Journal of Korean Library and Information Science Society
Vol. 51, No. 3

연구데이터 관리를 위한 OAK 메타데이터 확장 방안 연구*

A Preliminary Study on Extending OAK Metadata for Research Data

이 미 화(Mihwa Lee)**
이 은 주(Eun-Ju Lee)***
노 지 현(Jee-Hyun Rho)****

〈 목 차 〉

요 약 : 본 연구는 국립중앙도서관의 오픈액세스 리포지토리인 OAK에서 연구데이터를 기술할 수 있도록 OAK 메타데이터에 확장 방안을 제안하는데 목적이 있다. 이를 위한 연구방법으로 문헌연구, 사례조사, 관계자와의 면담을 실시하였다. 연구데이터 기술을 위한 기존 OAK 메타데이터의 확장 방안을 다음과 같이 도출하였다. 첫째, 연구데이터를 위한 모델링으로 컬렉션 〉 아이템 〉 파일로 구성된 기존 구조를 그대로 유지하되 컬렉션은 해당 연구데이터를 묶을 수 있는 상위 그룹으로 두고, 아이템에는 연구데이터의 메타데이터와 파일을 묶어 제공하는 구조를 제안하였다. 둘째, 표준, 사례 기관의 메타데이터를 기존 OAK 메타데이터와 매핑하여 연구데이터의 기술을 위해 OAK에 추가할 필요가 있다고 판단되는 요소를 선별하여 OAK 확장 요소를 도출하였다. 셋째, 구조화된 데이터를 통해 검색이나 추후 통계 등에 활용할 수 있도록 통제어휘집과 구문에 대한 사항도 제시하였다. 본 연구는 연구데이터의 기술을 위해 OAK 메타데이터를 확장함으로써 국내에서 산출되는 연구데이터가 공식적으로 수집・저장・활용될 수 있는 기반을 제공함으로써 국가적으로 연구의 중복을 방지하고 연구 산출물을 공유 및 재활용할 수 있는 정보환경을 구축하는데 기여하였다.

주제어 : 기관 리포지토리, 메타데이터, 연구데이터, 오픈액세스

ABSTRACT : This study aims to propose an extended OAK metadata for research data that would be described in OAK, an open access repository of the National Library of Korea. As a research method, literature review, case studies, and interviews with related parties were conducted. The method of extending the existing OAK metadata for research data was derived as follows. First, in modeling for research data, the structure of the collection〉 item〉 file is maintained, the collection is placed as a higher group to which the research data can be grouped, and item was combined metadata and files or digital objects of various formats together. Second, by mapping the metadata standard and case organizations with the existing OAK metadata, elements judged to need to be extended to OAK for research data were selected and reflected in the existing OAK. Third, the controlled vocabulary and syntax are also proposed so that it can be used for search or later statistics through structured data. By expanding the OAK metadata to describe research data, research data produced in Korea can be officially stored and used, which is the basis for preventing duplication of research and sharing and recycling research results nationally.

KEYWORDS : Institutional Repository, Metadata, Research Data, Open Access

* 본 연구는 2020년 국립중앙도서관『OAK 확장형 리포지터리 운영을 위한 연구데이터의 메타데이터 지침 개발』의 일부를 바탕으로 수정・보완한 것임.
** 공주대학교 문헌정보교육과 부교수(leemh@kongju.ac.kr / ISNI 0000 0004 6431 3495) (제1저자)
*** 동의대학교 문헌정보학과 조교수(ejulee@deu.ac.kr / ISNI 0000 0004 6335 8325) (공동저자)
**** 부산대학교 문헌정보학과 교수(jhrho@pusan.ac.kr / ISNI 0000 0004 6484 8385) (교신저자)

• 논문접수: 2020년 8월 13일 • 최초심사: 2020년 8월 25일 • 게재확정: 2020년 9월 7일
• 한국도서관・정보학회지, 51(3), 27-51, 2020. http://dx.doi.org/10.16981/kliss.51.3.202009.27

Ⅰ. 서 론

국립중앙도서관에서 추진하고 있는 OAK(Open Access Korea)는 오픈 액세스 방식의 학술정보 유통체제 구축을 목적으로 하는 지식협력체이다. 2009년 시작된 OAK 사업은 대학, 공공기관, 연구소, 기업체 등에서 생산한 디지털 지식정보를 체계적으로 수집・관리・공유할 수 있도록 기관 리포지토리(institutional repository)를 개발・보급하고, 수집된 정보에 대한 통합검색과 기관별・정보유형별 특화된 검색서비스를 제공하기 위해 OAK 포털을 구축하는데 주력하고 있다. 그러나 OAK 참여기관이 확대되고 연구 수행 과정에서 생산된 다양한 부산물인 연구데이터를 효율적으로 관리하고 공유할 필요성이 증대함에 따라 OAK 리포지토리는 기능의 확장이 불가피한 상황에 놓여 있다. 지금의 OAK 리포지토리는 단행본, 학위논문, 학술논문, 보고서 등 최종 산출물을 위주로 한 것이어서 다양한 유형의 연구데이터를 수용하기에는 한계가 있기 때문이다.

연구데이터는 연구를 수행하는 전 과정에서 생산된 데이터를 의미한다. 이는 연구 성과의 재현이나 진실성 확보를 위한 증거적 가치를 지니며, 후속 연구나 교육에 재활용될 수 있어 활용성 측면에서도 상당히 가치 있는 지식정보로 간주되고 있다. 이러한 이유로 최근 과학기술 분야의 국가 R&D 사업에서는 연구 개시 전에 체계적인 연구데이터 관리계획을 수립하고, 연구종료 후에 이를 다른 연구자가 활용할 수 있도록 공개할 의무를 연구책임자에게 부과하고 있다. 뿐만 아니라 연구데이터의 관리와 장기간 보존을 제공하고 다양한 연구데이터를 공유할 수 있는 안정적인 정보인프라를 마련할 것을 연구수행기관 측에 요구하고 있다(국가과학기술연구회 2019). 이러한 환경을 고려할 때, 연구의 전 과정에서 생산되는 연구데이터의 등록・수집・관리・보존을 체계적으로 지원하고, 이를 공유할 수 있는 플랫폼으로 널리 활용되도록 OAK 리포지토리의 기능 개선을 서두를 필요가 있다.

본 연구는 연구데이터의 효율적인 관리를 위해 기존 OAK 메타데이터 요소의 확장(안)을 도출하는데 목적을 두고 있다. 즉, OAK 리포지토리에 연구데이터를 수용할 수 있도록 연구데이터 기술을 위한 다양한 요소를 확장하는 방안을 마련하는데 목적을 두고 있는 것이다. 이를 위해 연구는 다음과 같은 순서로 진행하였다. 첫째, 연구데이터의 유형과 연구데이터에 적용되는 데이터 모델을 검토하였다. 둘째, 연구데이터에 적용되는 표준 메타데이터를 조사하였다. 셋째, 국내외 리포지토리에서 사용하는 연구데이터의 메타데이터 사례(요소, 입력지침, 입력사례)를 수집하여 분석하였다. 넷째, 연구데이터 기술을 위한 기존 OAK 메타데이터의 확장 방안을 도출하고 이에 대한 실무자 및 OAK 리포지토리 개발자들과의 검토 과정을 거쳐 최종 안을 확정하였다. 연구에 필요한 데이터는 문헌조사, 사례분석, 면담을 통해 수집하였다.

Ⅱ. 연구데이터의 개념과 데이터 모델

1. 연구데이터의 개념

연구데이터는 일반적으로 “연구를 수행하는 과정에서 생산되거나 수집된 자료”로 통칭된다. “매체나 형식에 상관없이 모든 기록된 정보”(University of Pittsburgh)로 포괄되지만, “각종 실험, 관찰, 조사, 분석 등을 통하여 산출된 자료로 연구 성과의 재현에 필수적이고 객관적인 사실 데이터”(국가과학기술연구회 2019)와 같이 그 의미를 구체화하고 있는 경우도 있다. 또한 리포지토리나 연구기관에 따라 논문이나 저술의 초안, 연구노트, 예비분석자료, 과학논문 초안, 장래 연구계획, 동료평가 또는 동료와의 커뮤니케이션, 실물자료 등 ‘연구데이터의 범주에 해당하지 않는’ 사례를 적시한 경우도 있다. 이처럼 연구데이터는 연구의 전 과정을 통해 생산되거나 수집된 정량적 또는 정성적 데이터로서 관리나 활용의 대상이 되는 모든 데이터라고 정의할 수 있다. 연구데이터에 해당하는 구체적인 유형은 다음과 같다(NC State University libraries 2019).

- 문서, 스프레드시트
- 실험노트, 현장노트, 일기
- 질문지, 채록지, 조사지, 코드북
- 실험데이터
- 필름, 오디오테이프, 비디오테이프
- 사진, 이미지 파일
- 단백질이나 유전자 서열
- 테스트 반응
- 슬라이드, 견본
- 연구과정에서 생성한 디지털 객체
- 데이터베이스의 콘텐츠(비디오, 오디오, 텍스트, 이미지 등)
- 모델, 알고리즘
- 응용프로그램의 콘텐츠(입력데이터, 산출데이터, 로그파일)
- 연구방법론, 작업 흐름도
- 표준작업절차서 등

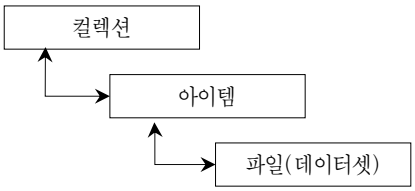
이러한 연구데이터는 데이터(파일) 자체 외에 다음과 같은 정보가 함께 수집되어 관리되어야 한다(Farnel and Shiri 2014): (1) 연구의 책임이 누구에게 있는지, (2) 데이터가 언제, 어디서,

왜 수집되었는지, (3) 데이터 수집을 위해 사용된 연구방법은 무엇인지, (4) 연구데이터는 어떻게 인용되어야 하는지, (5) 연구데이터에는 어떻게 접근하는지, (6) 이용의 제한사항은 무엇인지, (7) 데이터의 파일포맷은 무엇인지, (8) 데이터셋의 명칭은 무엇인지 등. 이들 정보는 메타데이터로 기술되는데, 요소에 따라 연구데이터 수집 프로세스에서 자동으로 생성되는 것도 있고 사람이 식별한 후 직접 입력해야 하는 것도 있다.

2. 연구데이터 기술을 위한 데이터 모델

연구데이터는 주로 해당 연구를 수행한 연구자가 등록하며, 하나의 연구로부터 생산되거나 수집된 다수의 연구데이터를 개별적으로 등록하기보다 연구와 관련된 여러 데이터를 하나의 데이터셋으로 묶어서 관리하는 방식을 취하고 있다. 이처럼 연구데이터를 어떠한 방식으로 관리할 것인지를 도식화한 것이 데이터 모델이다. 가령, 연구과제에 해당하는 '컬렉션'과 그로부터 생성된 연구데이터 파일 그룹인 '데이터셋', 그리고 개별 파일에 해당하는 '파일'과 같은 엔티티를 적용한 데이터 모델을 구성할 수 있다. 메타데이터는 이러한 데이터 모델에 근거하여 계층적으로 적용된다.

현재 연구데이터에 적용된 데이터 모델 사례를 살펴보면, '컬렉션 - 아이템 - 파일'과 같이 구조화하는 것이 일반적임을 알 수 있다(〈그림 1〉 참조). '아이템' 대신에 '오브젝트' 또는 '데이터셋'이라는 엔티티를 적용한 사례도 있지만 기본 구조는 동일하거나 유사하다. 또한 해당 아이템에 대한 파일이 복수일 때 이를 '데이터셋'으로 묶어 구성하기도 한다. 단위 기관에서 생성한 데이터를 모아서 통합검색을 제공하는 리포지토리에서는 '컬렉션'의 상위에 '리포지토리' 엔티티를 별도로 두어 4개의 엔티티를 적용한 사례도 있다.



- 컬렉션: 아이템을 그룹화하기 위한 논리적 그룹. 일반적으로 기관/부서, 연구과제명 등을 적용함
- 아이템: 연구데이터 파일의 그룹. 파일+메타데이터로 구성되거나 메타데이터로만 구성될 수도 있음
- 파일: 개별 단위의 연구데이터 파일이나 파일들의 집합(데이터셋)

〈그림 1〉 연구데이터 기술을 위한 데이터 모델 사례

3. 연구데이터의 주요 기술 요소 및 특징

연구데이터는 공식적으로 발표되는 최종 산출물에 대한 메타데이터와는 차별화되는 요소 및 기술방식을 가진다. 이에 연구데이터에 적용되는 기술 요소와 그 특징을 간략히 살펴보면 다음과 같다.

- 표제: 공식 출판된 자료와 달리 연구데이터의 경우 출판된 자료가 아니므로 표제를 고안해야 하는 어려움이 있다. 명확하게 데이터셋 전체를 포괄하는 표제를 만들지 않으면 데이터셋에 접근하기 어렵기 때문에 표제는 간략하면서 기술적이고 고유하게 작성되어야 하고 해당 분야에서 인지할 수 있는 언어로 작성해야 한다.
- 연구자: 연구데이터를 창작하는 데에는 다수의 개인이나 조직이 다양한 역할로 참여하고, 그 책임성의 수준도 다양하므로 책임성의 수준을 나타낼 수 있도록 역할어를 기술할 수 있어야 한다. 또한, 연구 참여자를 포괄적으로 기술할 것인지 주요 연구자만 기술할 것인지도 고려해 보아야 한다.
- 연구자의 연락처 정보: 연락처 정보는 다른 이용자가 그 연구데이터를 재사용할 수 있는지에 대한 추가 정보를 얻기 위해서는 매우 가치 있는 기술 요소이므로 가능한 기술한다. 연구자에 대한 연락처 정보로 전자메일이 가장 적합하고, 변경에 대비해 연구자마다 연락처 정보를 기술할 필요가 있다.
- 방법론: 연구데이터의 가장 큰 특징은 방법론이다. 데이터를 수집, 처리하는데 사용된 방법론은 연구 결과물의 내용을 파악하게 하므로 일반적으로 서술형식으로 기술되고 있다. 그러나 연구방법론이 검색이나 통계 처리를 위해서 사용될 수도 있으므로 통제어휘집 혹은 용어리스트를 사용해 기술하는 것도 고려하여야 한다. 연구방법론을 위한 통제어휘집의 용어는 다양한 관점을 적용하여 개발할 수 있다. 예를 들어, 특정 데이터 수집 방법을 중심으로 용어를 만들면 내용 분석, 실험, 관찰, 시뮬레이션, 서베이 등이며, 수집, 처리, 분석에 사용된 방법 측면에서는 소프트웨어 이름과 버전, 도구, 통계 테스트 등으로 용어를 개발할 수 있다.
- 데이터의 수록 범위: 데이터가 수집, 생성되고 처리되는 과정의 시간의 범위를 기술하는 것은 연구데이터에 매우 중요하다. 왜냐하면 관찰 데이터의 경우 관측일자, 탐사일자에 따라 수집된 데이터가 다를 수 있기 때문이다. 예를 들어, 관측을 중심으로 하는 천체데이터나 지질데이터는 계절에 따라 그 데이터 값이 다르므로 관측한 시간을 반드시 기술해야 한다.
- 라이선스: 라이선스 및 배포는 연구데이터의 재사용과 밀접하게 연관되는 사항으로 일반적으로 연구자가 직접 기술한다. 이 정보는 이용자에게 이 데이터를 공유할 수 있는 조건과 허가 없이 재사용할 수 있는 조건을 알려주고, 재사용을 위한 추가 조건이 있다면 연구자의 연락처 정보를 통해 사용가능 여부를 파악할 수 있도록 한다.
- 연구 지원 기관: 연구를 지원한 기관을 기술하며, 복수의 펀딩을 받은 연구에 대해서는 복수로 지원 기관을 입력할 수 있어야 한다.
- 키워드: 검색을 위해 학문이나 하위 학문 분야, 연구 토픽, 연구방법이나 도구, 데이터와 관련된 시간 범위, 탐사 장소 등 다양한 측면에서 키워드가 기술될 수 있도록 한다. 물론 연구방법은 연구방법 요소에, 시간범위는 데이터 수록범위 요소에도 기술될 수 있으나 검색을 위해

키워드에도 추가 기술할 수 있다.

- 언어: 언어는 데이터가 작성된 언어를 기술하는 것이지만 컴퓨터 프로그래밍을 사용하는 경우 C++, MATLAB, Python, XML 등과 같은 프로그래밍 언어를 기술할 수 있도록 한다.
- 관계: 관계는 연구데이터의 아이템을 포함하는 상위 컬렉션과 연계할 수도 있고, 컬렉션이 세분된 경우 컬렉션과 하위 컬렉션을 연계할 수도 있고, 연구데이터를 이용해 논문을 출판한 경우 이를 상호 연계하는 것도 고려되어야 한다.
- 인용: 연구데이터를 참고하거나 인용한 출판물은 참고 문헌에 연구데이터의 출처를 기술해야 한다. 연구데이터는 공식 출판 자료가 아니므로 인용 방식을 기술하는 것이 용이하지 않은데 일반적으로 다음과 같이 해당 리포지토리를 제시하고 URI, DOI 등 식별자를 기술한다(University of Michigan Deep Blue Home page).

〈연구데이터 인용 예시〉

Grundler, M., Grundler, M., Herrera, V. (2018). Video data of predation and parasitism by arthropods on small vertebrates in lowland Peruvian Amazon [Data set]. University of Michigan - Deep Blue. https://doi.org/10.7302/Z2862DP1

Ⅲ. 연구데이터 표준 메타데이터 및 구축 사례 분석

1. 연구데이터를 위한 표준 메타데이터

re3data(Registry of Research Data Repositories)에서는 전 세계의 연구데이터 리포지토리에 저장된 연구데이터를 손쉽게 검색할 수 있는 환경을 제공할 뿐만 아니라 연구데이터 관련 표준과 그 표준을 사용하는 기관의 현황을 제시하고 있다. re3data를 참고할 때, 현재 연구데이터에 가장 많이 사용되는 메타데이터는 Dublin Core (298개 기관) 〉 Data Documentation Initiative[1](169개 기관) 〉 DataCite Metadata Schema (165개 기관) 순이다. 이 중 특정 학문과 관련 없는 공통적인 메타데이터는 Dublin Core와 DataCite Metadata Schema으로, 본 연구에서는 두 표준을 중심적으로 살펴보았다.[2] Dublin Core는 다양한 정보자원에 폭넓게 사용하기 위해 개발된 표준으로써 DCMES 15개의 기본요소 외에 '컬렉션' 단위를 기술하기 위한 Dublin Core Collections Application

1) Data Documentation Initiative(DDI): 인문, 사회과학 연구데이터 기술을 위해 개발된 메타데이터 표준

2) 국내에는 2017년 제정된 "연구데이터 관리 및 공유를 위한 메타데이터 표준"(TTAK.KO-10.0976)이 있으나 이 표준을 분석한 결과 DataCite를 바탕으로 개발되었으므로 DataCite Metadata Schema의 메타데이터 요소와 중복적이라 판단하여 본 연구에서는 제외하였다.

Profiles(DC Collections AP)가 있다(DCMI 2007; 2012). 이를 통해 개별 자원에 대한 기술뿐 아니라 정보자원의 컬렉션과 컬렉션에 포함된 아이템의 관계, 혹은 컬렉션과 컬렉션 사이의 관계, 나아가 컬렉션과 다른 자원과의 관계 등을 기술할 수 있다.

그리고 DataCite Metadata Schema는 연구데이터를 보다 쉽게 접근·활용할 수 있도록 2009년에 설립된 국제 컨소시엄인 DataCite에서 개발한 메타데이터 표준이다. DataCite는 연구데이터를 정확하고 지속적으로 관리하기 위해 연구데이터의 영구적 식별자(DOI)를 제공하고 있으며, 객체에 할당된 각 DOI에 대한 메타데이터를 수집하여 관리함으로써 연구데이터를 체계적으로 관리할 수 있다. 또한 데이터 입력 시 각종 통제어휘집을 활용하거나 표준 어휘집을 적극적으로 사용함으로써 다양한 스키마들과 상호운용성을 확보할 수 있도록 노력하고 있다.

2020년 8월 현재 DataCite Metadata Schema는 4.3버전이 발행되어 활용되고 있으며, 4.3버전은 19개의 최상위요소와 66개의 하위요소를 가지고 있다. DataCite의 19개 최상위요소는 필수요소, 권장요소, 선택요소로 구분되며, 해당 요소는 다음 〈표 1〉과 같다. DataCite는 연구데이터의 관리를 위해 특별히 개발된 메타데이터이기 때문에 데이터 생산·수집과 관련된 지리적 공간 또는 장소와 관련된 'GeoLocation' 요소, 연구데이터를 생산하는 과정에서의 재정지원 기관에 대한 'FundingReference' 요소, 각종 사업과 관련된 사항을 입력하는 'award' 관련 요소 등을 갖는 것이 특징적이다.

〈표 1〉 DataCite Metadata Schema 4.3의 메타데이터 요소

필수요소	권장요소	선택요소
자원식별자(Identifier) 생산자(Creator) 제목(Title) 발행자(Publisher) 발행년(PublicationYear) 자원유형(ResourceType)	주제(Subject) 기여자(Contributor) 날짜(Date) 관련식별자(RelatedIdentifier) 기술(Description) 지리적 공간(GeoLocation)	언어(Language) 대체식별자(AlternateIdentifier) 크기(Size) 포맷(Format) 버전(Version) 권한(Rights) 재정지원 정보(FundingReference)

출처: DataCite Metadata Working Group(2019)

2. 국내외 리포지토리의 연구데이터 구축 사례

본 연구에서는 연구데이터 구축 사례로 한국연구재단의 기초학문자료센터(KRM), KISTI의 국가연구데이터플랫폼, UC San Diego Library Digital Collections, University of Michigan Deep Blue Data Repository를 선정하여 분석하였다. 국내 2곳은 대표적인 국내 연구데이터 수집 및 구축 기관이고, 국외의 경우 앞서 살펴본 DC와 DataCite를 적용한 대표적인 사례 기관이다.

구체적으로, UC San Diego의 경우 DataCite를 사용한 곳이고, University of Michigan의 경우는 DC를 중심으로 구축한 사례이다. 사례 분석은 개요, 데이터모델, 메타데이터, 통제어휘로 나누어 분석하였으며, 통제어휘집은 제시하지 않은 곳이 있어 제공되는 경우만 포함하였다.

가. 국내 사례

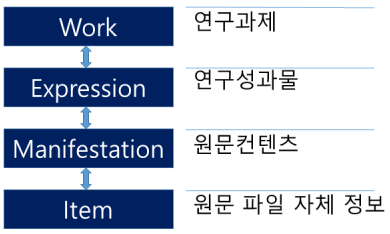
(1) 한국연구재단의 기초학문자료센터

(가) 개요

2007년에 개관한 기초학문자료센터(Korean Research Memory; 이하 KRM)는 학술연구지원사업의 연구성과물을 관리·공유·확산함으로써 중복연구 발생을 방지하고 기 수행 연구 자료를 동료 및 후속 연구자가 쉽게 열람할 수 있도록 하여 인문사회분야 발전의 기반을 구축하는 것을 목표로 삼고 있다. KRM에서는 인문사회분야 학술연구지원과제의 수행과정에서 생산된 연구성과물의 범위를, 단행본, 보고서, 논문, 조사자료(통계자료), 고문서, 고도서, 이미지, 동영상, 녹음자료, 웹사이트, 낱장자료, 원문콘텐츠로 구분하고, 자료의 성격(원자료, 중간산출물, 연구결과물)에 따라 수집한 후 메타데이터를 입력하고 있다.

(나) 데이터 모델

KRM에서는 연구데이터 기술을 위해 Functional Requirements for Bibliographic Records (FRBR) 모델을 활용하여 계층적, 연관적 메타데이터 구조를 응용하여 설계하였으며, 데이터베이스에 수록된 자료들은 다음 〈그림 2〉의 4단계로 계층화되어 있다. 이 데이터 모델을 하나의 연구과제에 여러 개의 연구데이터(가령, 학술논문, 사진, 인터뷰녹음자료)가 생산되는 상황에 적용해 본다면 연구과제에 대한 하나의 저작(Work)을 생성하고 논문, 이미지, 녹음자료 유형에 대한 메타데이터 요소를 적용한 표현형(Expression)을 생성한 뒤, 각 파일에 대한 정보(가령, 파일크기와 공개여부 등)를 입력한 구현형(Manifestation)을 생성·연결해 줄 수 있다.



〈그림 2〉 KRM 데이터 모델

(다) 메타데이터

연구과제(Work)에 대한 메타데이터 요소는 ① 과제명, ② 영문과제명, ③ 사업명, ④ 연구과제번호, ⑤ 선정년도, ⑥ 연구기간, ⑦ 연구책임자, ⑧ 연구수행기관, ⑨ 과제진행현황, ⑩ 과제신청시 연구개요, ⑪ 연구과제의 신청시 심사신청분야, 총 11개로 이루어져 있으며, 연구성과물(Expression)에 대한 메타데이터 요소는 자료유형에 따라 차이가 있지만, 공통적으로 ① 연구과

제번호, ② 제목, ③ 저자(제작자, 작성자), ④ 연구자 소속기관, ⑤ 발행일, ⑥ 발행처, ⑦ 키워드(색인어), ⑧ 초록, ⑨ 언어 요소가 적용되고 있다(한국연구재단 기초학문자료센터 홈페이지).

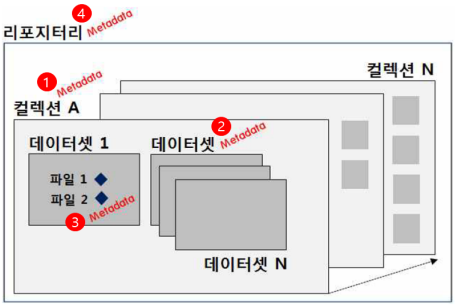
(2) 한국과학기술정보연구원(KISTI)의 국가연구데이터플랫폼(KORD)

(가) 개요

2019년 12월 한국과학기술정보연구원(KISTI)에서는 국내외 연구데이터 정보를 한 곳에서 서비스하는 것을 목표로 국가연구데이터플랫폼서비스(Korea Research Data Platform Service; 이하 KORD)를 개시하였다. 해당 서비스는 국가 R&D에서 나온 여러 데이터를 통합·연계하는 온라인 플랫폼으로, 연구데이터를 검색가능하고 접근가능하며 상호운영가능하고 재사용이 가능하도록 하는 FAIR(Findable, Accessible, Interoperable, Reusable) 원칙을 준용하고 있다(국가과학기술연구회 2019, 166). KORD에서는 정부출연 연구기관, 대학 등 각 기관에서 생산한 연구데이터를 등록, 관리, 검색, 다운로드할 수 있도록 지원하고 있으며, 관찰, 관측, 실험, 조사, 측량, 분석 등을 통하여 산출된 자료를 수집대상으로 삼고 있고, 데이터 형식이 미리 연구자 그룹에 의해 정의되어 있기도 하고, 특별한 형식 없이 스프레드시트, 이미지, 영상 등의 다양한 형태와 종류의 데이터를 포함할 수 있도록 제공하고 있다(한국과학기술정보연구원 2019, 3).

(나) 데이터 모델

KORD에서는 연구데이터 관리뿐 아니라 리포지토리 관리를 위해 '컬렉션 〉 데이터셋 〉 파일 〉 리포지토리' 구조를 따르는 데이터 모델을 적용하였다(〈그림 3〉 참조). 구체적으로, 리포지토리는 연구데이터들을 보관, 저장, 서비스하는 저장소(시스템)의 지칭하며, 컬렉션은 데이터셋을 그룹화하기 위한 논리적 그룹으로, 프로젝트, 부서, 연구과제 등과 같이 구축하고자 하는 의도에 맞게 자유롭게 설정할 수 있다. 그리고 데이터셋은 연구데이터의 관리 및 공유, 재사용성을 높이기 위해 파일을 묶어 놓은 단위로, 과학탐사 과정에서 발생한 관측데이터와 관련 연구자들과 면담데이터는 연구방법에 따라 관측 데이터셋과 면담 데이터셋으로 그룹핑할 수 있다. 마지막으로, 파일은 관리 및 공유, 재사용의 가치가 있는 개별 단위의 연구데이터를 지칭한다.



〈그림 3〉 KORD 데이터 모델
출처: 한국과학기술정보연구원 2019, 8

(다) 메타데이터

KORD는 데이터 모델에서 나타난 리포지토리, 컬렉션, 데이터셋, 파일 수준에 따라 기술 요소에

차이가 존재한다. 각 단계별 요소는 필수요소, 권고요소, 선택요소, 해당시 필수 요소로 구분되어 있는데, 데이터 모델에 따른 필수요소를 중심으로 살펴보면 컬렉션, 데이터, 파일은 공통적인 요소가 많이 존재하는 것을 확인할 수 있다(〈표 2〉 참조). 그러나 리포지토리 메타데이터 요소는 리포지토리 접근(DatabaseAccessType), 연구데이터에 적용되는 접근 유형(DataAccessType), 라이선스(DataLicenseName, DataLicenseURI), 데이터 제출유형(DataUpload), 데이터와 출판물의 상호참조(EnhancedPublication), 데이터 및 메타데이터 품질관리 여부(QualityManagement) 등과 같은 독특한 요소들이 존재함을 알 수 있다.

〈표 2〉 KORD 메타데이터 중 필수 요소

리포지토리	컬렉션	데이터셋	파일
RepositoryURI	Identifier	Identifier	Identifier
Identifier	IdentifierType	IdentifierType	IdentifierType
IdentifierType	Title	Title	Title
RepositoryName	DateType	DateType	DateType
Type	SubjectName	SubjectName	subjectName
RepositoryLanguage		Creator	Creator
Subject		creatorName	
SubjectName		nameType	
InstitutionName		Publisher	Publisher
InstitutionCountry		PublicationYear	PublicationYear
DatabaseAccessType		Language	
DataAccessType		contributorType	contributorType
DataLicenseName			
DataLicenseURI			
DataUpload			
Versioning			
EnhancedPublication			
QualityManagement			

(라) 통제어휘

KORD에서는 re3data.org의 통제어휘리스트(dataAccessType, subjectScheme, databaseAccessType, dataAccessRestriction, dataUploadType, enhancedPublication, qualityManagement, responsibilityType, versioning 등)와 DCMI(FileType)를 활용하지만, 가장 적극적으로 활용하는 통제어휘는 DataCite의 통제어휘이다. DataCite에서 제공하는 contributorType, dateType, ResourceTypeGeneral, relationType, relatedIdentifierType 등의 세부적인 값을 활용하고 있다(한국과학기술정보연구원 2019, 63-67).

나. 국외 사례

(1) UC San Diego Library Digital Collections

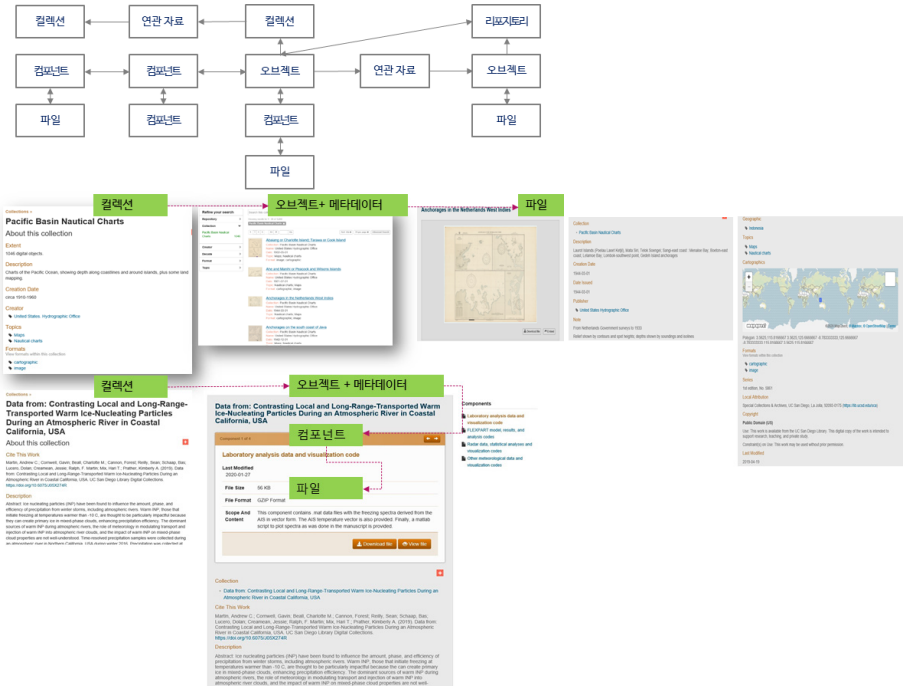
(가) 개요

UC San Diego Library의 리포지토리 Digital Collections(https://library.ucsd.edu/dc/)에서는 대학에서 구축한 기존 리포지토리와 과학데이터 연구소와 통합하여 검색 서비스를 제공한다. 즉 예술, 영화, 음악, 역사 및 인류학 분야 컬렉션의 디지털 자료를 대상으로 하는 도서관 컬렉션(Library Digital Collections)과 UC San Diego 연구자에 의해 생성된 연구데이터 컬렉션(UCSD Research Data Collections)의 2가지로 구성된다. 기존 데이터는 DC 기반이며, 연구데이터는 DataCite 스키마를 바탕으로 개발되었다(Pennington 2020).

연구데이터에는 설정데이터(Configuration data), 소스 코드, 가공되지 않은 데이터, 과학데이터, 통계데이터, 데이터베이스, 아카이브데이터, A/V 데이터, 표준 문서, 이미지 등이 포함된다.

(나) 데이터 모델

디지털 객체는 단순한 파일이 아니고 계층에 따른 위계와 데이터요소가 달리 제공된다. '커뮤니티 혹은 컬렉션 〉 오브젝트 〉 컴포넌트 〉 파일'의 구조를 따라 계층화된다. 〈그림 4〉에서 첫 번째



〈그림 4〉 UCSD 데이터모델 및 예시
출처: University of California in San Diego Library Digital Collections Home page

예시와 같이 컬렉션 내에 여러 오브젝트가 있고, 오브젝트는 메타데이터와 파일로 구성되어 있다. 만일 오브젝트 내에 여러 파일이 구성된 경우는 〈그림 4〉의 두 번째 예시와 같이 오브젝트 다음에 컴포넌트가 추가되고 컴포넌트 아래에 다수의 파일이 묶이게 된다.

따라서 컬렉션에는 1개 이상의 오브젝트가 포함되어 있으며, 이러한 오브젝트는 컬렉션 내의 아이템이라고 할 수 있다. 오브젝트는 메타데이터와 컴포넌트(component) 혹은 파일로 구성되고, 오브젝트는 1개 이상의 이미지, 비디오, 텍스트, 데이터 등의 파일을 가질 수 있다. 오브젝트에 파일이 하나인 경우 컴포넌트 없이 파일 하나만 제공된다. 예를 들어, 과학 탐사의 경우 탐사를 통한 이미지, 탐사 관련 관측데이터 등이 발생하며, 이를 각각의 파일로 저장하고, 이러한 여러 파일을 하나로 묶을 수 있는 컴포넌트를 구성하게 된다. 오브젝트는 파일과 메타데이터 혹은 컴포넌트와 메타데이터로 구성되며, 컴포넌트 하에는 여러 파일이 묶여있다.

(다) 메타데이터

DataCite Metadata Schema + Dublin Core + 리포지토리 자체 개발 메타데이터 스키마를 사용하며, 입력형식은 RDF data model로 RDF/XML을 준용하여 궁극적으로 링크드데이터를 목적으로 한다. 특히, 관계정보는 오브젝트가 속하는 컬레션, 오브젝트에 포함되는 파일 정보가 상호 관계로 표시되고, 다른 컬렉션과도 연계될 수 있다.

컬렉션, 오브젝트, 파일 수준에 따라 기술되는 메타데이터에 차이가 있다. 컬렉션에서는 표제, 크기, 주기, 일자, 창작자, 토픽, 포맷이 해당된다. 오브젝트에서는 컬렉션 수준의 정보를 계승하면서 표제, 일자, 출판사, 일반주기, 주제 등 상세한 정보가 제공된다. 파일 수준의 정보에는 파일과 관련된 테크니컬한 정보가 포함되며 대부분 자동으로 생성된다.

오브젝트 수준에서 기술되는 메타데이터 요소는 크게 표제, 에이전트, 날짜, 주기, 식별자, 주제, 언어, 관련 자원, 기타 저작권 및 접근관련 요소로 나눌 수 있다. 이 중 특징적인 요소만 살펴보면, 에이전트라는 행위자를 두되, 창작자, 기여자, 출판사는 세구분하고, 나머지는 행위자는 역할어(role)를 이용해 에이전트의 역할을 식별한다. 날짜 중에서는 date.event를 두어 탐사일이나 특정 이벤트 일자를 기술하도록 하였다. 주기는 세분화되어 있는데 일반주기, 목차, 펀드 등의 도서관 분야의 주기 이외에도 연구데이터 관련 주기로 탐사 수심, 발견사항, 제한사항, 연구방법, 기술적인 세부사항을 두어 연구데이터 기술시 사용한다. 식별자 정보는 다양하게 사용되며 고고학에서 사용되는 edm 식별자도 포함된다. 주제와 관련된 사항으로 해부학(anatomy) 관련 주제를 별도로 세부 기술하며, 수집된 암석 층위학 명칭(lithology)을 기술할 수 있고, 과학적 학명(scientific name)을 기술하거나 탐사선(cruise)도 기술한다. 접근 및 저작권 관련하여 저작권 상태, 저작권 관련 기술, 저작권 관련 주기, 저작권 소유자, 라이선스, 시스템 오버라이드(rights override), 엠바고 만료일, 엠바고 기간 중에 객체의 접근 상태, 엠바고 만료 후 객체의 접근 상태, 저작권 관할

지역 등이 있다.

파일 수준의 기술 요소는 파일 수정 일자, 파일 생성일자, 파일 레이블, 파일크기, 파일명, 파일 포맷명, 파일 버전정보, 객체 유형(비트스트림, 파일, 오브젝트인지 구분), 파일의 보존 수준, 아날로그 소스 자원의 매체, 파일 생산 기관, 파일을 생성시 사용된 장치, 파일의 상영 시간, 파일 품질, 원본 소스 파일의 위치이다.

(라) 통제어휘

에이전트 유형으로 개인, 조직, 그룹, 가족을 구분하고, 에이전트의 역할어로 artist, art director, applicant, architect 등 250가지 통제어휘를 개발하였다. 자원유형은 RDA 자원유형을 참조하여 cartographic, mixed material, moving image, notated movement, notated music, sound recording, sound recording-musical, sound recording-nonmusical, still image, text, three dimensional object, data, multimedia, software로 구성된다. 이외에도 라이선스, 저작권 상태, 엠바고 관련 사항, 접근방식 등에 대한 세부적인 값을 제시하였다.

(2) University of Michigan Deep Blue Data repository

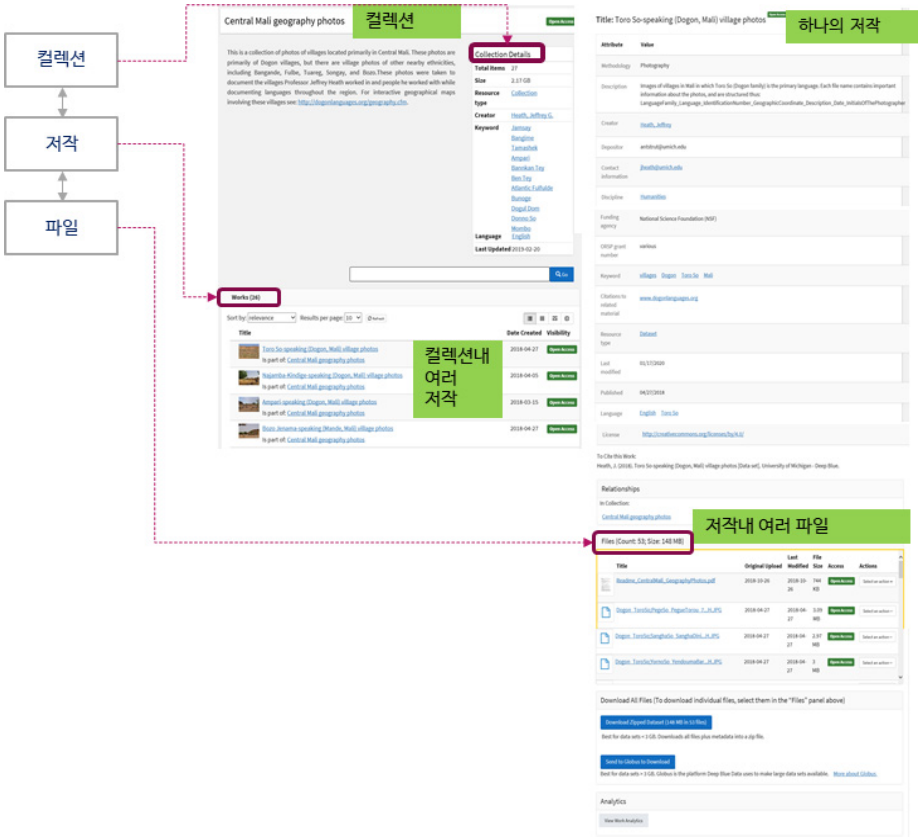
(가) 개요

University of Michigan의 기관 리포지토리인 Deep Blue는 2개의 리포지토리를 운영하는데, 하나는 DSpace를 바탕으로 하는 전통적인 기관 리포지토리인 Deep Blue로 이는 논문, 책의 장, 비디오, 학위논문 등 기관에서 나온 출판 결과물에 대한 접근을 제공하고, 다른 하나는 연구활동의 과정에서 개발되거나 사용되는 연구데이터에 대한 리포지토리인 Deep Blue Data이다. 연구데이터를 이용하려면 메인화면(https://deepblue.lib.umich.edu)에서 연구데이터 웹사이트(https://deepblue.lib.umich.edu/data)를 선택해 이용한다. 연구데이터 리포지토리는 2016년 2월 29일 시작하였으며 정식 오픈은 2016년 9월 20일 시작하였다. SAMBERA 프로그램을 사용하는데 이는 DC 기반이다.

연구데이터의 자료유형은 데이터셋과 컬렉션으로만 구분되어 있고, 수록된 연구데이터는 재사용이 가능하며 관찰, 서베이 결과, 시뮬레이션, 실험 데이터와 같이 컴퓨터를 통해 열람하거나 사용되는 연구데이터를 대상으로 하고 있다(University of Michigan Deep Blue Home page).

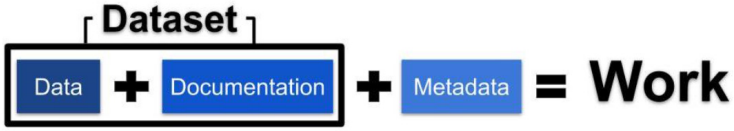
(나) 데이터 모델

데이터 모델은 '커뮤니티 혹은 컬렉션 〉 저작 〉 데이터셋(파일 포함)'으로 구성된다(〈그림 5〉 참조). 컬렉션은 관련 저작의 네비게이션이 용이하도록 그룹핑한 것으로 이러한 그룹핑은 연구자나 연구데이터 서비스 담당자에 의해서 이루어진다. 동일 컬렉션 내에서는 관련 저작을 함께 모아 보여준다.



〈그림 5〉 University of Michigan Deep Blue Data 리포지토리
출처: University of Michigan Deep Blue Home page

저작은 리포지토리 내에 핵심 정보 조직 단위로 데이터셋을 디포짓할 경우 저작을 우선 작성하고, 저작에는 모든 데이터와 데이터셋 내의 파일을 포함한다. 저작은 Deep Blue 데이터 리포지토리에서 주된 조직 단위로 하나의 웹페이지로 표현된다. 데이터, 다큐멘테이션, 메타데이터를 디포짓하면 이러한 부분이 결합해 하나의 저작을 형성하게 되는데 저작은 대중적으로 이용가능한 데이터셋으로 제시된다(〈그림 6〉 참조).



〈그림 6〉 저작, 데이터셋, 메타데이터와 관계
출처: University of Michigan Deep Blue Home page

데이터셋은 연구데이터를 수록한 모든 파일이나 다큐먼트와 파일의 내용을 접근하거나 파일의 내용을 이해할 수 있도록 하는 다큐멘테이션을 추가 구성요소로 포함한다. 다큐멘테이션은 데이터셋을 이해하고, 신뢰하고, 이용하는데 필요한 정보이다. 이는 데이터가 수집, 처리, 분석된 사항에 대해 좀 더 상세한 정보를 제공하므로 데이터셋 메타데이터의 정보를 확장한다. 다큐멘테이션은 여러 형태로 나타나는데 일부는 README 문서(텍스트 파일이나 PDF 형태) 혹은 코드북(codebooks)과 같이 별도 파일이거나 데이터파일 자체에 포함되어 있기도 하다. 연구데이터는 연구 과정 중에 생겨난 데이터로 하나의 파일일 수도 있지만, 여러 파일로 구성되기도 하는데 여러 파일로 구성된 경우 이를 이용하기 위한 보조로 다큐멘테이션을 사용한다.

메타데이터는 데이터셋에 관한 핵심 속성을 정의한 기술 정보이다. 데이터셋 메타데이터를 통해 이용자는 쉽게 연구데이터를 발견하고 자신의 연구과의 관련성 정도도 파악할 수 있다. 메타데이터는 온라인 검색 엔진을 통해 발견되도록 고안되어야 하며 메타데이터를 준비하는 연구자는 다른 이용자가 찾을 수 있도록 상세하게 기술하고, 데이터셋과의 연계성을 고려해 포괄적으로 기술해야 한다.

(다) 메타데이터

DC를 기반으로 하되 필요한 요소는 자체 개발하였다. 자체 개발한 요소는 RDF 구축을 위해 속성을 predicate:::RDF::Vocab::DC.alternative와 같이 표시한다. 자체 개발된 요소에는 〈표 3〉과 같이 파일명, 파일 위치, 레포지토리 접근, 지원금 제공기관, 지원금번호, 연구방법, 출판일자가 있으며, 이 중 연구방법론(methodology)은 데이터 수집과 처리에 사용된 방법론으로 내용분석, 실험, 관찰, 시뮬레이션, 서베이 등이 포함되며, 데이터 수집, 처리, 분석에 사용된 도구로 소프트웨어명과 버전, 도구, 통계 테스트 등도 기술할 수 있다(Carlson 2020).

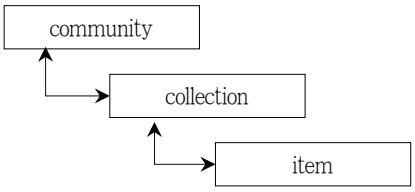
〈표 3〉 UM 자체개발 메타데이터 요소

속성	설명	공개한 자체 개발 요소
label	파일명	predicate:ActiveFedora::RDF::Fcrepo::Model.downloadFilename
relative_path	파일 위치	predicate:::RDF::URI.new('http://scholarsphere.psu.edu/ns#relativePath')
access_deepblue	레포지토리 접근	predicate:::RDF::URI.new('https://deepblue.lib.umich.edu/data/help.help#access_deepblue')
fundedby_other	지원	predicate:::RDF::URI.new('https://deepblue.lib.umich.edu/data/help.help#fundedby_other')
grantnumber	지원금 번호	predicate:::RDF::URI.new('http://purl.org/cerif/frapo/hasGrantNumber')
methodology	연구방법론	predicate:::RDF::URI.new('http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/schemas/datacollection_xsd/elements/DataCollectionMethodology.html')

Ⅳ. 연구데이터 관리를 위한 OAK 메타데이터 확장(안)

1. OAK 메타데이터 확장을 위한 설계 방향

OAK 리포지토리에 연구데이터를 수용하기 위해서는 현행 OAK의 데이터 구조 및 메타데이터 요소에 대한 분석이 선행될 필요가 있다. 현재 OAK에 적용된 데이터 모델은 community - collection - item 등 3개의 엔티티로 구성되어 있다(〈그림 7〉 참조). community 엔티티에는 일반적으로 데이터를 구축하는 기관에 대한 정보가, collection 엔티티에는 기관에서 구축한 디지털 컬렉션에 관한 정보가 수록된다. collection 엔티티는 데이터를 구축하는 기관에서 자체적으로 정의할 수 있지만, 주로 자원유형(학위논문, 보고서, 단행본, 프리젠테이션 등)으로 구분하여 적용하고 있다. 마지막으로 item 엔티티에는 해당 자원에 대한 메타데이터와 파일이 함께 수록된다. 가령, A연구소에서 구축한 '학위논문', '보고서' 등은 collection으로 구분되고, 개별 학위논문에 대한 메타데이터와 파일은 학위논문 collection 내의 item에 저장되는 구조이다.



〈그림 7〉 OAK 데이터 구조

현행 OAK 메타데이터는 DSpace 요소를 기본으로 하여 OAK에서 자체적으로 추가한 요소로 구성되어 있다. 16개의 상위요소와 67개의 요소구분(한정어)로 구성되어 있으며, community, collection, item 엔티티에 사용가능한 요소가 별도로 지정되어 있다. 예를 들어, title 요소는 community, collection, item 엔티티에 모두 적용 가능하며, subject 요소는 item 엔티티에만 적용할 수 있다. 또한 각 요소에 적용하는 인코딩 스킴으로 subject type, contributor type, resource type, identifier type 등이, 통제어휘 구문으로 날짜, 파일포맷, 언어, 표준식별자 등이 정의되어 있다.

이러한 OAK 구조와 메타데이터를 참조할 때, 연구데이터를 수용하기 위한 방안으로 현행 데이터 구조를 유지하면서 연구데이터를 기술할 수 있도록 메타데이터 요소나 구문을 확장 전개하는 것이 합리적일 것으로 판단된다. 현행 데이터 구조를 변경하지 않고서도 collection에 연구데이터를, item에 이에 대한 메타데이터와 관련 파일을 수용하는 구조의 적용이 가능하며, 무엇보다 기 구축된 데이터와의 일관성을 유지하고 입력자나 관리자의 혼란과 번거로움을 최소화하는데 최선의 방안이라 판단되기 때문이다.[3] 또한 특정 연구로부터 생성된 연구데이터 전체에 대한 정보를 하나의 메타데이터로 표현하되, 관련된 파일이 복수일 경우 미시건대학의 Deep Blue 사례와

3) 앞서 연구데이터를 구축하는 리포지토리 사례를 분석한 결과에서도, 한 연구로부터 생성된 연구데이터는 하나의 메타데이터에 다수의 파일을 포함하는 구조를 지향하면서 연구데이터 등록 시 입력 중복이나 번거로움을 덜어주고 있었다.

같이 데이터셋으로 묶어 메타데이터와 함께 item 엔티티에 수용할 수 있다는 점도 현행 데이터 구조의 유지가 가능한 이유라 할 수 있다.

이에 기존 구조 속에서 연구데이터를 위한 메타데이터 요소를 확장 전개하는 방안을 모색하였다. 먼저, 연구팀에서는 연구데이터 표준의 준수 및 관련 사례와의 상호호환성 확보가 중요하다고 생각하여, 관련 표준과 사례 분석에서 도출된 요소를 기존 OAK 메타데이터와 매핑하는 과정을 거쳤다. 이 과정에서 이미 OAK에 존재하거나 불필요하다고 판단되는 요소는 제외하고, 연구데이터의 기술을 위해 OAK에 추가할 필요가 있다고 판단되는 요소만을 우선적으로 선별하였다. 다음 〈표 4〉는 요소 매핑 및 선별 과정의 예시이다. 〈표 4〉와 같이 date 요소 및 요소구분(한정어)을 동일 항목별로 매핑한 다음, 현행 OAK 메타데이터와 비교하였다. 그 결과, 현행 OAK에 관련 요소들이 이미 존재하고 있으며, OAK에 없는 DataCite의 dateInformation(날짜와 관련된 세부 정보)과 UC San Diego의 date:collection(연구자가 연구데이터를 수집한 날짜)은 기존 요소로도 충분히 대체 가능하므로 확장할 필요가 없는 요소로, 그리고 연구데이터가 생성된 날짜를 기술하는 date.event는 별도로 추가할 필요가 있는 요소로 판명되었다. 이러한 선별은 연구팀의 브레인스토밍과 연구데이터의 사례 검토를 통해 진행되었다. 통제어휘집과 구문도 동일한 방식을 적용한 후 수정 또는 확장 방안을 도출하였다. 이렇게 일차적으로 도출된 안에 대해 다시 사례를 대입해 보고, 최종적으로 OAK 실무진 및 시스템 개발자들과의 세밀한 검토를 거친 후 최종 안을 확정하였다.[4)]

〈표 4〉 OAK 메타데이터 추가 및 미적용 요소 선별 예시

관련 표준		관련 사례					OAK
DataCite	DC	KISTI	KRM	UC San Diego	Cornell University	University of Michigan	
date	**date**	**date**	**발행일/촬영/제작일**	**date**		**date_created**	**date**
dateType		DateType		date:created			date.created
dateInformation							미적용
				date:event			[추가] date.event
				date:collection			미적용
				date:copyrighted			date.dateCopyright
							date.valid
							date.available
							date.accepted
							…

4) 연구팀에서 최종 도출한 안에 대한 OAK 실무진과 시스템 개발자의 의견은 크게 다음 세 가지로 구분되었다. 첫째, 연구데이터는 동일한 데이터라 하더라도 최초 원시데이터에서부터 이를 가공한 또는 최종 인증된 데이터까지 정제 정도에 따라 다양한 버전의 데이터가 존재하므로 이를 반영할 필요가 있다. 둘째, 연구데이터의 특성상 연구데이터의 유형(예: 실험데이터, 조사데이터 등)이 중요한 검색요소가 될 수 있으므로 이에 대한 적용할 필요

2. 연구데이터 기술을 위한 OAK 메타데이터 확장(안)

〈표 5〉는 연구데이터의 기술을 위해 OAK 메타데이터를 확장한 결과이다(음영 처리된 부분이 확장 요소). 기존 OAK 메타데이터 요소와 비교하여 16개 상위요소는 변함이 없지만, 요소구분은 129개로 대폭 확장되었다. 요소구분이 확장된 상위요소는 subject, description, contributor, date, type, format, relation, coverage, rights, citation 등 총 10개 요소이며, 그 외 〈funderIdentifierType〉, 〈dataVersionType〉, 〈subResourceTypes〉, 〈rightOverrideType〉 등과 같은 통제어휘가 새로 추가되었다. 이외에도 각 엔티티에 적용할 요소에 대한 수정이 이루어졌다.

특징적인 요소만 살펴보면, 연구지원이나 후원과 관련하여 '연구지원기관'(funderName), '연구과제명'(awardTitle), '연구과제번호'(awardNumber) 등이 추가되었으며, 연구데이터의 정제 정도를 표현하기 위해 'dataVersion' 요소구분을 추가하고 해당 값의 표현을 위해 '미가공'(raw), '정제'(refinded), '신뢰'(trusted) 등과 같은 통제어휘를 적용하였다. 또한, 연구방법론을 기술할 수 있는 'methods'와 연구데이터의 유형 구분을 위해 'subType'을 신설함으로써 연구를 수행하는데 사용된 방법론과 그로부터 획득 또는 생성된 연구데이터의 유형(관찰 데이터, 실험 데이터, 추출/가공 데이터, 테스트 시뮬레이션 데이터, 재활용 데이터)을 구체적으로 기술할 수 있도록 하였다. 자원의 이용권한과 관련하여서도, 시스템에서의 접근 제한을 표현하는 'rightsOverride'나 엠바고 기간(embargoReleaseDate), 엠바고 기간 중 접근상태 표시(visibilityDuringEmbargo), 엠바고 만료 후 접근상태 표시(visibilityAfterEmbargo) 등을 두어 연구데이터의 이용과 접근권한을 확실하게 표현할 수 있도록 하였다.

〈표 5〉 연구데이터 수용을 위한 OAK 메타데이터 확장(안)

구분	요소명	정의	통제어휘	필수 여부	반복 여부	적용수준
1	title	자원의 표제나 제목, 명칭	-	M	N	item file collection
1.1	title.alternative	공식적으로 사용되는 표제와 다른 문자나 형식으로 된 표제 혹은 부차적 표제	-	O	Y	item collection
1.2	title.original	원본의 표제	-	O	Y	item
1.3	title.partNumber	표제를 세분하는 권차, 회차, 연차, 편차	-	O	Y	item
1.4	title.partName	권차, 회차, 연차, 편차에 부여된 표제나 제목	-	O	Y	item

가 있다. 셋째, 가급적 다양한 유형의 실제 연구데이터를 입력하면서 최종 점검할 필요가 있다.
이러한 검토 의견을 반영하여 먼저, 연구데이터의 정제 정도를 기술하기 위한 'dataVersion' 요소를 description 요소 아래에 추가하고 데이터 값을 raw, refinded, trusted 세 가지로 구분 적용하는 방안을 보완하였다. 또한 연구데이터의 유형을 일반적인 자원유형(예: image, text 등)과 구분하여 보다 세부적으로 기술하기 위한 방안으로써 type 요소 아래에 subResourceType을 추가하고, 관련 사례를 참고하여 연구데이터의 유형에 적용할 통제어휘로 'observational', 'experimental', 'derived', 'simulation', 'reference'를 추출하였다. 마지막으로 연구팀에서 도출한 안을 중심으로 다양한 사례를 적용해 봄으로써 요소 또는 요소값의 적합성을 검증하였다.

구분	요소명	정의	통제어휘	필수 여부	반복 여부	적용수준
2	creator	자원에 책임이 있거나 기여한 개인이나 단체	-	미사용	미사용	미사용
3	subject	자원의 주제나 토픽, 학문분야	〈subjectType〉 KDC DDC LCC UDC LCSH MESH NLSH scientificName local other	O	Y	item
3.1	subject.schemeURI	표준 분류체계의 URI	-	O	Y	item
3.2	subject.keyword	주제어, 키워드, 핵심어 등 저자나 입력자가 임의 부여한 통제되지 않은 자연어	-	O	N	item collection
4	description	자원에 관한 설명어구	-	O	Y	item file collection
4.1	description.abstract	요약이나 초록	-	O	Y	item collection
4.2	description.tableOfContents	목차나 본문의 구성, 순서	-	O	Y	item collection
4.3	description.statementOfResponsibility	자원에 기재된 저자 표시	-	O	Y	item
4.4	description.patentClaim	특허와 관련된 청구항	-	O	Y	item
4.5	description.sponsorship	지원·후원기관	-	O	Y	item
4.5.1	description.sponsorship.funderName	지원·후원한 기관의 이름	-	O	Y	item
4.5.2	description.sponsorship.funderIdentifier	지원·후원한 기관의 식별자	-	O	Y	item
4.5.3	description.sponsorship.awardTitle	지원·후원한 사업명이나 과제명	-	O	Y	item
4.5.4	description.sponsorship.awardNumber	지원·후원과 관련된 사업번호나 과제번호	-	O	Y	item
4.5.5	description.sponsorship.awardURI	지원·후원한 사업명이나 과제명의 URI	-	O	Y	item
4.6	description.degree	학위의 종류	〈degreeType〉 doctoral master	O	N	item
4.7	description.eprintVersion	원문의 버전	〈eprintType〉 preprint postprint published	O	N	item
4.8	description.dataVersion	데이터셋의 정제 정도 표현	〈dataVersionType〉 raw refinded trusted	O	N	item
4.9	description.provenance	자원의 관리내역	-	O	Y	item collection
4.10	description.biography	자원과 관련된 인물 이력	-	O	Y	item
4.11	description.edition	자원의 판이나 버전	-	O	Y	item
4.12	description.materialDetails	자원의 물리적 특성 표현	-	O	Y	item
4.13	description.methods	연구 수행 방법	-	O	Y	item
4.14	description.technicalDetails	자원의 기술적 요건	-	O	Y	item
5	publisher	자원의 발행처나 배포처	-	O	Y	item
5.1	publisher.location	발행처나 배포처가 위치한 국가나 지역	-	O	Y	item
6	contributor	자원의 생산자 또는 기여자 유형	〈contributorType〉 author advisor editor translator illustrator	MA (그룹)	Y (그룹)	item collection

구분	요소명	정의	통제어휘	필수 여부	반복 여부	적용수준
			examiner department reviewer other			
6.1	contributor.contributorName	자원의 생산자 또는 기여자의 이름	-			item collection
6.1.1	contributor.nameIdentifier	이름 식별자	〈nameIdentifierType〉 orcid viaf isni scopusId reseracherId localId other			item
6.1.2	contributor.schemeURI	이름 식별자의 URI	-			item
6.1.3	contributor.alternativeName	이름의 상이한 표현 형식	-			item
6.2	contributor.affiliation	소속기관명	-			item collection
6.2.1	contributor.affiliationIdentifier	소속기관 식별자	〈affiliationIdentifierType〉 rorId scopusId other			item
6.2.2	contributor.affiliation.schemeURI	소속기관 식별자의 URI	-			item
6.3	affiliatedAuthor	해당 기관의 소속 저자 표시				item
6.4	approver	데이터를 입력한 사람이나 단체	-	O	Y	item
7	date	자원과 관련된 날짜	-	O	Y	item
7.1	date.created	자원을 창작, 제작, 생산한 날짜 또는 파일 생성일	-	O	N	item file collection
7.2	date.valid	자원의 유효한 날짜나 날짜범위	-	O	N	item
7.3	date.available	자원을 이용할 수 있는 날짜	-	O	N	item
7.4	date.issued	자원을 공식적으로 발행 또는 배포한 날짜	-	O	N	item
7.5	date.modified	자원의 내용이 변경된 날짜 또는 파일 수정일	-	O	N	item file collection
7.6	date.dataaccepted	리포지토리에 자원을 접수한 날짜	-	O	N	item
7.7	date.accessioned	리포지토리에 자원을 등록한 날짜	-	O	N	item
7.8	date.datecopyright	저작권 일자	-	O	N	item
7.9	date.datesubmitted	리포지토리에 자원을 제출한 날짜	-	O	N	item
7.10	date.awarded	학위를 수여한 날짜	-	O	N	item
7.11	date.application	특허를 요청한 날짜	-	O	N	item
7.12	date.registration	특허를 등록·공시한 날짜	-	O	N	item
7.13	date.event	연구 이벤트와 관련된 날짜	-	O	N	item
8	type	자원의 유형	〈resourceType〉 DCMI Type MARC genre DSpace RDA dataCite local other	M	Y	item
8.1	type.subType	자원의 하위 유형으로 연구데이터 데이터셋의 유형 구분	〈subResourceType〉 observational experimental derived simulation reference	M	Y	item
9	format	파일 형식	-	O	Y	item file

구분	요소명	정의	통제어휘	필수 여부	반복 여부	적용수준
9.1	format.medium	물리적 매체	-	O	Y	item file
9.2	format.extent	파일의 크기나 재생시간 등	-	O	Y	item file collection
9.3	format.version	파일의 버전	-	O	Y	item
10	identifier	자원 식별자	〈identifierType〉 ISBN ISSN LISSN ISMN ISTC DOI UCI URI govdoc patentRegistrationNumber patentApplicationNumber SICI PMID scopusid wosid ark arXiv localId	MA	Y	item
10.1	identifier.bibliographicCitation	해당 자원에 대한 형식화된 인용표시	-	O	Y	item
11	source	자원이 유래한 자원에 대한 정보	-	미사용	미사용	file
12	language	자원의 언어	-	O	Y	item
13	relation	다른 자원과의 관계	〈identifierType〉 ISBN ISSN LISSN ISMN ISTC DOI UCI URI govdoc patentRegistrationNumber patentApplicationNumber SICI PMID scopusid wosid ark arXiv localId	O	Y	item
13.1	relation.isPartOf	상위자료		O	Y	item
13.2	relation.hasPart	포함자료		O	Y	item
13.3	relation.isFormatOf	다른 형태의 자료(이전)		O	Y	item
13.4	relation.hasFormat	다른 형태의 자료(이후)		O	Y	item
13.5	relation.isVersionOf	이전 버전		O	Y	item
13.6	relation.hasVersion	최신 버전		O	Y	item
13.7	relation.replaces	선행자료		O	Y	item
13.8	relation.isReplacedBy	후속자료		O	Y	item
13.9	relation.references	참고자료		O	Y	item
13.10	relation.isReferencedBy	인용된 자료		O	Y	item
13.11	relation.require	필수자료		O	Y	item
13.12	relation.isRequiredBy	요구자료		O	Y	item
13.13	relation.conformsTo	관련표준		O	Y	item
13.14	relation.isPartOfSeries	관련 총서		O	Y	item
13.15	relation.isSupplementTo	부록자료		O	Y	item
13.16	relation.isSupplementedBy	기본자료		O	Y	item
13.17	relation.isContinuedBy	계속자료(이전)		O	Y	item
13.18	relation.continues	계속자료(이후)		O	Y	item
13.19	relation.hasMetadata	자원의 메타데이터		O	Y	item
13.20	relation.isMetadatafor	메타데이터의 대상 자원		O	Y	item
13.21	relation.isNewVersionOf	과거버전 자원	-	O	Y	item
13.22	relation.isPreviousVersionOf	최신버전 자원	-	O	Y	item
13.23	relation.isDocumentedBy	설명 자원	-	O	Y	item
13.24	relation.document	설명 대상 자원	-	O	Y	item
13.25	relation.isCompiledBy	편집 자원		O	Y	item
13.26	relation.compiles	편집 대상 자원		O	Y	item
13.27	relation.isIdenticalTo	동일 내용 자원		O	Y	item

구분	요소명	정의	통제어휘	필수 여부	반복 여부	적용수준
13.28	relation.isReviewedBy	리뷰 자료		O	Y	item
13.29	relation.reviews	검토 대상 자료		O	Y	item
13.30	relation.isDerivedFrom	원본 자원		O	Y	item
13.31	relation.isSourceOf	파생 자원		O	Y	item
14	coverage	자원의 시간적, 공간적 범위	-	O	Y	item
14.1	coverage.spartial	자원과 관련된 공간적 범위	-	O	Y	item
14.1.1	coverage.spartial.geoLocationPoint	자원과 관련된 공간적 범위(특정 지점)	-	O	Y	item
14.1.2	coverage.spartial.geoLocationBox	자원과 관련된 공간적 범위(사각형 박스)	-	O	Y	item
14.1.3	coverage.spartial.geoLocationPolygon	자원과 관련된 공간적 범위(다각형 박스)	-	O	Y	item
14.2	coverage.temporal	자원과 관련된 시간적 범위	-	O	Y	item
15	rights	자원의 이용, 소유, 접근 등의 권한	-	O	Y	item collection
15.1	rights.rightsHolder	저작권 소유자	-	O	Y	item
15.2	rights.copyrightStatus	저작권 상태	〈copyrightStatusType〉 copyrighted public domain unknown	O	N	item
15.3	rights.rightsStatement	저작권 정보	-	O	Y	item
15.4	rights.rightsOverride	시스템 에서 접근 불가	〈rightOverrideType〉 suppress-discovery metadata-only culturally-sensitive	O	Y	item
15.5	rights.license	라이선스(이용 권한)	-	O	N	item collection
15.6	rights.accessRights	접근권한	〈accessRightsType〉 free access free access with registration available for purchase available by subscription limited free access	O	Y	item file
15.7	rights.embargoReleaseDate	엠바고 만료 일자	-	O	Y	item
15.8	rights.visibilityDuringEmbargo	엠바고 기간 중 접근상태 표시	〈visibilityDuringEmbargoType〉 private institute	O	Y	item
15.9	rights.visibilityAfterEmbargo	엠바고 기간 후 접근상태 표시	〈visibilityAfterEmbargoType〉 public	O	Y	item
16	citation	인용 표시	-	O (그룹)	Y	item
16.1	citation.title	자원이 수록된 저널명 등	-		Y	item
16.2	citation.volume	자원이 수록된 저널 등의 권	-		Y	item
16.3	citation.number	자원이 수록된 저널 등의 호	-		Y	item
16.4	citation.date	자원의 발행일자(URL 인용일자)	-		Y	item
16.5	citation.startPage	자원이 수록된 저널 등의 시작페이지	-		Y	item
16.6	citation.endPage	자원이 수록된 저널 등의 종료페이지	-		Y	item
16.7	citation.conferenceName	자원이 수록된 회의자료의 제목	-		Y	item
16.8	citation.conferenceNumber	자원이 수록된 회의자료의 회차	-		Y	item
16.9	citation.conferencePlace	자원이 수록된 회의자료의 개최지	-		Y	item
16.10	citation.conferenceDate	자원이 수록된 회의자료의 개최일자	-		Y	item
16.11	citation.author	자원이 수록된 저널 등의 저자	-		Y	item
16.12	citation.edition	자원이 수록된 저널 등의 판	-		Y	item
16.13	citation.place	자원이 수록된 저널 등의 장소	-		Y	item
16.14	citation.publisher	자원이 수록된 저널 등의 발행처	-		Y	item
16.15	citation.URL	자원의 URL	-		Y	item

* 필수여부: M-필수, MA-해당시 필수, O-선택 ** 반복여부: Y-반복가능, N-반복불가

Ⅴ. 결 론

본 연구에서는 국립중앙도서관의 오픈액세스 리포지토리인 OAK에 연구과정에서 생산되는 다양한 연구데이터를 기술할 수 있도록 연구데이터를 위한 요소를 조사한 후 OAK 메타데이터 확장 방안을 제안하였다. 이를 위한 연구방법으로 문헌연구, 사례조사, 관계자들과의 면담을 실시하였다.

요소 안을 도출하기 위해 연구데이터를 위한 표준 메타데이터를 분석하고, 연구데이터를 구축・관리하고 있는 국내 2개 기관 및 국외 2개 기관을 대상으로 연구데이터의 메타데이터 요소, 입력지침, 입력사례를 수집하여 분석하였다. 이를 바탕으로 최종적으로 연구데이터 기술을 위한 OAK 메타데이터의 확장 방안을 다음과 같이 도출하였다.

우선, 연구데이터를 위한 모델링 측면에서 컬렉션 〉 아이템 〉 파일의 구조를 그대로 유지하고 컬렉션은 해당 연구데이터를 묶을 수 있는 상위 그룹으로 두고, 아이템에는 연구데이터의 메타데이터와 파일을 묶어 연구과정 중에 생성되는 다양한 포맷의 파일이나 디지털 객체를 함께 제공하는 구조를 제안하였다.

둘째, 표준, 사례 기관의 메타데이터를 기존 OAK 메타데이터와 매핑하여 연구데이터의 기술을 위해 OAK에 추가할 필요가 있다고 판단되는 요소를 선별하여 기존 OAK에 반영하였다. 그 결과 기존 OAK 메타데이터 요소와 비교할 때 subject, description, contributor, date, type, format, relation, coverage, rights, citation 등에서 요소 구분이 확장되었다. 특히, 연구지원기관(funderName), 연구과제명(awardTitle), 연구과제번호(awardNumber), 연구데이터의 정제 수준(dataVersion), 연구방법론(methods), 연구데이터의 세부 유형 구분(subType)을 신설하고, 자원의 이용권한과 관련한 시스템 접근 제한(rightsOverride), 엠바고 기간(embargoReleaseDate), 엠바고 기간 중 접근상태(visibilityDuringEmbargo), 엠바고 만료 후 접근상태(visibilityAfterEmbargo) 등을 두어 연구데이터의 이용과 접근 권한을 표현할 수 있도록 하였다.

셋째, 구조화된 데이터를 통해 검색이나 추후 통계 등에 활용할 수 있도록 통제어휘집과 구문에 대한 사항도 제시하였다. 특히 통제어휘집으로 〈funderIdentifierType〉, 〈dataVersionType〉, 〈subResourceTypes〉, 〈rightOverrideType〉를 추가하였다.

OAK에서 연구데이터를 기술할 수 있도록 메타데이터를 확장 방안을 제안한 본 연구는 국내에서 산출되는 연구데이터가 공식적으로 저장되어 활용될 수 있는 기반을 마련했다고 할 수 있다. 이는 국가적으로 연구의 중복을 방지하고, 연구 산출물을 공유 및 재활용할 수 있는 인프라 구축에 기여한 것으로 의의를 가진다. 본 연구에서 제안한 확장(안)을 적용하여 OAK에서 연구데이터를 구축해 나가더라도 현장의 다양한 의견을 폭넓게 수렴하는 한편, 보다 발전적인 연구데이터의 유지 관리를 위해 추가적인 연구도 필요할 것이다.

참 고 문 헌

국가과학기술연구회. 2019.『출연(연) 연구데이터 관리・활용 방안 연구』. 서울: 동연구회, 2018-09.

한국과학기술정보연구원. 2019.『메타데이터 설계 지침서』. 대전: 동연구원.

한국연구재단 기초학문자료센터 홈페이지.
〈https://www.krm.or.kr/〉 [인용 2020. 1. 5].

DataCite Metadata Working Group. 2019. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data (Version 4.3)*.
〈https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf〉 [cited 2020. 1. 9].

DCMI 2007. *Dublin Core™ Collection Description Application Profile*.
〈https://www. dublincore.org/specifications/dublin-core/collection-description/collection-ap-summary/2007-03-09/〉 [cited 2020. 1. 9].

DCMI. 2012. *Dublin Core™ Metadata Element Set, Version 1.1: Reference Description*.
〈https://www.dublincore.org/specifications/dublin-core/dces/〉 [cited 2020. 1. 9].

Farnel, Sharon and Ali Shiri. 2014. "Metadata for Research Data: Current Practices and Trends." *Proceedings of International Conference on Dublin Core and Metadata Applications 2014*. 〈https://dcpapers.dublincore.org/pubs/article/view/3714〉 [cited 2020. 1. 9].

NC State University Libraries. 2019. *Defining Research Data*.
〈https://www.lib.ncsu.edu/data-management/define〉 [cited 2019. 12. 27].

University of California in San Diego Library Digital Collections Home Page.
〈https://library.ucsd.edu/dc/〉 [cited 2020. 1. 20].

University of Michigan Deep Blue Home Page.
〈https://deepblue.lib.umich.edu/〉 [cited 2020. 1. 20].

University of Pittsburgh. *Guidelines on Research Data Management*.
〈http://www.provost.pitt.edu/documents/RDM_Guidelines.pdf〉 [cited 2019. 12. 27].

Pennington, A. 2020, January 8, 〈apennington@UCSD.EDU〉 [E-mail Interview].

Carlson, J. 2020. January 15, 〈jakecar@umich.edu〉 [E-mail Interview].

• **국한문 참고문헌의 영문 표기**
(English translation / Romanization of references originally written in Korean)

Korea Institution of Science and Technology Information. 2019. *Guideline for the Metadata of Research Data*. Daejon: Korea Institution of Science and Technology Information.

Korean Research Memory Homepage. 〈https://www.krm.or.kr/〉 [cited 2020. 1. 5].

National Research Council of Science & Technology. 2019. *Research Data Management and Utilization for Government-Funded Research Institutes in Korea*. Seoul: National Research Council of Science & Technology, 2018-09.