/SV0301000.do?page=&schM=view&editorId=12&dataIdx=2080317
[24] 국립중앙도서관, "근현대 한글 관련 잡지". https://koreanmemory.nl.go.kr/USER/contents/SV0301000.do?page=&schM=view&editorId=12&dataIdx=2079101
[25] 국립중앙도서관. "공유서재". https://www.nl.go.kr/aiocr/
[26] 국립중앙도서관. "대한민국 신문 아카이브". https://www.nl.go.kr/newspaper/index.do
[27] 국립중앙도서관. "도서관정보나루". https://www.data4library.kr/
[28] 국립중앙도서관. "OAK 국가리포지터리". https://oak.go.kr/main/main.do

참고문헌

Sagan, Carl. 홍승수 역(2014). <코스모스>. 사이언스북스. http://lod.nl.go.kr/resource/KMO200702553

Moller, Violet. 김승진 역(2023). <지식의 지도>. 마농지. http://lod.nl.go.kr/resource/KMO202344235

Vitiello, G. (2001). "National libraries: the concept and the practice, 1700–2000". *Alexandria.* 13-3. 139-151. https://doi.org/10.1177/095574900101300302

Henry, A., Molinari, F., Peralta, A., Scupola, A., & van der Vaart, B. (2024). [Deliverable] D3. 2 Digital Transformation and ICT v. 2.0. D3. 2 Digital Transformation and ICT v2. 0. *Horizon Europe Project LibrarIN.* https://librarin.eu/resources/public-deliverables/

Library of Congress (2023). *Library of Congress Annual Report of the Librarian of Congress 2023.* https://www.loc.gov/static/portals/about/reports-and-budgets/documents/annual-reports/fy2023.pdf

National Library of the Netherlands (2023). KB Policy Plan 2023–2026. https://www.kb.nl/sites/default/files/documents/tg_-_kb_beleidsplan2326_en_v3_11072023.pdf.