JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 12, pp. 3089-3097, Dec. 2023

도서관 관련 공공데이터의 관리와 활용에 대한 탐색적 분석

송 채 은[1] · 김 학 래[2*]
[1]중앙대학교 문헌정보학과 박사과정
[2*]중앙대학교 문헌정보학과 교수

Exploratory Analysis on the Management and Utilization of Public Library Dataset

Chaeeun Song[1] · Haklae Kim[2*]

[1]Master's Course, Department of Library and Information Science, Chung-Ang University, Seoul 06974, Korea
[2*]Professor, Department of Library and Information Science, Chung-Ang University, Seoul 06974, Korea

[요 약]

도서관은 데이터 기반 의사결정을 지원하기 위해 새로운 핵심 정보자원으로서 도서관 관련 데이터를 제공하고 있다. 도서관 관련 데이터는 도서관의 장서목록, 사용자 통계, 위치 정보 등 다양한 정보를 공개하고 있으며, 주로 공공데이터 형태로 제공된다. 그러나, 서로 다른 데이터 포털에 분산되어 있어 일관된 관리체계를 마련하기 어려우며, 데이터의 접근성을 저해한다. 본 연구는 분산된 도서관 데이터를 통합적으로 분석하기 위해 19개 데이터 포털을 대상으로 개방 현황을 조사하고, 데이터의 관리 현황과 활용성을 정량적으로 분석하는 방안을 제안한다. 첫째, 도서관 서비스를 개선하기 위한 지표로 개별 데이터 포털의 분류체계와 메타데이터를 표준화해 데이터 포털별 특성을 조사한다. 둘째, 수집한 메타데이터 중 '수정일', '업데이트 주기', '조회 수', '다운로드 수'를 활용한 탐색적 분석을 수행한다. 연구 결과, 4,848건의 도서관 데이터를 수집했으며, 도서관 데이터의 메타데이터와 분류체계를 통합하여 데이터 포털별 업데이트 준수율과 활용도 평가를 진행했다.

[Abstract]

Public libraries offer key information resources for data-driven decision-making through open access to various types of library-related data, primarily in the form of public data. However, due to dispersion across different data portals, establishing a consistent management system proves challenging, thereby impeding data accessibility. This study comprehensively analyzed 19 data portals to assess the current status of openness, aiming to integrate dispersed library data systematically. First, proposing the standardization of classification systems and metadata across data portals, we investigated library data characteristics. Second, we conducted an exploratory analysis using collected metadata, focusing on parameters such as "modification date," "update cycle," "number of views," and "number of downloads" to evaluate the dynamic aspects of library data. The study resulted in the collection of 4,848 library data, the integration of metadata and the classification system, and the evaluation of update compliance and utilization by data portal.

색인어 : 도서관, 데이터 포털, 공공데이터, 도서관 데이터, 데이터 분석
Keyword : Library, Data Portal, Public Data, Library Data, Data Analysis

http://dx.doi.org/10.9728/dcs.2023.24.12.3089

Received 18 October 2023; **Revised** 23 November 2023
Accepted 24 November 2023

***Corresponding Author; Haklae Kim**

Tel: +82-2-820-5561
E-mail: haklaekim@cau.ac.kr

Ⅰ. 서 론

거대언어모델을 활용한 생성형 인공지능 기술의 발전은 다양한 분야의 정보서비스를 변화시키고 있다. ChatGPT는 일반지식에 대한 질의응답으로 시작하여 코드 생성과 도메인에 차별화된 영역으로 서비스 범위를 확대하고 있다. 예를 들어, 생성형 인공지능을 적용한 정보서비스는 사용자의 질문에 대한 답변, 관련 있는 정보의 요약을 포함하는 기능을 제공한다. 도서관에서 생성형 인공지능의 활용은 광범위할 수 있다[1]. 메타데이터의 입력과 관리와 같은 노동집약적 업무는 적정 수준의 인공지능 기술을 통해 보완할 수 있고, 서지 정보의 요약이나 추천은 비교적 빠른 시간에 확산될 수 있는 정보서비스의 유형이다[2],[3]. 이러한 데이터 기반의 접근은 도서관이 보유한 다양한 자료와 사용자의 행동 패턴 등을 종합적으로 이해하고 분석할 수 있게 한다. 전통적으로 도서관은 장서 관리와 온라인으로 제공하는 전자자료까지 광범위한 정보서비스를 포함하고 있고, 그 핵심에 체계화된 데이터가 있다. 데이터 중심의 도서관 서비스가 시작됨에 따라 관련 연구도 지속적으로 수행되고 있다. 문화체육관광부와 한국과학기술정보연구원은 2010년부터 2017년까지 '도서관 빅데이터 분석활용 체계 구축' 연구를 통해 도서관 빅데이터를 활용하는 방안을 모색하고, 데이터 중심의 도서관 환경 변화와 사서의 업무 생산성 향상을 이끌어 냈다. 일반적으로 도서관 데이터는 도서관의 소장정보, 서지정보, 대출 정보, 이용자 정보와 관련된 데이터로 구분한다. 이수상[4]은 도서관 데이터를 '도서관 정형 데이터', '도서관 반정형 데이터', '도서관 비정형 데이터'로 구분하고, 온정미, 박성희[5]는 도서관 빅데이터의 개념을 문헌정보학 영역에서 다루어지는 생성 주체에 따라 원시 데이터, 콘텐츠 데이터, 사회적 데이터 3가지로 구분하고 있다.

최근 도서관 관련 데이터는 공공데이터포털 또는 지방자치단체의 데이터 포털을 통해 공개되고 있다. 국립중앙도서관은 도서관 빅데이터 플랫폼인 '도서관 정보나루'를 운영하고 있다. 도서관 정보나루는 전국 공공도서관 1,490개의 소장목록과 대출빈도에 대한 데이터세트를 주기적으로 관리하고, 수집한 도서관 데이터세트를 제공한다. 공공데이터포털, 지방자치단체별 데이터 포털, 문화체육관광부의 문화 공공데이터광장, 한국문화정보원의 문화 빅데이터 플랫폼은 대규모 도서관 데이터를 제공하는 대표적인 정보원이다.

도서관 데이터는 도서관의 운영과 정보 서비스를 이해하는 핵심 자산으로 인식된다. 그러나, 도서관 데이터는 서로 다른 플랫폼에서 파편적으로 제공되고 있고, 제공하는 메타데이터 요소와 데이터 값을 정의하는 체계가 미흡하다. 예컨대, 도서관 데이터는 도서를 분류하기 위한 일관적인 분류체계를 적용하지 않고 있다. 개별 데이터 포털은 표준화된 데이터 체계가 없기 때문에 일관적인 데이터의 관리와 연계가 쉽지 않다[6],[7].

본 연구는 서로 다른 도서관 데이터를 연계·활용하기 위한 방안을 제안한다. 이를 위해 국내에서 활용되고 있는 도서관 데이터 현황을 검토한다. 특히, 도서관 데이터의 관리와 활용을 정량적으로 분석할 수 있는 지표를 통해 도서관 데이터의 활용을 위한 개선 방안을 제안한다.

본 논문의 구성은 다음과 같다. 2장은 도서관 데이터와 데이터 포털에 대한 연구를 소개하고, 3장은 데이터 수집과 메타데이터 분석을 포함한 연구 방법을 요약한다. 4장은 도서관 데이터의 개방현황과 활용도를 정량적 지표로 분석한 결과를 요약한다. 5장은 연구 결과와 공헌점을 기술하고, 향후 연구를 소개한다.

Ⅱ. 관련 연구

웹상에 개방된 도서관 데이터는 데이터 개방 현황, 데이터 제공기관, 데이터 품질 평가 등 다양한 관점에서 연구되고 있다. 조재인[8]은 공공데이터포털에 개방된 도서관 데이터 현황을 조사하기 위해 도서관 데이터의 공개 주체를 공공기관과 지방자치단체로 구분해 분석했다. 연구 결과, 2018년 기준 공공데이터포털에 개방된 도서관 데이터의 약 70%는 지방자치단체에서 개방하지만, 수요 관점에서 전국 단위의 데이터를 보유한 공공기관의 데이터가 비교적 사용성이 높음을 설명한다. 따라서, 일반적인 도서관 현황에 대한 데이터보다 이용자의 수요에 맞춘 데이터를 발굴할 것을 제안한다. 김혜선, 김완종[9]은 도서관 정보나루에서 서비스하는 API 데이터를 수집해 데이터의 완전성과 정확성을 검증했다. 데이터 품질 진단 결과, 데이터의 공백값과 부정확한 값이 포함되어 있어 품질 개선을 위한 데이터 스키마를 표준화할 것을 제안한다. 도서관 데이터의 통합된 체계를 마련하기 위한 연구도 이루어지고 있다. 박진호[10]는 도서관이 다뤄야 할 새로운 지식정보자원으로 데이터세트를 제안하고, 데이터 카탈로그를 위한 온톨로지 어휘인 DCAT (Data Catalog Vocabulary)을 도서관 데이터에 적용했다. DCAT 기반의 디지털 도서관 데이터세트 서비스를 설계하기 위해 어휘의 클래스와 속성을 분석하고, 외부 데이터를 연계할 수 있는 이용자 서비스를 제안한다.

도서관 데이터는 도서관 정보나루, 공공데이터포털뿐 아니라 개별 지방자치단체에서 운영하는 데이터 포털을 통해서도 제공된다. 선행된 도서관 데이터 연구는 대부분 하나의 데이터 포털을 대상으로 수행되었기 때문에, 분산적으로 존재하는 도서관 데이터의 통합적인 관리를 위해 데이터 포털과 데이터세트 수준의 연구가 필요하다. 데이터 포털은 이용자와 데이터 제공자의 상호작용 지점이자, 데이터 활용을 위한 기반이 된다.

데이터 포털의 증가는 데이터 획득의 용이성을 증진시키지만, 데이터의 통합성과 일관성 유지가 어렵다는 단점이 있다. 이를 개선하기 위해, 국내외 개방 데이터 포털(Open Data

Portal)에 대한 연구가 다양한 관점에서 수행되고 있다. Anastasija Nikiforova, Keegan McBride[11]는 개방 데이터 포털 사이의 상호운용을 개선하기 위한 연구를 진행했다. 41개 데이터 포털을 선별한 뒤, 데이터 포털의 메타데이터를 수집해 3가지(Open dataset specification, Open dataset feedback, Open dataset request) 범주로 나누어 평가 프레임워크를 설계했다. 설계한 지표를 기준으로 서로 다른 메타데이터를 연계했기 때문에, 사용자 관점에서 데이터 포털을 정량적으로 평가할 수 있는 기반을 마련했다.

데이터 포털은 데이터를 개방하는 것뿐 아니라, 데이터 활용을 위한 관리체계를 마련해야 한다[12]. Vetrò et al.,[13]은 개방형 데이터 포털의 데이터 품질을 평가하기 위해 7가지 지표(Traceability, Currentness, Expiration, Completeness, Compliance, Understandability, Accuracy)를 제안했으며, 송채은, 김학래[14],[15]는 Expirtation(만료) 지표를 활용해 데이터 포털과 데이터세트의 업데이트 준수율을 평가했다. 공공데이터포털과 지방자치단체별 데이터 포털의 공공데이터를 대상으로, '수정일', '업데이트 주기' 메타데이터를 활용해 데이터 포털별 업데이트 준수율을 평가했다. 윤상오, 현지우[16]는 공공데이터포털의 국가중점데이터를 대상으로 데이터 개방 실태를 분석했다. 가용성(데이터의 양, 데이터의 다양성), 사용용이성(데이터 제공유형), 활용도(데이터의 조회 수 순위와 다운로드 수 순위) 영역에 초점을 두고 기술통계방법을 활용한 평가 프레임워크를 설계했다. 김동준 외[17]는 공공데이터포털에서 제공하는 데이터세트를 지방자치단체 관점에서 분석하고, 데이터 활용도(DPV, Download Per View)와 관리 현황을 평가해 개선방안을 제안했다. 데이터 활용도는 사용자의 조회와 다운로드 수치로 측정하고, 데이터의 관리 현황은 등록일, 수정일, 업데이트 주기를 분석해 진단했다.

국내외 도서관은 서지정보에 대한 통합적 관리를 위한 연구는 활발하게 진행되고 있지만[18]-[24], 데이터형태의 자원을 관리하기 위한 방안은 미흡한 실정이다. 데이터는 도서관의 새로운 핵심 자원이 될 수 있기 때문에 이를 관리하기 위한 메타데이터 표준과 연계방안이 필요하다. 본 연구의 목적은 서로 다른 데이터 포털에 존재하는 도서관 데이터를 통합적으로 관리하고 활용하기 위한 기반을 마련하는 것이다. 따라서, 도서관 데이터의 개방 현황을 조사하고, 데이터 분류체계에 따른 특성을 분석한다. 데이터 포털별 관리와 활용성 분석은 메타데이터를 표준화한 뒤, 데이터의 업데이트 준수율과 활용도를 정량적으로 평가한다.

Ⅲ. 연구방법

그림 1은 도서관 데이터 수집과 개방 현황 분석에 대한 일련의 과정을 도식화한 것이다. 먼저, 도서관 관련 공공데이터를 수집하기 위해 데이터 포털을 선별한다. 선별된 19개의 데이터 포털에서 전체 데이터세트를 크롤링 방식으로 수집하고 (242,767건), 도서관 관련 데이터를 추출하기 위해 데이터의 제목, 설명, 키워드에 '도서'와 '도서관'이 존재하는 데이터를 필터링한다. 도서관 데이터의 관리와 활용을 분석하기 위해, 개별 데이터 포털의 상이한 메타데이터와 분류체계 용어를 의미상 같도록 매핑하는 작업을 수행한다.

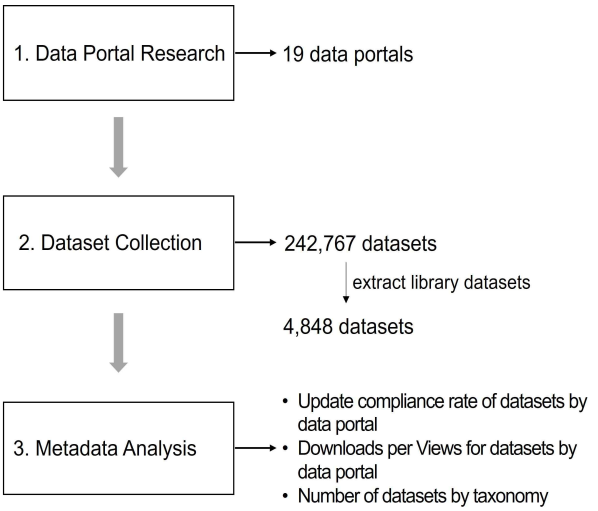

그림 1. 도서관 데이터 수집과 분석
Fig. 1. Collection and analysis of library data

표준화된 메타데이터와 분류체계는 도서관 데이터의 현황을 분석하는 기반이 된다. 매핑된 분류체계를 통해 도서관 데이터의 특징과 분류체계별 데이터 개방 현황을 분석한다. 수집한 메타데이터는 데이터의 관리와 활용을 평가할 수 있는 지표가 된다. 수집한 메타데이터 중 '수정일', '업데이트 주기'는 데이터세트가 주기적으로 관리되는지 파악할 수 있는 정보다. 데이터세트의 업데이트 준수율은 데이터 서비스 측면에서 데이터 최신성을 유지하는 기준으로 평가할 수 있다. 송채은, 김학래[15] 연구를 기반으로 '만료(Expiration)' 지표를 참조해 도서관 데이터를 평가한다. '만료'는 데이터세트의 업데이트가 만료된 비율을 계산하는 것으로, 데이터세트가 업데이트 되는 주기(예: 주, 월, 년)와 데이터세트의 게시일을 비교한다. 본 연구는 선행연구를 참조해 지표의 대상을 복수의 데이터 포털로 확장하여 업데이트 준수율을 평가한다(수식 1 참고).

$$U_{portal} = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} PU_i & \text{if}\,(SD_i - CD_i \geq 1) \\ DU_i & otherwise \end{cases} \quad (1)$$

U_{portal}은 데이터 포털의 평균 업데이트 준수율이고, SD_i (the date that scheduled for a Dataset update)와 CD_i (the date that a Dataset is collected)는 데이터세트의 업데이트 예정일과 데이터를 수집한 날짜를 의미한다. $SD_i - CD_i$의 값이 1보다 같거나 큰 경우, 업데이트 보류 중인 데이터(PU_i, pending registration)로 해석할 수 있다. 반대의 경우,

업데이트 주기를 미준수한 데이터(DU_i, delayed update)이다.

데이터 포털은 사용자의 목적에 적합한 데이터를 제공할 수 있도록 사용자가 데이터의 가치를 평가할 수 있는 지표를 제공해야 한다[21]-[24]. 예를 들어, 데이터 포털은 가장 많이 검색된 데이터세트, 상위 다운로드 수 10개 데이터세트에 대한 정보를 제공해 데이터 활용도를 표현해야 한다. 본 연구에서 데이터 활용도는 '조회 수', '다운로드 수'를 통해 이용자가 데이터를 조회하고 다운로드하는 일련의 과정을 정량적으로 분석한다. 송채은, 김학래[15]는 데이터 활용도를 정량적으로 평가하기 위해 조회 수 대비 다운로드 수(Download Per View, DPV)를 계산한 평가지표를 제안했다.

$$DPV_{portal} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{V_i}\right)\times 100 \quad (2)$$

DPV_{portal}은 데이터 포털이 제공하는 데이터세트의 조회 수 대비 다운로드 수를 산출하는 데이터 활용률이다(수식 2 참고). D_i와 V_i는 각각 다운로드 수와 조회 수를 의미한다. 데이터 포털이 제공하는 DPV_{portal}은 모든 데이터세트의 $DPV_{dataset}$ 평균으로 계산된다.

Ⅳ. 연구 결과

4-1 도서관 데이터 개방 현황

2022년 8월 1일 기준으로 데이터 포털이 보유한 전체 데이터 수량, 도서관 데이터 수량, 메타데이터와 분류체계 현황을 조사한다. 데이터 포털은 총 19개로 공공데이터포털, 지방자치단체가 운영하는 데이터 포털, 도서관 정보나루, 문화 공공데이터포털, 문화 빅데이터 플랫폼을 포함한다.

데이터 포털의 유형은 통합, 문화, 도서관 3가지로 구분할 수 있다. 첫째, 다양한 주제의 데이터를 제공하는 통합형으로 공공데이터포털과 지방자치단체가 운영하는 데이터 포털이 존재한다. 공공데이터포털은 행정안전부에서 관리하는 데이터 포털로 공공기관이 보유·관리하는 데이터를 제공한다. 지방자치단체는 총 15개의 데이터 포털을 관리하고 있다. 대전광역시와 세종특별자치시는 데이터 포털을 개별적으로 운영하지 않아 제외됐다.

둘째, 문화와 관련된 데이터만 제공하는 문화 유형은 문화 빅데이터 플랫폼, 문화 공공데이터광장이 해당한다. 문화와 관련된 데이터만 제공하는 포털인 문화 공공데이터광장과 문

표 1. 데이터 포털별 데이터세트 개방 현황
Table 1. Dataset provided by each data portal

Data potal	The number of total datasets	The number of library datasets	The number of metadata element	URL
Public data portal	98,996	2,318	26	https://data.go.kr
Gangwon data portal	497	1	13	https://data.gwd.go.kr/index
Gyeonggi data portal	1,513	20	12	https://data.gg.go.kr/portal/mainPage.do
Gyeongnam data portal	2,154	39	17	https://bigdata.gyeongnam.go.kr/index.gn
Gyeongbuk data portal	1,080	5	4	https://gb.go.kr/Main/open_contents/section/datastat/index.html
Gwangju data portal	86,698	245	11	https://bigdata.gwangju.go.kr
Daegu data portal	11,122	92	11	http://data.daegu.go.kr/open/main.do
Busan data portal	3,839	207	16	https://data.busan.go.kr/index.nm
Seoul data portal	5,843	62	14	https://data.seoul.go.kr/
Ulsan data portal	1,240	45	4	http://data.ulsan.go.kr/index.ulsan
Incheon data portal	4,384	61	17	https://www.incheon.go.kr/data/index
Jeonnam data portal	171	2	11	https://data.jeonnam.go.kr/index.do
Jeonbuk data portal	2,232	50	14	http://www.bigdatahub.go.kr/index.jeonbuk
Jeju data portal	790	2	6	https://www.jejudatahub.net/data/list
Chungnam data portal	148	0	4	http://www.chungnam.go.kr:8100/cnnet/board.do?mnu_cd=CNNMENU02498
Chungbuk data portal	18,865	52	11	https://data.chungbuk.go.kr/index.do
Library data portal	1,501	1,501	5	https://data4library.kr/
Culture data portal	414	51	7	https://www.culture.go.kr/data/main/main.do
Culture bigdata portal	1,280	95	6	https://www.bigdata-culture.kr
Total	242,767	4,848	-	-

화 빅데이터 플랫폼은 체육, 예술, 도서관 데이터를 개방한다. 문화 공공데이터광장은 문화체육관광부가 운영하고, 문화체육관광부 소속기관과 타 부처 기관에서 생산하는 문화 공공데이터를 통합적으로 수집해 개방하는 플랫폼이다. 문화 빅데이터 플랫폼은 공공기관과 민간기업으로 구성된 데이터 센터로 데이터의 유통거래를 목적으로 운영한다.

마지막으로, 도서관 데이터만 제공하는 도서관 유형의 도서관 정보나루가 있다. 도서관 정보나루는 국립중앙도서관이 운영하는 플랫폼으로, 전국 공공도서관에서 수집한 데이터를 개방하고 있다. 일반적인 데이터 포털과 다르게 도서에 대한 데이터 서비스가 중점적이다. 도서관 정보나루는 전국의 도서관 데이터를 사용해 참여 도서관 목록, 장서/대출 데이터, 인기대출도서, 도서별 이용분석, 대출 급상승 도서, 지역별 비교분석, 이달의 키워드 서비스를 제공한다.

표 1은 19개 데이터 포털에 개방된 도서관 데이터 현황을 정리한 것이다. 전체 데이터 포털 중 가장 많은 도서관 데이터를 개방한 데이터 포털은 공공데이터포털(2,318건)이고, 가장 적은 도서관 데이터를 개방한 데이터 포털은 충청남도 데이터 포털(0건)이다. 데이터 포털의 도서관 데이터 현황을 분석할 때, 도서관 데이터의 개방 수량만을 고려하는 것은 충분하지 않다. 특정 분야의 데이터는 전체 데이터 수량과 데이터 포털 유형에 따라 영향을 받기 때문이다[12].

따라서, 데이터 포털을 분류한 3가지 기준에 따라, 전체 데이터 수량 대비 도서관 데이터의 비율을 계산해 분석한다. 통합형 데이터 포털에 해당하는 16개의 포털의 전체 데이터 대비 도서관 데이터 개방 수량을 계산한 결과, 도서관 데이터의 비율은 약 1.4%다. 통합형 데이터 포털 중 도서관 데이터의 비율이 가장 높은 데이터 포털은 부산 데이터 포털(약 5.4%)이고, 낮은 비율의 데이터 포털은 충청남도 데이터 포털(0%)과 강원도 데이터 포털(약 0.2%)이다. 한편, 광주 데이터 포털은 전체 데이터 수량은 86,698건으로 19개의 데이터 포털 중 2번째로 많지만, 도서관 데이터의 수량은 245건으로 약 0.3%의 비율을 차지한다. 즉, 통합형 데이터 포털은 개방 수량에 따라 도서관 데이터가 비례하지 않는다. 다양한 분야의 데이터를 개방하기 때문에, 해당 분야와 관련된 데이터 제공기관의 방향성에 따라 도서관 데이터 개방 현황이 상이하다.

문화 유형 데이터 포털은 2개의 데이터 포털이 해당하며, 도서관 데이터는 평균 약 9.8%를 차지한다. 문화 유형 데이터 포털은 통합형 데이터 포털에 비해 문화와 관련된 데이터만 제공하므로 도서관 데이터의 비율이 비교적 높다. 문화 빅데이터 플랫폼은 문화체육관광부 산하의 공공기관인 한국문화정보원이 운영하고, 문화 공공데이터 광장은 문화체육관광부가 운영한다. 문화체육관광부는 공공도서관과 작은도서관의 예산과 정책을 집행하는 기관으로 도서관과 관련된 정보를 보유하고 있다. 문화 유형의 데이터 포털은 운영기관과 데이터 주제의 밀접한 연관성으로 인해 도서관 데이터를 타 유형의 데이터 포털에 비해 적극적으로 개방하고 있다.

도서관 유형은 도서관 정보나루만 해당하며, 도서관에 대한 정보만 제공하기 때문에 모든 데이터가 도서관 데이터에 해당한다.

표 2. 분류체계별 도서관 데이터 수량
Table 2. Library datasets quantity by category

Category	Number of datasets	Mapped classifications
Education	1,463	Education, education and employment, education/training
Cultural tourism	1,128	Cultural tourism, tourism, culture, and sports, culture and tourism, culture/leisure, culture/tourism, culture, tourism, and sports, tourism, culture and arts, cultural heritage, cultural industry, sports, books, library, policy support, environment creation, cultural promotion, promotion support
Public administration	397	Public administration, tax and legal administration, general public administration and population, general society, national accounts, government/finance, regional statistics, administrative law, women/family, administration
Land management	80	Land management, regional development, construction, land use, urban management, urban construction
Science & technology	47	Science and technology, information and communication, science/technology
Industrial employment	44	Industrial employment, industrial economy, industry/small and medium-sized enterprises, labor, mining/manufacturing, corporate management, retail/services, industry/economy
Healthcare	26	Healthcare, health
Agriculture & fisheries	19	Agriculture and fisheries, agriculture, forestry, livestock, and fisheries, agriculture, forestry, marine, and fisheries, agriculture, forestry, and fisheries
Transportation & logistics	13	Transportation and logistics, transportation, construction, and environment, transportation and traffic, trade/international balance of payments, transportation
Disaster safety	3	Disaster safety, fire and disaster safety, public order and safety, crime safety, safety
Law	2	Law
Unknown	1,626	-

4-2 도서관 데이터의 분류체계

데이터 분류체계는 데이터를 주제별로 분류해 데이터 탐색의 접근점이 되므로 데이터를 파악하는 데 중요한 요소다. 하지만, 도서관 데이터는 데이터 포털마다 다른 분류체계로 데이터를 관리하고 있으므로, 이종의 데이터 포털이 보유한 데이터를 주제별로 연계하기 위한 작업이 필요하다. 공공데이터포털은 총 16개의 분류체계를 사용해 데이터를 구분하고 지방자치단체 중 강원도, 경상남도, 대구광역시, 부산광역시, 전라북도, 충청북도는 공공데이터포털과 같은 분류체계를 사용한다. 하지만, 다른 지방자치단체는 최소 8개(경기도, 인천광역시, 제주특별자치도)에서 최대 30개(광주광역시)의 분류체계를 사용한다. 문화 공공데이터광장과 문화 빅데이터플랫폼은 같은 문화 유형 데이터 포털임에도 분류체계의 항목명이 세부적으로 다르다. 예를 들어, '문화예술', '문화유산', '문화산업', '관광', '체육'은 동일한 어휘를 사용하지만, '정책지원(문화 공공데이터광장)-여건조성(문화 빅데이터 플랫폼)', '문화홍보(문화 공공데이터광장)-홍보지원(문화 빅데이터 플랫폼)'은 상이한 용어로 표현한다.

도서관 데이터의 분류체계를 의미적으로 같은 항목끼리 매핑하여 분류체계별 데이터 수량을 집계한 결과는 표 2와 같다. 분류체계를 제공하지 않는 데이터 포털의 데이터 1,626건은 제외하고, 도서관 데이터가 가장 많은 주제 분류는 '교육(1,463)', '문화관광(1,128)'이고, '공공행정(397)', '국토관리(80)', '과학기술(47)', '산업고용(44)', '보건의료(26)', '농축수산(19)', '교통물류(13)', '재난안전(3)', '법률(2)' 순으로 존재한다. 데이터 제공기관의 특성이 데이터를 분류하는 기준으로 고려될 수 있다. 가장 많은 데이터가 포함된 '교육' 분야 도서관 데이터의 제공기관은 지방자치단체뿐 아니라 교육청 산하의 공공도서관이 존재하고, '문화관광' 분야 도서관 데이터는 문화체육관광부 산하의 공공기관이 제공기관으로 존재한다.

4-3 도서관 데이터의 관리와 활용

데이터 포털은 메타데이터(예: 파일데이터명, 제공기관 등)를 통해 데이터를 관리하고, 이용자에게 데이터 정보를 제공한다. 분류체계와 마찬가지로, 도서관 데이터를 관리하는 메타데이터 표준이 존재하지 않아 데이터의 상호운용이 어렵다. 데이터 포털의 데이터 제공 형태와 메타데이터 체계가 서로 다르므로 분산적으로 존재하는 도서관 데이터의 관리현황을 파악하는 데 어려움이 있다. 공공데이터포털은 데이터 형식에 따라 약 24개의 메타데이터를 제공하고, 도서관 정보나루는 8개의 메타데이터를 제공한다. 데이터 포털에 따라 달라지는 메타데이터는 웹상에 분산된 데이터를 일관적으로 탐색할 수 없도록 접근성을 저해한다. 예를 들어, 공공데이터포털은 데이터를 등록한 날짜를 '등록'으로 표현하지만 다른 데이터 포털은 '등록일', 'DATA개방일', '데이터개방일', '데이터등록일' 등 상이하게 표현한다.

그림 2는 가장 많은 메타데이터를 제공하는 공공데이터포털을 기준으로, 19개의 데이터 포털이 사용하는 메타데이터를 의미상 같도록 매핑하여 집계한 것이다. 개별 데이터 포털이 공통적으로 사용하는 메타데이터는 '제공기관(18개)', '분류체계(16)', '등록일(16개)', '수정일(15개)', '업데이트 주기(14개)', '관리부서명(12개)', '파일데이터명(11개)', '이용허락범위(10개)', '원본시스템(9개)', '관리부서 전화번호(8개)', '다운로드수(8개)', '키워드(8개)', '서비스 유형(8개)', '차기 등록 예정일(7개)', '확장자(7개)', '설명(7개)', '매체 유형(6개)', '제공 형태(6개)', '비용부과유무(5개)', '조회수(4개)', '비용부과기준 및 단위(4개)', '이용허락범위(4개)' 순으로 조사되었다.

메타데이터를 표준화해 연계하면, 데이터 포털마다 분산적으로 존재하는 데이터의 관리와 활용성을 정량적으로 평가할 수 있다. 데이터 관리 측면에서, '수정일', '업데이트 주기'는

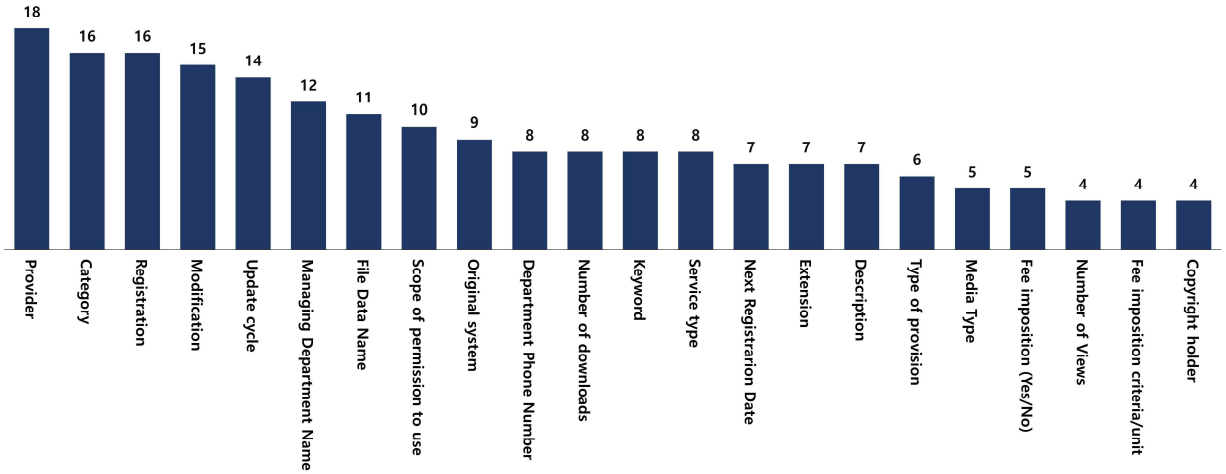

그림 2. 데이터 포털의 메타데이터 사용 현황
Fig. 2. Metadata status in data portal

데이터 최신성과 관련된 요소다. 이는 데이터 관리를 위해 업데이트 일정을 공지한 것으로, 해당 날짜에 데이터의 업데이트가 필요하다는 의미다. '업데이트 주기'를 준수한 데이터세트는 최신성을 유지하고 있으며, 해당 데이터세트에 대한 관리가 수행되고 있다고 해석할 수 있다. '수정일'과 '업데이트 주기' 메타데이터를 모두 제공하는 13개의 데이터 포털을 대상으로 업데이트 현황을 검토해 업데이트 준수율을 산출한 결과는 표 3과 같다. 업데이트 주기를 특정할 수 없는 데이터('수시', '일회성', '알수없음' 등)는 제외하고, '매일', '매주', '매달', '반년', '분기', '매년'으로 구분해 분석했다. 업데이트를 준수한 데이터의 주기는 '매달(1,497건)', '매년(751건)', '반년(29건)', '분기(27건)', '매일(27건)' 순으로 집계됐고, 업데이트를 미준수한 데이터의 주기는 '매년(103건)', '매달(45건)', '반년(26건)', '분기(20건)', '매주(1건)' 순으로 집계됐다. 데이터 포털의 평균 준수율은 0.78이며, 경상남도(1), 광주광역시(1), 서울특별시(1)의 데이터 포털에 존재하는 도서관 데이터는 모든 데이터가 업데이트 주기를 준수한다. 평균보다 낮은 업데이트 준수율을 보인 데이터 포털은 공공데이터포털(0.76), 대구광역시(0.74), 경기도(0.6), 문화 빅데이터 플랫폼(0.46)이다. 낮은 업데이트 준수율을 보인 공공데이터포털, 대구광역시, 모두 업데이트 주기가 '매년'인 데이터의 준수율이 가장 낮다.

표 3. 데이터 포털의 업데이트 준수 현황
Table 3. Data portal update compliance status

Data portal	Compliant	Non-compliant	Compliance rate
Public data portal	438	135	0.76
Gangwon data portal	0	1	0
Gyeonggi data portal	9	6	0.6
Gyeongnam data portal	25	0	1
Gwangju data portal	235	0	1
Daegu data portal	32	11	0.74
Busan data portal	32	2	0.94
Seoul data portal	30	0	1
Incheon data portal	22	3	0.88
Jeonbuk data portal	12	1	0.92
Chungbuk data portal	5	1	0.83
Library data portal	1,472	24	0.98
Culture bigdata portal	19	22	0.46
Average	179	16	0.78

데이터 활용도 측면에서, '조회 수'와 '다운로드 수'는 데이터 제공기관이 이용자의 수요에 맞는 데이터를 제공하는지 판단할 수 있는 유용한 지표다. 일반적으로 '조회'는 검색 결과에서 특정 데이터세트를 선택하고 메타데이터와 인스턴스 수준의 데이터를 확인하는 정보이며, '다운로드'는 사용자가 데이터 포털에서 특정 데이터세트를 얻는 것을 의미한다. '조회'는 특정 데이터세트의 사용 여부에 대한 정보를 포함하지 않으므로, 데이터세트의 조회 수와 다운로드 수가 항상 비례하는 것은 아니다. 따라서, 데이터 활용도는 '조회 수' 대비 '다운로드 수'를 분석하여 이용자들이 데이터를 조회하고 다운로드하는 일련의 과정을 정량적으로 분석한다. 특정 데이터세트의 조회 수는 높지만, 다운로드 수가 낮다면 이용자에게 활용되지 않은 데이터로 해석할 수 있다. $DPV_{dataset}$가 높다면 이용자가 사용할 수 있거나 관심을 가질 수 있는 데이터를 효과적으로 제공하는 것이다.

'조회 수'와 '다운로드 수'를 제공하는 모두 제공하는 공공데이터포털의 DPV_{portal}를 계산한 결과, 약 0.14의 값이 도출됐다. 공공데이터포털에 존재하는 도서관 데이터는 평균적으로 100회 조회 당 1-2회 다운로드된다고 해석할 수 있다. 개별 데이터세트의 활용도인 $DPV_{dataset}$가 1보다 크거나 같은 경우, 데이터의 사전 미리보기를 수행하지 않고 다운로드하거나 조회 시, 반드시 데이터가 다운로드되는 것을 의미한다. 공공데이터포털에서 제공하는 도서관 데이터 34건의 데이터세트만 1보다 큰 $DPV_{dataset}$값을 가지고, 1보단 작은 데이터세트는 2,284건이다. 즉, 대부분의 도서관 데이터는 조회 수 대비 다운로드 수가 낮으므로 데이터 활용도가 낮다고 해석할 수 있다. 데이터세트의 메타데이터 요소로서 $DPV_{dataset}$는 활용도가 높은 데이터세트의 특징을 다차원적으로 분석할 수 있는 기반이 된다. 예를 들어, 도서관 데이터 중 $DPV_{dataset}$가 1보다 높은 데이터세트를 1개 이상 제공하는 기관은 국립중앙도서관(4건), 국립장애인도서관(3건), 경상북도교육청구미도서관(3건), 대구광역시(3건), 중랑구시설관리공단(2건)이다. $DPV_{dataset}$가 높은 데이터세트를 많이 제공하는 기관일수록 이용자의 수요와 맞는 유용한 데이터를 개방한다고 해석할 수 있다.

V. 결 론

본 연구는 국내에서 오픈 데이터로 제공하는 도서관 데이터의 현황을 분석하고, 데이터 관리와 활용을 정량적으로 평가했다. 도서관 데이터는 공공기관이 운영하는 데이터 포털 또는 빅데이터 플랫폼을 통해 공개되고 있다.

개별 데이터 포털에서 제공하는 도서관 데이터는 다른 유형의 데이터와 비교하면 전체 데이터세트의 규모가 적지만, 강원을 제외한 대부분의 데이터 포털에서 관련 데이터세트를 제공하고 있다. '데이터 수집일', '수정일', '업데이트 주기' 등 데이터의 최신성 평가는 데이터 제공자가 적절한 시점에 해당 데이터를 갱신하는 것을 의미한다. 데이터의 최신성 지표는 전체 평균이 0.78로 비교적 높은 수준이다. 반면, 데이터의 활용도(DPV)는 평균 0.14로 매우 낮은 수준이다. 요약하면, 도서관 데이터는 다양한 기관에서 제공하고 있고, 다른 데이터세트와 비교하면 적절한 시점에 관리되고 있다. 그러나, 도서관 데이터는 일부 데이터를 제외하면 활용이 되지 않고 있다.

공공데이터의 활용이 낮은 이유는 도서관 데이터에 한정되지 않는다. 공공데이터포털을 포함한 대부분의 데이터 포털에서 제공하는 데이터의 활용은 매우 낮은 수준이고, 데이터 활용 활성화를 위해 대규모 해커톤, 아이디어 발굴, 경진대회와 같은 다양한 이벤트를 시행하고 있다. 도서관 데이터는 활용 대상과 목적이 한정적이고, 개방할 수 있는 데이터가 어느 정도 범주화될 수 있다. 따라서, 도서관 데이터의 활용을 높이기 위해 데이터 관리체계를 정교하게 수립하고, 활용 대상에 맞는 데이터를 제공하는 것이 바람직하다. 도서관 데이터는 도서관이 소속된 공공기관의 데이터 포털에 개방되어 분산적이고 파편화된 특성을 갖고 있다. 한편, 개방된 데이터의 구조와 관리체계가 다르기 때문에 데이터의 연계가 어렵다는 한계가 있다. 이를 개선하기 위해, 제공하는 데이터의 구조를 표준화하고, 데이터값을 일관적으로 기술하기 위한 가이드라인을 제공할 필요가 있다. 문헌정보 분야는 더블린코어, MARC, MODS 등 다양한 형식의 메타데이터 표준을 사용하고 있는데, 공공데이터로 제공하는 데이터세트에 적용하기 위한 메타데이터 규격을 설계하고, 데이터 제공자가 공통으로 활용할 수 있는 체계를 제공해야 한다.

운영적 관점으로 보면, 데이터 제공자는 도서관 데이터를 공개하는 시점에서 동일한 분류체계 또는 키워드를 적용하는 것이 필요하다. 공공데이터포털과 지방자치단체가 운영하는 데이터 포털에서 사용하는 분류체계가 통일되어 있지 않고, 새로운 분류체계를 지원하지 않을 수 있다. 이런 경우, 키워드를 통제어와 유사하게 사용할 수 있는 방안을 검토할 필요가 있다. 기술적 측면에서 온톨로지 어휘를 적용하여 메타데이터를 기계가 읽을 수 있는 형식으로 표현하고, 서로 다른 데이터 포털에서 제공하는 데이터세트의 연계를 위한 방안을 검토해야 한다. RDF 어휘로 표현된 데이터는 서로 동일한 체계를 적용할 경우 분산된 데이터 포털에 있는 데이터라도 의미적으로 연결할 수 있기 때문에, 도서관 데이터를 통합적으로 연결하는 데 활용할 수 있다.

본 연구는 도서관 데이터의 개방 현황과 데이터의 활용에 대한 분석을 수행했다. 향후 연구는 도서관 데이터의 품질을 진단하기 위한 방안과 도서관 데이터의 연계를 위해 공통 메타데이터의 개발, 지식그래프 기술을 적용하는 방안을 포함한다. 한편, 본 논문에서 제안한 DPV 지표는 조회수와 다운로드 수를 조합하고 있어 사용자가 실제 데이터세트의 활용을 측정하는 데 한계가 있다. 따라서, 데이터의 활용 지표는 DPV 지표와 사용자가 실제 활용한 수치를 조합하는 방식으로 평가할 수 있도록 검토할 필요가 있다.

감사의 글

이 논문 또는 저서는 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A6A3A01078538)

참고문헌

[1] B. D. Lund and T. Wang, “Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries?,” *Library Hi Tech News*, Vol. 40, No. 3, pp. 26-29, 2023. http://doi.org/10.1108/LHTN-01-2023-0009

[2] P. E. Schreur, “The Use of Linked Data and Artificial Intelligence as Key Elements in the Transformation of Technical Services,” *Cataloging & Classification Quarterly*, Vol. 58, No. 5, pp. 473-485, 2020. http://doi.org/10.1080/01639374.2020.1772434

[3] S. Vijayakumar and K. N. Sheshadri, “Applications of Artificial Intelligence in Academic Libraries,” *International Journal of Computer Sciences and Engineering*, Vol. 7, No. 16, pp. 136-140, September 2019.

[4] S. Lee, “Library and Big Data Analysis,” *KLA (Korean Library Association) Journal*, Vol. 55, No. 8, pp. 14-25, August 2014.

[5] J. On and S. H. Park, “Big Data Analysis for Public Libraries Utilizing Big Data Platform: A Case Study of Daejeon Hanbat Library,” *Journal of the Korean Society for Information Management*, Vol. 37, No. 3, pp. 25-50, August 2020. http://doi.org/10.3743/KOSIM.2020.37.3.025

[6] R. Abraham, J. Schneider, and J. vom Brocke, “Data Governance: A Conceptual Framework, Structured Review, and Research Agenda,” *International Journal of Information Management*, Vol. 49, pp. 424-438, December 2019. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[7] H. Kim, “Analysis of Standard Vocabulary Use of the Open Government Data: The Case of the Public Data Portal of Korea,” *Quality & Quantity*, Vol. 53, No. 3, pp. 1611-1622, May 2019. https://doi.org/10.1007/s11135-018-0829-z

[8] J. Cho, “A Study about Library-Related Open Data through Public Data Portals,” *Journal of the Korean Biblia Society for Library and Information Science*, Vol. 29, No. 2, pp. 35-56, June 2018. https://doi.org/10.14699/kbiblia.2018.29.2.035

[9] H.-S. Kim and W.-J. Kim, “A Study on Library Data Open Status and Improvement Strategies,” in *Proceedings of the 23rd Conference of Korean Society for Information Management*, Seoul, pp. 77-80, August 2016.

[10] J. H. Park, “Designing Dataset Management and Service System for Digital Libraries Using DCAT,” *Journal of the Korean Society for Library and Information Science*, Vol. 53, No. 2, pp. 247-266, May 2019. https://doi.org/10.4275/KSLIS.2019.53.2.247

[11] A. Nikiforova and K. McBride, “Open Government Data Portal Usability: A User-Centred Usability Analysis of 41

Open Government Data Portals," *Telematics and Informatics*, Vol. 58, 101539, May 2021. http://doi.org/10.1016/j.tele.2020.101539

[12] D. Wang, C. Chen, and D. Richards, "A Prioritization-Based Analysis of Local Open Government Data Portals: A Case Study of Chinese Province-Level Governments," *Government Information Quarterly*, Vol. 35, No. 4, pp. 644-656, October 2018. https://doi.org/10.1016/j.giq.2018.10.006

[13] A. Vetrò, L. Canova, M. Torchiano, C. O. Minotas, R. Iemma, and F. Morando, "Open Data Quality Measurement Framework: Definition and Application to Open Government Data," *Government Information Quarterly*, Vol. 33, No. 2, pp. 325-337, April 2016. https://doi.org/10.1016/j.giq.2016.02.001

[14] C. Song and H. Kim, "Improvements of Public Data Policy through Data Portal Analysis of Local Governments," *Journal of Digital Contents Society*, Vol. 23, No. 4, pp. 697-705, April 2022. https://doi.org/10.9728/dcs.2022.23.4.697

[15] C. Song and H. Kim, "Considerations in Releasing Public Data: The Case of Local Governments in Korea," *Journal of Information Science*, July 2022. https://doi.org/10.1177/01655515221106636

[16] S. Yun and J. Hyeon, "An Analysis of Open Data Policy in Korea: Focused on National Core Data in Open Data Portal," *Korean Public Management Review*, Vol. 33, No. 1, pp. 219-247, March 2019. http://doi.org/10.24210/kapm.2019.33.1.010

[17] D. Kim, H. Kim, C. Song, J. Yang, and H. Kim, "Methods for Utilising Local Government's Public Data Released to the Public Data Portal," *Journal of Digital Contents Society*, Vol. 22, No. 3, pp. 445-452, March 2021. http://doi.org/10.9728/dcs.2021.22.3.445

[18] J.-H. Rho, "The Current State and Challenges of Linked Data in Library Cataloging," *Journal of Korean Library and Information Science Society*, Vol. 50, No. 3, pp. 71-95, September 2019. http://doi.org/10.16981/kliss.50.3.201909.71

[19] H. J. Yi, A Study on the Implementation of Multi-Source Search System for Linked Data Utilization, Ph.D. Dissertation, Chung-Ang University, Seoul, February 2015.

[20] H.-K. Moon and S.-K. Han, "A Study of Reference Model of Smart Library Based on Linked Open Data," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 20, No. 9, pp. 1666-1672, September 2016. https://doi.org/10.6109/JKIICE.2016.20.9.1666

[21] C. J. Godby, S. Wang, and J. K. Mixter, *Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description*, Cham, Switzerland: Springer, 2015.

[22] N. Freire, R. Voorburg, R. Cornelissen, S. de Valk, E. Meijers, and A. Isaac, "Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network," *Information*, Vol. 10, No. 8, 252, July 2019. https://doi.org/10.3390/info10080252

[23] R. P. Lourenço, "An Analysis of Open Government Portals: A Perspective of Transparency for Accountability," *Government Information Quarterly*, Vol. 32, No. 3, pp. 323-332, July 2015. https://doi.org/10.1016/j.giq.2015.05.006

[24] A. Quarati, M. De Martino, and S. Rosim, "Geospatial Open Data Usage and Metadata Quality," *International Journal of Geo-Information*, Vol. 10, No. 1, 30, January 2021. https://doi.org/10.3390/ijgi10010030

송채은(Chaeeun Song)

2021년 : 중앙대학교 대학원
(문헌정보학 석사)

2016년～2021년: 중앙대학교 문헌정보학과
2021년～2023년: 중앙대학교 문헌정보학과 정보학전공 석사
2023년～현 재: 중앙대학교 문헌정보학과 정보학전공 박사과정
※관심분야 : 지식그래프, 공공데이터, 데이터 포털 등

김학래(Haklae Kim)

2010년 : 아일랜드 국립대학교 (공학박사)

2004년～2009년: Digital Enterprise Research Institute, Ireland
2009년～2016년: 삼성전자
2017년～2019년: 한국과학기술정보연구원
2019년～현 재: 중앙대학교 문헌정보학과 교수
※관심분야 : 지식그래프, 인공지능, 데이터 사이언스 등