Journal of Korean Library and Information Science Society
Vol. 51, No. 4

생태 분야 연구데이터를 위한 메타데이터 설계*

- DCAT을 중심으로 -

Metadata Design for Ecological Research Data: Focused on DCAT

김 주 섭(Juseop Kim)**, 윤 희 남(Heenam Yoon)***
권 용 수(Yong-su Kwon)****, 김 선 태(Suntae Kim)*****

〈 목 차 〉

요 약 : 본 연구의 목적은 생태 분야에서 생산되는 연구데이터의 관리 및 공유를 위한 메타데이터를 설계하기 위함이다. 특히, 메타데이터는 웹에 게시된 데이터 카탈로그 사이의 상호운용성을 용이하게 하도록 설계된 DCAT을 기반으로 설계되었다. 또한, 생태분야의 특성을 반영하기 위하여 ABCD, Darwin Core 그리고 EML 등을 분석하여 메타데이터 설계에 적용하였다. 연구 결과, 생태 분야 연구데이터를 관리하기 위한 메타데이터 요소는 전체 51개가 도출되었으며 필수 요소는 6개, 권고 요소 23개 그리고 선택요소 22개가 포함되었다. 본 연구 결과는 메타데이터 요소를 기반으로 연구데이터 관리 및 공유시스템 구축 시 DB를 설계할 수 있으며, 타 시스템과 연계 시 메타데이터 교환 형식을 제시하는데 사용할 수 있을 것이다.

주제어 : 메타데이터, 연구데이터, DCAT, 국립생태원, 생태학

ABSTRACT : The purpose of this study is to design metadata for management and sharing of Yeogu data produced in the ecological field. Specifically, metadata is designed based on DCAT, which is designed to facilitate interoperability between data catalogs published on the web. In addition, ABCD, Darwin Core, and EML were analyzed and applied to metadata design to reflect the characteristics of the ecological field. As a result of the research, a total of 51 metadata elements were derived for managing research data in the ecological field, and 6 essential elements, 23 recommended elements, and 22 optional elements were included. The results of this study can be used to design a DB when building a research data management and sharing system based on metadata elements, and to suggest a metadata exchange format when linking with other systems.

KEYWORDS : Metadata, Research Data, DCAT, National Institute of Ecology, Ecology

---

* 본 논문은 환경부의 재원으로 국립생태원의 지원을 받아 수행하였음(NIE-전략연구-2020-01).
이 논문은 2020년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
** 전북대학교 문헌정보학과 강사(kimjuseop@jbnu.ac.kr / ISNI 0000 0004 7492 1806) (제1저자)
*** 국립생태원 생태정보연구실 에코뱅크팀(ecospace@nie.re.kr / ISNI 0000 0004 7357 8209) (공동저자)
**** 국립생태원 생태정보연구실 에코뱅크팀(kwonys@nie.re.kr / ISNI 0000 0004 9218 3421) (공동저자)
***** 전북대학교 문헌정보학과 조교수(kim.suntae@jbnu.ac.kr / ISNI 0000 0004 6492 6355) (교신저자)
• 논문접수: 2020년 11월 26일 • 최초심사: 2020년 11월 26일 • 게재확정: 2020년 12월 14일
• 한국도서관·정보학회지, 51(4), 249-278, 2020. http://dx.doi.org/10.16981/kliss.51.4.202012.249

Ⅰ. 서 론

1. 연구의 필요성 및 목적

국가의 세금으로 생산된 데이터의 원활한 접근을 가능하게 하는 오픈사이언스의 확장은 연구데이터에 대한 접근 및 공유를 넘어서 재사용에 대한 관심으로 확대되고 있다. 국내의 경우, 최근 2019년 9월 1일부터 시행되도록 개정된『국가연구개발사업의 관리 등에 관한 규정』에 따르면 중앙행정기관의 장이 필요하다고 인정하는 연구개발과제의 경우, 연구개발과제의 선정 시 데이터관리계획에 따른 연구데이터 생산・보존・관리의 충실성 및 공동활용 가능성을 검토하도록 하고 있다. 여기에 선제적으로 대응하기 위해 한국한의학연구원 및 한국지질자원연구원에서는 기관에서 생산되는 연구데이터의 수집 및 공유를 위한 리포지토리 시스템을 개발하여 운영 중에 있다. 이러한 추세는 국가출연연구소를 비롯하여 각 연구기관으로 확산되고 있다.

인텔에서는 2020년까지 하루에 인터넷 이용자가 1.5기가바이트를, 자율주행 자동차는 4 테라바이트, 스마트 공장은 1 페타데이트를 마지막으로 유튜브와 같은 클라우드 비디오 프로바이더는 750 페타데이터를 생산한다고 예견하였다(Data Center Frontier 2017). 이러한 데이터의 폭발적 증가는 데이터에 대한 접근을 어렵게 만들고 있으며 데이터의 체계적인 관리가 필요한 동인으로 이끌고 있다. 이에 따라 연구 중에 생산되는 데이터의 체계적인 관리, 공유 및 재사용을 위해 메타데이터의 개발은 필수적이라 할 수 있다.

생태분야에서 생산되는 데이터는 타 학문분야보다 복잡하며 다양한 분야의 연구를 지원할 필요가 있어 그 범위가 매우 넓고 데이터양도 상당하다. 또한 같은 데이터를 개별 연구자들이 각자 상이하게 표현하는 경우도 많을 뿐만 아니라 데이터 해석을 위해 맥락정보의 필요성도 제기되고 있다(Aleksejs 2007; DBBM 2001). 뿐만 아니라 전 세계적으로 생태분야에서 생산되는 데이터가 무수히 많은 웹사이트나 리포지토리를 통해 저장 및 보존되고 있어 이기종의 플랫폼에서 생산되는 데이터 카탈로그에 대한 공유 및 재사용을 위해서라도 DCAT이라는 국제 표준에 기반한 메타데이터 스키마 개발이 필요할 것으로 판단된다.

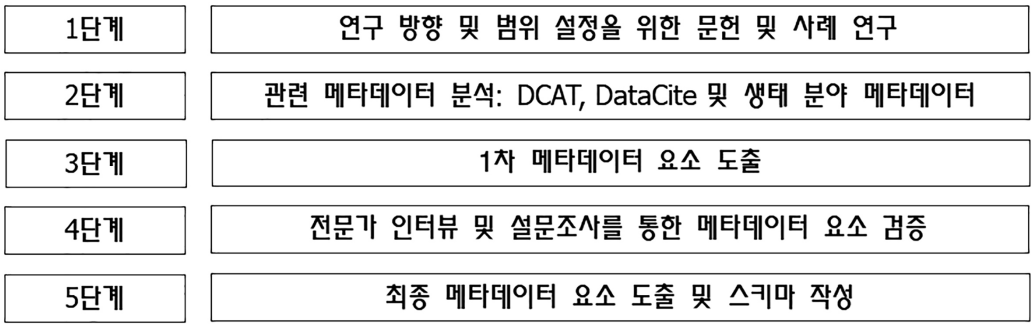
Data Catalog Vocabulary(이하 DCAT)은 2014년 W3C가 승인한 권고표준으로, 웹에 게시된 데이터 카탈로그 사이의 상호운용성을 용이하게 하도록 설계된 RDF(Resource Description Framework) 어휘이다. 이러한 DCAT은 카탈로그에서 데이터셋과 데이터서비스를 포함하고 설명할 수 있도록 RDF 클래스와 속성을 제공하고 있다. 또한 여러 카탈로그에서 메타데이터를 쉽게 사용하고 수집할 수 있는 표준 모델 및 어휘를 사용하여 카탈로그에서 데이터셋 및 데이터 서비스를 기술할 수 있다. DCAT의 유연한 확장성으로 인해 CKAN, DKAN, Socrata 등 대표적인 오픈데이터 플랫폼에서 DCAT을 지원하고 있으며 CKAN은 대표적인 오픈데이터 플랫폼 중 하

나로서 CKAN 하베스팅과 DCAT 하베스팅을 지원하여 플랫폼 간에 데이터셋 정보를 공유하고 있다. 여기에 국가의 대표적인 공공데이터 포털인 data.gov와 data.gov.uk에서도 DCAT 하베스팅 기능을 제공하여 이기종의 플랫폼과의 데이터 공유를 지원하고 있는 추세이다(박경현, 원희선, 류근호 2018).

이에 따라 본 연구에서는 생태분야에서 생산되는 데이터를 관리하기 위하여 DCAT을 기반으로 한 메타데이터 스키마를 개발하고자 한다. 개발된 생태분야 메타데이터를 이용하여 관련 기관에서 생산되는 연구데이터를 관리 및 재사용하기 위한 플랫폼 설계가 가능할 것이다. 또한, 제안되는 메타데이터 요소를 기반으로 연구데이터 관리 및 공유시스템 구축 시 데이터베이스를 설계할 수 있으며, 타 시스템과 연계 시 메타데이터 교환 형식을 제시하는데 사용할 수 있다. 마지막으로 생태 분야의 표준화된 데이터 제공으로 데이터의 활용도를 제고할 수 있다.

2. 연구 방법 및 절차

본 연구의 목적을 달성하기 위해 사용된 연구 방법 및 절차는 다음 〈그림 1〉과 같다.



〈그림 1〉 연구방법 및 절차

첫 번째 단계에서는 연구의 전반적인 방향과 범위를 설정하고 두 번째 단계에서는 생태 분야 메타데이터를 도출하기 위하여 DCAT 및 관련 메타데이터 표준을 분석하였다. 분석 대상인 생태 분야 메타데이터에는 ABCD, DarwinCore 그리고 EML 등 3개의 메타데이터 표준이 포함되었다. 세 번째 단계에서는 생태 분야 연구데이터 관리를 위한 메타데이터 요소를 도출하였다. 메타데이터 요소를 도출하기 위해 2단계에서 분석한 메타데이터 표준을 대상으로 크로스워크를 실시하였다. 네 번째 단계에서는 3단계에서 도출된 메타데이터 요소 검증을 위해 생태 분야 이해관계자와 메타데이터 전문가를 대상으로 인터뷰와 설문조사를 실시하였다. 마지막으로 4단계에서 진

행된 검증을 통해 생태 분야 연구데이터를 관리하기 위해 메타데이터 요소 도출 및 스키마를 작성하였다.

특히 메타데이터 요소를 도출하기 위해 사용된 크로스워크는 메타데이터 간 상호운용성을 확보하기 위한 방법으로 이번 생태 분야의 메타데이터 요소 도출에 필요한 절차라고 할 수 있다. 또한 크로스워크 진행 시 각 메타데이터의 모든 요소를 대상으로 하기에는 한계가 있어 최상위요소 및 데이터셋을 기술하기 위한 메타데이터 요소만을 매핑했다. 이 점은 본 연구의 제한점으로 제시될 수 있을 것이다.

Ⅱ. 이론적 배경

1. DCAT

DCAT(Data Catalog Volcabulary, 이하 DCAT)은 웹에서 발행된 데이터 카탈로그들 간 상호운용성 향상을 위해 설계된 RDF(Resource Description Framework) 어휘로 W3C에서 2014년 1월 16일 웹 표준으로 권고 승인되었다. 오리지널 DCAT 어휘([VOCAB-DCAT-20140116])는 DERI(Digital Enterprise Research Institute)가 개발하였으며 이를 eGov Interest Group가 개선하여 2014년에 Goverment Linked Data(GLD) Working Group이 표준화하였다. DCAT 1.0은 유럽, 미국, 호주 등 전 세계 오픈 데이터 포털에서 활용하고 있으며, 다양한 분야에서 DCAT을 확장하거나 독립적으로 설계하여 응용 프로파일을 개발하고 있다. 이러한 DCAT은 메타데이터를 RDF 형태로 정의하여 데이터셋과 데이터 서비스를 기술할 수 있게 하며, 데이터의 접근과 활용을 보장한다. 또한, DCAT은 Dublin Core, FoaF, SKOS 등 기존의 어휘에서 차용한 용어들을 사용하고 있다. 표준 모델 및 어휘를 사용함으로써 여러 카탈로그들로부터 온 메타데이터의 이용과 통합을 지원하고 데이터셋 및 데이터서비스의 검색 가능성을 증대시킨다. 여기에 산재되어 있는 데이터셋의 통합검색을 가능하게 한다. 이렇게 모인 DCAT 메타데이터는 디지털 보존 프로세스의 일부로서 매니페스트 파일(manifest file)의 역할을 할 수 있다. 본래 DCAT 어휘는 Digital Enterprise Research Institute(DERI)에서 개발하였으나 이를 eGov Interest Gropu에서 정제하고 2014년에 Government Linked Data(GLD) 워킹 그룹에서 표준화하였다. Dataset Exchange 워킹 그룹은 2014년 권고 표준 발표 이후 사람들의 경험을 통해 수집된 사례와 요건을 반영하여 DCAT 개정판을 개발하였다. 이후 2019년에 제약조건을 완화하고 새로운 클래스와 속성을 추가한 DCAT V2를 발표하였다. 다음의 〈표 1〉은 DCAT V2의 13개 클래스를 정리한 것이다(W3C 2019).

〈표 1〉 DCAT V2의 클래스

명칭	설명	메인 클래스	비고
dcat:Catalog (카탈로그)	데이터셋 및 데이터서비스에 관한 선별된(curated) 메타데이터 집합	O	• dcat:Dataset의 서브 클래스
dcat:Resource (리소스)	단일 에이전트에 의해 게시되거나 선별된 자원	O	• DCAT V2에서 추가됨
dct:Dataset (데이터셋)	단일 기관에 의해 게시되거나 선별된 데이터의 집합으로 하나 이상의 표현형으로 다운로드하거나 액세스할 수 있음	O	• dcat:Resource의 서브 클래스
dcat:Distribution (배포)	데이터셋에 대한 구체적인 표현으로 데이터셋을 이용하기 위한 정보를 서술	O	
dcat:DataService (데이터서비스)	하나 이상의 데이터셋 또는 데이터 처리 기능에 대한 액세스를 제공하는 기능모음	O	• dcat:Resource의 서브 클래스 • dctype:Service의 서브 클래스 • DCAT V2에서 추가됨
dcat:CatalogRecord (카탈로그 레코드)	누가, 언제 항목을 추가하였는가와 같은 등록 정보를 주로 나타내는 카탈로그의 메타데이터 항목을 서술	O	
skos:ConceptScheme (분류체계)	카탈로그에서 데이터셋의 테마/카테고리를 표현하기 위해 사용된 지식구조체계(KOS)		
skos:Concept (카테고리)	카탈로그에서 데이터셋을 기술하기 위해 사용된 테마 또는 카테고리		
foaf:Person / foaf:Organization (인물/기관)	인물은 foaf:person, 정부 기관 및 기타 엔티티는 foaf:organization로 나타냄		
dcat:Relationship (관계)	DCAT 자원 간 관계에 추가적인 정보를 첨부하기 위한 연관 클래스		• prov:EntityInfluence의 서브 클래스 • DCAT V2에서 추가됨
dcat:Role (역할)	자원 속성 또는 자원 관계의 맥락에서 다른 자원과 관련한 자원 또는 기관의 기능		• skos:Concept의 서브 클래스 • DCAT V2에서 추가됨
dct:PeriodOfTime (기간)	시작과 끝으로 명명되거나 정의된 시간의 간격		• DCAT V2에서 추가됨
dct:Location (장소)	공간 영역 또는 명명된 장소		• DCAT V2에서 추가됨

DCAT V2는 Catalog, Cataloged Resource, Catalog Record, Dataset, Distribution, Data Service, Concept Scheme, Concept, Organization/Person, Relationship, Role, Period of Time 그리고 Location 등의 13개 클래스로 구성되어 있으며 이 중에서 메인 클래스는 'dcat:Catalog', 'dcat:Resource', 'dcat:Dataset'. 'dcat:Distribution ', 'dcat:DataService' 그리고 'dcat:CatalogRecord' 등 6개이다. 또한, DCAT V2에서는 이전 버전과 달리 Resource, DataService, Relationship, Role, PeriodOfTime 그리고 Location 등 6개 클래스가 새롭게 추가되었다. 다음의 〈표 2〉는 DCAT V2의 메인 클래스를 나타낸 것이다(W3C 2019).

〈표 2〉 DCAT V2의 메인 클래스

클래스명	내용
dcat:Catalog	• 정보자원을 기술하기 위한 메타데이터 • dcat:Catalog의 범위는 데이터셋 또는 데이터서비스 관련 메타데이터의 집합
dcat:Resource	• 카탈로그에서 메타데이터 레코드에 의해 기술되는 데이터셋, 데이터서비스 등의 어떤 정보자원을 나타냄 • dct:Resource 클래스는 직접적으로 사용되는 것은 아니며 dcat:Dataset, dcat:DataService, dcat:Catalog 클래스의 모체 클래스임 • 카탈로그의 구성 항목은 DCAT 프로파일 또는 DCAT 어플리케이션 프로파일에서 정의된 dcat:Resource 클래스의 하위 클래스 또는 그 이하의 구성 요소 중 하나여야만 함
dcat:Dataset	• 단일 에이전트에 의해 발행된 데이터의 집합 • 데이터는 데이터셋으로 수집될 수 있는 잠재적인 모든 형태의 유형으로부터 올 수 있음
dcat:Distribution	• 다운로드 파일과 같은 데이터셋의 접근가능한 형식을 나타냄
dcat:DataService	• 인터페이스를 통해 접근 할 수 있는 기능의 집합 • 데이터 프로세싱 기능 또는 하나 또는 그 이상의 데이터셋으로의 접근을 제공
dcat:CatalogRecord	• 카탈로그 내 메타데이터 항목을 나타내는 것 • 주로 등록 정보와 같이 카탈로그 항목에 대한 출처 정보를 명시적으로 나타낼 때 사용됨 • 해당 클래스는 선택적(optional)으로 사용함

위 〈표 2〉에서 나타난 바와 같이 6개의 주요 클래스 중 DCAT V2에서 새롭게 추가된 메인 클래스는 dcat:Resource 클래스로서 카탈로그된 모든 정보자원의 클래스이자, dcat:Dataset, dcat:DataService, dcat:Catalog의 슈퍼 클래스이다. 이 클래스는 데이터셋과 데이터 서비스를 포함하는 모든 카탈로그된 정보자원에 공통적인 속성으로 적용된다. 카탈로그에서 사용되는 항목은 DCAT 또는 DCAT 어플리케이션 프로파일의 dcat:Resource 클래스 하위의 속성 또는 클래스에서 명시되어야 한다. W3C는 dcat:Resource 클래스 아래 더 구체적인 하위 클래스를 사용할 것을 권고하고 있다.

다음의 〈표 3〉은 DCAT을 기반으로 개발된 DCAT 응용 프로파일(Application Profile)을 나타낸 것이다(W3C 2019).

〈표 3〉 DCAT 응용 프로파일

분야	프로파일명	연도	비고
공통	DCAT- 응용 프로파일	2018	유럽에서 개발한 DCAT 프로파일로 정부 데이터 카탈로그 및 과학 데이터를 위한 메타데이터 교환 형식으로 사용됨
지리	GeoDCAT- 응용 프로파일	2016	지리공간 데이터를 기술하기 위한 DCAT-응용 프로파일의 확장
통계	StatDCAT- 응용 프로파일	2016	통계데이터를 기술하기 위한 DCAT-응용 프로파일의 확장
연구데이터	CiteDCAT- 응용 프로파일	2019	다학문적 연구 데이터를 위해 설계된 DCAT-응용 프로파일 확장으로 Zenodo에서 지원됨
연구데이터	DCAT- 응용 프로파일-JRC	2019	JRC에서 개발한 DCAT-응용 프로파일로 다학문적 연구 데이터의 문서화를 위해 JRC 데이터 카탈로그에서 사용됨

DCAT 응용 프로파일은 2014년 이후 유럽 전역에서 정부 데이터 카탈로그와 과학데이터를 위한 메타데이터 교환 형식으로 사용되는 DCAT 프로파일이다. DCAT 응용 프로파일을 확장하여 통계(StatDCAT), 지리(GeoDCAT) 등의 분야에서 DCAT 응용 프로파일이 개발되었다. 이 외에도 타 메타데이터 표준(DataCite)과 매핑한 CiteDCAT-응용 프로파일, 다학문 연구데이터를 위해 DCAT-응용 프로파일을 확장한 DCAT-응용 프로파일-JRC 등이 개발되었으며 특히, CiteDCAT-응용 프로파일은 유럽에서 가장 광범위하게 사용되는 리포지토리이자 연구데이터 카탈로그인 Zenodo에서 지원된다.

2. 선행 연구

이번 절에서는 DCAT을 기반으로 설계된 메타데이터와 관련한 국내외 논문 8편에 대하여 살펴보고자 한다.

〈표 4〉 DCAT 관련 선행연구 목록

제목	발행연도	비고
DCAT을 활용한 디지털도서관 데이터셋 관리와 서비스 설계	2019	디지털 도서관
교통 분야 Data 관리와 공유를 위한 DCAT 기반 Data Catalogue 표준 연구: DCAT-Trans	2019	교통 분야
DCAT Application Profile for data portals in Europe	2018	유럽 데이터포털
연구데이터 관리를 위한 온톨로지 설계에 대한 연구	2018	연구데이터
오픈데이터 플랫폼의 상호운용성을 위한 DCAT 기반 메타데이터 변환도구 설계 및 구현	2018	오픈데이터
BotDCAT-AP: An Extension of the DCAT Application Profile for Describing Datasets for Chatbot Systems	2017	Chatbot 시스템
GeoDCAT-AP: Representing geographic metadata by using the “DCAT application profile for data portals in Europe”	2017	지리공간정보
StatDCAT-AP, A Common Layer for the Exchange of Statistical Metadata in Open Data Portals	2016	통계데이터

박진호(2019)는 DCAT을 활용하여 디지털도서관에서 데이터셋을 관리 및 서비스할 수 있는 시스템을 제안하였다. DCAT 구조 및 구성요소를 분석하여 4개의 주 클래스와 그에 속하는 속성과 기술방법에 대한 상세한 설명을 제시하였다. 해당 시스템은 원천데이터, 데이터셋 관리, 링크드 데이터 연결, 이용자 서비스로 구성된다. 원천데이터는 디지털도서관이 직접 수집하는 데이터셋, CKAN과 DKAN 등 데이터셋 관리 프레임워크, 자체 개발 운영 중인 데이터 플랫폼의 데이터 등을 대상으로 하였다. 이 중 데이터셋 관리는 입수, 정제, 연동 그리고 DCAT 매핑관리로 이루어

져 있으며 DCAT을 중심으로 다양한 어휘들과 매핑할 수 있는 매핑 관리 기능이 핵심이다. CKAN, DKAN의 데이터 요소와 DCAT을 매핑한 결과, CKAN은 DCAT의 모든 클래스와 속성에 대응이 가능하나 사용하는 네임스페이스와 필드명이 달라 반드시 매핑 작업이 요구되며 DKAN의 경우 DCAT의 일부만 지원하고 있는 것으로 나타났다.

신도겸 외(2019)는 메타데이터를 관리하기 위해 DCAT을 기반으로 교통 데이터의 데이터 카탈로그 작성 표준인 DCAT-Trans를 제안하였다. 연구방법은 DCAT을 적용한 해외 데이터 카탈로그 작성 사례를 분석하고 DCAT 표준 1.0과 2.0의 개정사항을 분석하였다. DCAT을 적용한 해외 데이터 카탈로그 작성 사례를 조사한 결과, 대부분의 데이터 포털은 DCAT에서 필요한 속성을 취사선택하여 사용하고 있으며 모든 속성을 사용하는 데이터 포털은 존재하지 않았다. 해당 연구에서는 기준 표준을 활용하여 교통 데이터의 특성에 맞게 DCAT의 클래스와 속성을 변경하거나 새롭게 제안하였다. DCAT의 Location 클래스를 개선하고 Relationship 클래스의 하위 속성을 변경하였으며 Dataset 클래스에 포함된 dcat:theme을 Taxonomy 클래스로 분리하였다.

DCAT 응용 프로파일은 유럽 데이터 포탈을 위한 DCAT 어플리케이션 프로파일로 W3C가 개발한 Data Catalogue Vocabulary(DCAT)을 기반으로 기존 메타데이터를 매핑하고 기존 통제어휘를 재사용하여 다른 어플리케이션과 상호운용성을 확보하면서 유럽의 데이터포털의 어플리케이션 수요를 충족시키기 위한 목적으로 개발되었다. 유럽 내 공공분야 데이터셋을 기술할 수 있는 공통 명세를 제공함으로써 데이터 포털 간 데이터셋 설명(description)의 교환을 지원한다. 데이터 재사용자는 타 회원국 소속 특정 데이터셋의 존재 여부와 공공 관리가 유지되고 있는지 등에 대한 정보를 파악할 수 있으며 데이터 공급자는 하나 이상의 데이터 포탈에 의해 접근 및 검색 가능하게 만들면서 데이터셋의 재사용을 장려할 수 있다. 해당 응용 프로파일은 총 27개의 클래스로 5개의 필수 클래스(Agent, Catalogue, Dataset, Litera, Resource), 4개의 권고 클래스(Category, Category scheme, Distribution, Licence document) 그리고 18개의 선택 클래스(Catalogue Record, Data Service, Checksum, Document, Frequency, Identifier, Kind, Linguistic system, Location, Media type, Period of time, Publisher type, Relationship, Rights statement, Role, Standard, Status, Provenance Statement)로 구성하였다(Nuffelen 2020).

박옥남(2018)은 국가과학기술정보서비스(NTIS)에 적용할 수 있는 연구데이터 관리를 목적으로 온톨로지를 설계하고 그 예시를 제시하였다. 연구방법은 연구데이터 관련 선행연구를 조사하고 더블린 코어, DDI 메타데이터, DataCite 메타데이터 스키마 4.1, DCAT 등 연구데이터 관련 메타데이터 표준을 비교・분석하였으며 온톨로지 예시를 제시하였다. 또한, ICPSR 데이터 아카이브, Harvard Dataverse 그리고 Dryad 등 연구데이터 리포지토리 사례를 조사하였으며 메타데이터 표준과 리포지토리 분석을 통해 시사점을 도출하고 온톨로지 설계에 적용하였다.

박경현, 원희선, 류근호(2018)는 공공데이터 포털과 같은 오픈데이터 플랫폼 간의 원활한 데이터 공유 및 상호운용성을 위해 DCAT을 기반으로 메타데이터 변환도구를 설계하였다. 오픈데이터 플랫폼은 데이터소유자의 데이터 유통을 지원하고 데이터 사용자의 데이터 검색 및 활용을 가능하게 한다. 또한, 유연한 확장성을 특징으로 하는 DCAT은 CKAN, DKAN 그리고 Socrata와 같은 대표적인 오픈데이터 플랫폼에서 지원하고 있으며 영국의 data.gov.uk와 미국의 data.gov에서도 DCAT 하베스팅 기능을 제공한다. 유럽의 데이터포털은 공공데이터 게시 및 연동을 위해 DCAT 응용 프로파일인 DCAT-AP를 개발하였다.

Cappello, Comerio, Celino(2017)는 챗봇 어플리케이션의 기존 데이터 소스를 원활한 재사용을 위해 챗봇을 위한 데이터셋을 기술하는 BotDCAT-응용 프로파일을 제안하였다. BotDCAT 응용 프로파일은 유럽 데이터 포털의 DCAT 응용 프로파일을 준수하여 확장하였으며 bot:Intent와 bot:EntitiesCatalog을 추가하여 챗봇과 상호작용을 위해 이용자가 달성하고자 하는 목적 및 목적과 연관된 정보인 엔티티를 기술할 수 있도록 하였다. BotDCAT 응용 프로파일은 웹 상에 게시된 데이터셋과 데이터 카탈로그만을 기술할 수 있는 DCAT 응용 프로파일과 달리 bot:hasMethodURL, bot:hasAssetURL, bot:hasDocumentation 등을 이용하여 데이터셋에 접근하는 방법을 정의할 수 있는 것이 특징이다.

Perego et al.(2017)은 유럽 데이터 포털을 위한 DCAT 응용 프로파일을 준수하여 RDF 기반 지리공간정보 메타데이터의 표현을 제공할 목적으로 유럽데이터포털의 DCAT-AP의 확장인 GeoDCAT 응용 프로파일을 개발하였다. GeoDCAT 응용 프로파일은 INSPIRE 인프라를 통해 지리공간정보 메타데이터의 공유를 가능하게 하며 지리 공간 메타데이터 소유자에게 플랫폼 간 공유 및 재사용을 지원한다. INSPIRE 메타데이터 및 ISO 19115:2003의 핵심 프로파일의 모든 요소를 다루면서 지리 공간적 레코드의 DCAT 응용 프로파일 표현과의 조화를 제공하기 위해 INSPIRE 메타데이터, ISO 19115:2003 핵심 프로파일 그리고 DCAT 응용 프로파일을 매핑하였다.

Dekkers et al.(2016)은 통계 데이터셋의 재사용 및 탐색을 향상시키기 위해 유럽 데이터포털의 DCAT 응용 프로파일(DCAT Application Profile for Data Portals in Europe)의 확장 프로파일인 StatDCAT 응용 프로파일를 개발하였다. StatDCAT 응용 프로파일은 DCAT 응용 프로파일을 기반으로 하여 DCAT 응용 프로파일 버전 1.1을 전적으로 준수하면서 통계 데이터 탐색 서비스를 개선하고 오픈 데이터 포털과 기존 통계 데이터 포털의 통합을 촉진하도록 하였다. StatDCAT 응용 프로파일 데이터 모델은 DCAT 응용 프로파일의 Catalogue, Catalogue Record, Dataset, Distribution 등 4개의 주요 엔티티와 통계적 데이터셋의 기술을 위해 추가된 stat:attribute, stat:dimension, dqv:hasQualityAnnotation 등을 포함한다.

이상 8편의 국내외 선행연구를 분석한 결과 DCAT을 기반으로 여러 분야에서 AP가 개발되고

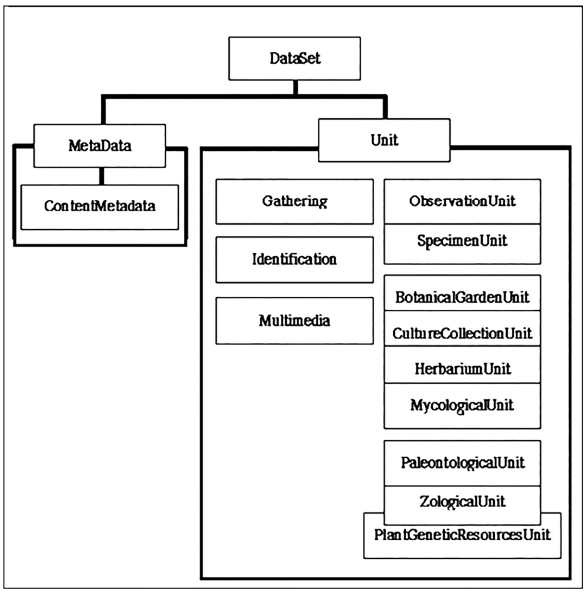
있음을 알 수 있었다. 유럽의 경우, 유럽 위원회를 중심으로 시스템 간 상호운용성을 보장하기 위하여 지리, 통계 그리고 연구데이터에 대한 응용 프로파일이 개발되었으며 국내에서도 디지털 도서관, 교통 그리고 연구데이터를 중심으로 응용 프로파일이 개발된 것을 확인할 수 있었다. 이에 따라 본 논문에서는 DCAT을 기반으로 기존 생태분야의 메타데이터인 ABCD, DarwinCore 그리고 EML 등과 범용적 연구데이터 기술을 위한 표준 메타데이터인 DataCite 4.3을 매핑하여 생태분야에 적합한 메타데이터 요소를 도출하고 스키마를 설계하고자 한다.

Ⅲ. 생태분야 메타데이터 분석

1. Access to Biological Collection Data (ABCD)

ABCD 스키마는 2001년에 개발이 시작되어, 2005년에 TDWG[1] 표준으로 XML 스키마로 비준되었다. ABCD 스키마는 TDWG에서 만든 표본과 관측에 대한 데이터(일차 생물 다양성 데이터)를 접근 및 교환하기 위해 개발된 종합 표준이다. 이러한 ABCD는 다양한 DB의 데이터를 지원하기 위해 포괄적으로 구조화되어 있으며 사용하는 요소와 개념은 HISPID, Darwin core 등과 같은 생물 수집 데이터 분야의 다른 표준과 최대한의 호환성을 제공한다. 또한, ABCD 스키마는 표본이 없는 현장 관측과 더불어, 살아있고 보존된 표본들을 포함한 생물학적 수집 단위를 대상으로 한 공통 데이터 표준이다. 전 세계의 생물학 데이터들은 표본의 특징 요소(예시: 분류군, 고도, 개체의 암수)와 컬렉션(예시: 소장처 등)을 포함한다. 사용되는 요소들은 컬렉션마다 다르며, ABCD는 과학자와 큐레이터들이 사용할 수 있도록 요소명과 정의 표준을 제공하고 있다. ABCD의 설계 목표는 포괄적이고 일반적인 개념으로, 컬렉션 DB에서 사용 가능한 광범위한 개념을 포함하지만, 메타데이터 기능에 필요한 최소한의 요소만 의무화했다. ABCD는 의도적으로 분류학적 데이터, 분포 범위, 지표값과 같은 정보를 스키마에 포함하지 않도록 하였다. 때문에 수많은 메타데이터 스키마 중 최소로 사용하는 요소는 17개에 지나지 않는다. 최소한의 스키마만 의무화했음에도, 다양한 데이터의 종류를 소화하기 위해, 종류마다 메타데이터 스키마를 구성해놓았다. 다른 분야에 응용한다면, 이러한 범용성을 살려 보다 많은 데이터를 수용할 수 있다. 다음의 〈그림 2〉는 ABCD 스키마의 계층 구조를 나타낸 것이다(Access to Biological Collection Data task group 2007).

1) TDWG(Taxonomic Databases Working Group 혹은 Biodiversity Information Standards)는 생물 다양성 정보학을 촉진하는 협회로 Darwin core를 만든 곳이기도 하다. 따라서 ABCD와 Darwin Core는 높은 호환성을 가진다.



〈그림 2〉 ABCD 요소의 계층 구조

위 〈그림 2〉에서 나타난 바와 같이 ABCD는 최상위 요소로 DataSet, MetaData 그리고 Unit 요소로 구성되어 있다. MetaData에서는 ContentMetadata가 직접 연결되어 있다. Unit 하위에는 Gathering, Identification, Multimedia, ObservationUnit, SpecimenUnit, BotanicalGardenUnit, CultureColletcionUnit, HerbariumUnit, MucologicalUnit, PaleontologicalUnit, ZologicalUnit, PlantGeneticResourceUnit이 있다. 다음의 〈표 5〉는 ABCD의 최상위 요소인 DataSet의 하위 요소를 나타낸 것이다.

〈표 5〉 DataSet의 하위 요소

하위요소	설명
GUID	(Globally unique identifier) 데이터셋의 세계적 고유 식별자
ID	데이터셋의 고유 식별자
ResourceURI	데이터셋의 고유 식별 URI
TechnicalContact	데이터베이스 설치 담당자 (Contact 속성 사용)
ContentContact	데이터셋 작성자 (Contact 속성 사용)
DataCenter	데이터셋가 보관되는 기관 또는 데이터 센터의 이름(소유기관과는 다를 수 있음)
OtherProvider	이 데이터셋을 서비스하는 다른 정보제공자
Metadata	전체 데이터 컬렉션의 주요한 항목을 설명하는 메타데이터
Unit	Unit을 설명하는 모든 데이터를 담은 컨테이너
DataSetExtension	데이터셋 레벨 메타데이터에 데이터를 추가할 수 있는 컨테이너 요소

DataSet은 해당 데이터를 직접 기술하기 위한 요소로서 DataSet의 하위 요소에는 GUID, ID, ResourceURI, TechnicalContact, ContentContact, DataCenter, OtherProvider, Metadata, Unit 그리고 DataSetExtensin 등이 포함되어 있다.

2. DarwinCore

DarwinCore는 Z39.50 Biology Implementer Group(ZBIG)이 미국 NSF에서 연구비를 지원받아 1998년에 처음 만들어졌으며, 2009년 10월 9일에 메타데이터 공식 표준으로 발표되었다. 기본적으로 DarwinCore는 자연에서 수집하는 표본을 바탕으로 하며, 그 외에도 표본 관련 자료 및 관찰 자료도 포함하고 있다. 이 표준에는 이러한 용어의 관리 방법, 새로운 목적을 위해 용어의 집합을 확장할 수 있는 방법, 용어의 사용 방법을 설명하는 문서가 함께 들어있어 이를 바탕으로 사용할 수 있다. DarwinCore는 단순 DarwinCore와 일반 DarwinCore로 나뉘는데, 단순 DarwinCore는 표본과 그 수집상황에 관한 데이터를 간단한 구조 방식으로 공유하기 위한 하나의 특정한 방법에 대한 규격이다. 이름 그대로 단순성과 유연성을 특징으로 가진다. 단순 DarwinCore는 속성과 값, 필드 및 레코드로 사용되는 행과 열의 개념을 넘어서 구조가 제한되지 않으며, 용어와 필드명이 동일하여 단순한 특징을 가지고 있다. 또한, 단순 DarwinCore에는 필요한 최소한의 필드에 대한 제한이 없다. 필수 필드 제한이 없어 단순 DarwinCore를 사용하여 상황에 적절한 필드들의 조합을 만들고 공유할 수 있다. 이러한 유연성은 다양한 서비스에 대한 용어 및 공유 메커니즘의 재사용을 촉진한다(TDWG 2009).

DarwinCore는 식별자, 라벨 및 정의 그리고 예제 및 해설을 제공하여 생물학적 다양성에 대한 정보의 공유를 촉진하기 위한 용어집(다른 맥락에서 속성, 요소, 필드, 열, 속성 또는 개념)이다. 더불어 DarwinCore는 평범한 정보 공유를 넘어 생물학적 다양성을 보장하는 정보 공유를 위해 안정적이고 표준적인 참조를 제공하는 것을 목표로 한다. 용어집으로서 기능하는 DarwinCore는 다양한 맥락에서 최대한으로 재사용할 수 있도록 기능하고 있으며, 명확한 시맨틱 정의를 제공한다. 이는 DarwinCore를 이전에 사용했던 것과 같이 사용할 수 있을 뿐더러, 용어집의 공통된 어휘를 통해 상호운용성을 보장하며, 보다 세부적이고 복잡한 교환 포맷을 구축하는 기초가 될 수도 있다. 이러한 DarwinCore는 현재 GBIF 네트워크[2] 안에서 대부분의 표본 및 관측 자료 메타데이터를 만드는 데 사용되고 있다. 2012년 기준 43개국 340개 이상이 사용하고 있으며, 자연사적인 수집을 넘어, 관련 커뮤니티까지 확장되고 있다(Wieczorek et al. 2012).

다음의 〈표 6〉은 DarwinCore의 최상위 클래스를 나타낸 것이다.

2) GBIF 네트워크: 정부 간 협력체계이며, 생물 다양성 데이터에 무료로 접근을 도와주는 국가, 경제 및 국제기구이다. 2019년에는 97개국이 참여하고 있다.

〈표 6〉 DarwinCore 최상위 클래스

클래스	설명	비고
레코드 레벨 용어(Record-level Terms)	Dublin Core 용어, 기관, 컬렉션, 데이터 레코드의 성질	단순 다윈 코어
종(Occurrence)	자연에 있는 동식물종의 증거, 관찰자, 행동, 관련 매체, 참조	
유기체(Organism)	분류학적으로 같다고 간주되는 특정 유기체 또는 정의된 유기체 그룹을 기술하는 요소	
재료 샘플(MaterialSample)	샘플링 작업의 결과물 및 재료 표본	
이벤트(Event)	샘플 수집 프로토콜 및 날짜, 시간, 필드 노트	
위치(Location)	지질학적, 지역성 설명, 공간 데이터	
식별자(Identification)	표본과 종(Occurrence)의 관련성	
표본(Taxon)	학명, 자국어 명칭, 명칭 용법, 표본 개념 등 사이의 관계	
지질학적 맥락(GeologicalContext)	지질학적, 연대적 생물층서학적, 지층학적 시간	
리소스 관계(ResourceRelationship)	식별된 리소스 간의 명시적 관계	유전 다윈 코어
측정 및 사실(MeasurementOrFact)	측정, 사실, 특성, 주장, 참조	

DarwinCore의 상위 클래스에는 Record-level Terms, Occurrence, Organism, MaterialSample, Event, Location, Identification, Taxon, GeologicalContext, ResourceRelationship 그리고 MeasurementOrFact가 포함된다. 이 중 Record-level Terms, Occurrence, Organism, MaterialSample, Event, Location, Identification, 그리고 Taxon 총 7개는 단순 DarwinCore에 포함되고, GeologicalContext와 ResourceRelationship는 유전 DarwinCore에 추가로 사용되는 클래스이다. 다음의 〈표 7〉은 DarwinCore의 상위 클래스 중 하나인 Record-level Terms의 하위 요소를 나타낸 것이다.

〈표 7〉 Record-level Terms의 하위 요소

하위요소	설명
type	리소스의 특성 및 장르
modified	가장 최근에 수정된 날짜
language	리소스에 쓰인 언어
license	생물자원의 이용을 허가하는 공식적인 법적 문서
accessRights	리소스에 접근할 수 있는 사용자 또는 보안 상태에 관한 정보
bibliographicCitation	이 레코드를 인용하는 방법
references	이 리소스가 참조하거나 인용한 관련된 다른 리소스
institutionID	이 레코드에서 언급된 객체 또는 정보를 갖고 있는 기관의 식별자
collectionID	이 레코드가 속해있는 컬렉션 또는 데이터셋의 식별자
datasetID	데이터셋의 식별자, 기관에서 사용하는 특정한 식별자도 될 수 있다.
institutionCode	레코드에서 언급된 객체 또는 정보를 갖고 있는 기관의 이름
collectionCode	이 레코드가 속해있는 컬렉션 또는 데이터셋의 이름, 약어, 이니셜
datasetName	레코드가 속해 있는 데이터셋의 이름
ownerInstitutionCode	레코드에서 언급된 객체 또는 정보의 소유권을 가진 기관의 이름(또는 약어)
basisOfRecord	데이터 레코드의 특성
informationWithheld	추가 정보 기술
dataGeneralizations	공유한 데이터를 원래 형식보다 구체적으로 만들지 않게 하기 위한 조치
dynamicProperties	레코드에 관한 추가적인 측정자료, 특성 또는 관련 주장

Record-level Terms 클래스는 실제 레코드 단위의 데이터를 기술하기 위한 클래스로서 하위 요소에는 Dublin Core의 요소를 포함하여 기관, 컬렉션 및 데이터 레코드의 성질을 기술하는 클래스로, 총 19개의 하위 요소를 포함하고 있다.

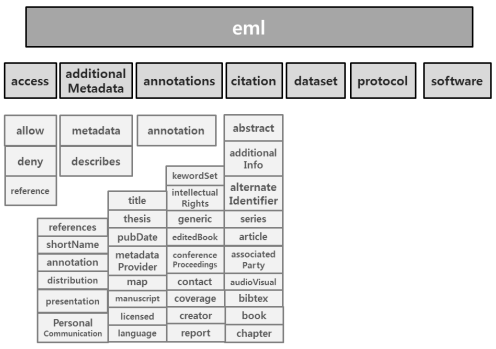
3. Ecological Metadata Language (EML)

EML은 Knowledge Network for Biocomplexity(KNB)에서 구축한 XML 기반의 메타데이터 표준으로, 미국 생태 학회가 수행한 작업을 바탕으로 생태 분야의 연구데이터를 다루고 공유하기 위해 개발되었다. EML은 NCEAS[3] 소속의 Matthew Jones의 「미래 장기적 생태학 데이터에 대한 ESA 위원회」라는 보고서와 Michener 등이 저술한 「생태학을 위한 범지역적 메타데이터」를 기반으로 개발되었다. Version 1.0은 1997년 NCEAS에 의해 발표되었으며, 콘텐츠 구현에 FLED 권장 사항을 엄격히 준수한 Version 1.2, 1.3, 그리고 1.4가 발표되었다. 이후의 Version 2는 단체에서 유지, 관리하는 공개 사양이 되었다. EML 2.0의 핵심적인 수정 및 변경은 NCEAS의 이전 사양을 통한 경험과 생태학 커뮤니티(특히 장기 생태 연구 네트워크)의 정보 관리자와의 피드백을 통해 이루어졌고, Version 2.1과 2.2는 글로벌, 시멘틱 주석과 데이터 논문 지원 등 중요하면서도 새로운 특징을 포함하고 있다. 이러한 EML은 데이터 패키지를 식별하고 인용하는 모듈, 데이터의 시공간적, 분류학적, 주제별 범위를 기록하는 모듈, 연구 방법 및 프로토콜을 기술하는 모듈, 복잡한 데이터 패키지 내의 구조 및 내용을 해명하는 모듈 그리고 시멘틱 어휘로 데이터에 정확한 주석을 달기 위한 모듈을 가진다. 또한 EML은 생태, 지구 및 환경과학 등의 관찰 연구 관련 데이터를 기록하기 위한 메타데이터 규격(사양)을 오픈소스로 제공하고 있으며, 커뮤니티 중심으로 운영되고 있다(The Ecological Society of America 2013; William 1997).

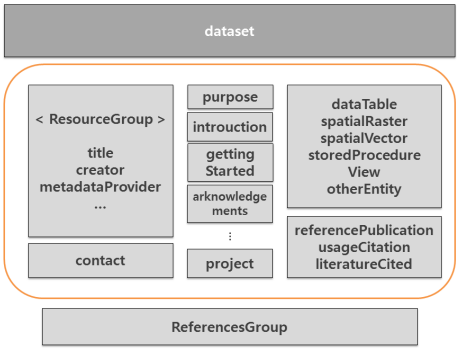
다음의 〈그림 3〉과 〈그림 4〉는 EML의 최상위 모듈과 최상위 모듈 중의 하나인 dataset의 하위요소를 나타낸 것이다(Jones et al. 2019).

EML의 최상위 모듈은 The EML Module이며, 하위 요소로 access, additionalMetadata, annotations, citation, dataset, protocol, 그리고 software가 있다. eml-dataset 모듈은 데이터셋 정보의 개요, 즉 데이터 자체의 title, abstract, keywords, contacts, maintenance history, purpose, 그리고 distribution 등의 광범위한 정보를 제공한다. 또한 이 모듈은 데이터셋을 자세하게 설명하는 다른 많은 모듈을 가져온다. 특히 데이터셋 수집과 처리에 사용되는 방법론을 기술하는 eml-methods 모듈, 많은 것과 관련되는 연구 맥락 정보와 실험 설계 정보를 기술하는 eml-project 모듈, 데이터와 메타데이터의 접근 제어 규칙을 정의하는 eml-access 모듈, 데이터셋 논리 구조의 상세한 정보를

3) 국립 생태 분석 및 합성 센터(National Center for Ecological Analysis and Synthesis)의 약어로서 1995년에 설립되었다.



〈그림 3〉 eml 최상위 모듈



〈그림 4〉 데이터셋의 하위 요소

제공하는 eml-entity 모듈이 있다. 데이터셋은 특정한 무결성 제약 조건으로 연결된 일련의 데이터 엔티티로 구성된다. 다음의 〈표 8〉은 ResourceGroup의 하위 요소를 나타낸 것이다.

〈표 8〉 ResourceGroup 모듈의 하위 요소

요소명	의미
초록(abstract)	자원의 간략한 개요
추가정보(additionalInfo)	자원의 다른 추가 정보
대체식별자(alternateIdentifier)	엔티티의 두번째 식별자
주석(annotation)	자원의 시멘틱 선언문
관련자(associatedParty)	자원과 연관된 다른 사람/조직
범위(coverage)	자원의 범위
생산자(creator)	자원을 작성한 사람/단체
배포(distribution)	자원의 온/오프라인 배포 방법
지적재산권(intellectualRights)	자원의 사용 및 라이선스 관련 정보
키워드셋(keywordSet)	기술된 자원의 키워드 정보
언어(language)	자원이 작성된 언어
라이센스(licensed)	메타데이터/데이터의 잘 알려진 라이선스를 식별하는 정보
메타데이터 제공자(metadataProvider)	자료의 문서/메타데이터 제공자
발행일(pubDate)	자원을 발행한 일자
시리즈(series)	자원의 연작(시리즈)
약어(shortName)	기술된 자원의 약칭, 때때로 파일명
제목(title)	타 자원과 구별되는 자원의 약술

대상 자원을 기술하는 일반적인 설명을 나타내는 〈ResourceGroup〉은 abstract, additionalInfo, alternateIdentifier, annotation, associatedParty, coverage, creator, distribution, intellectualRights, keywordSet, language, licensed, metadataProvider, pubDate, series, shortName, 그리고 title 요소를 포함한다.

4. DataCite 4.3

DataCite는 2009년에 설립된 범용적 연구데이터를 학문적으로 커뮤니케이션하기 위해 합법적으로 기여하는 국제적인 컨소시움이자, 표준 메타데이터이다. DataCite는 연구데이터셋에 대한 영구 식별자(DOI)를 할당하고 데이터 인용뿐만 아니라 검색 및 접근에 대한 간단하며 대단히 효과적인 방법을 지원하는 환경을 제공한다. DataCite는 구축한 DOI 인프라를 이용자에게 제공하는데, 이 DOI 명칭은 과학 저널 및 기사에서 널리 사용되는 식별자로서, 대부분의 이용자(연구자, 출판사 등)에게 익숙하다. DataCite는 개방형 접근방식을 채택하고 있으며, 그 목적을 위해 식별자 시스템과 서비스를 지속적으로 개발하는 비영리 단체이다. DataCite의 기능적인 목표는 다음과 같다(Max 2010).

- 인터넷 환경에서 과학 연구데이터에 보다 쉽고 간단하게 접근
- 과학 레코드에 대한 합법적이고 인용 가능한 기여로 연구데이터의 수용성 증가
- 향후 연구를 위해 결과 검증 및 용도 변경이 가능한 데이터 아카이빙 지원

다음의 〈표 9〉는 DataCite 4.3의 최상위 요소에 대한 정의와 적용방식을 나타낸 것이다.

〈표 9〉 DataCite 메타데이터 스키마 4.3의 최상위 요소

요소	정의	적용방식
Identifier	자원을 식별하는 고유한 식별자	M
Creator	데이터 생산 또는 출판에 기여한 주요 연구자	M
Title	자원이 알려진 이름 또는 제목 (연도 4자리 기술)	M
Publisher	자원을 보유, 보관, 게시, 인쇄, 배포, 공개, 발행 또는 생성하는 엔티티의 이름	M
PublicationYear	데이터가 공개되었거나 공개 될 연도	M
Subject	자원을 설명하는 주제, 키워드, 분류 코드 또는 주요 문구	R
Contributor	자원의 수집, 관리, 배포 또는 기타 기여에 책임이 있는 개인 또는 기관	R
Date	업무와 관련된 다른 날짜 (YYYY를 기본으로 상세 기술)	R
Language	정보자원의 기본 언어	O
ResourceType	정보자원에 대한 기술 및 설명	M
AlternateIdentifier	등록 중인 자원에 적용된 기본 식별자 이외의 식별자	O
RelatedIdentifier	관련된 자원의 식별자	R
Size	크기 (예: 바이트, 페이지, 인치 등) 또는 기간 (범위) 예: 정보자원의 시간, 분, 일 등. “45 minutes”	O
Format	자원의 기술적 포맷(구성 방식)	O
Version	자원의 버전 번호	O
Rights	자원에 대한 모든 권리 정보.	O
Description	다른 범주에 맞지 않는 모든 추가 정보 (기술적 정보)	R
GeoLocation	데이터가 수집되었거나 데이터가 초점을 둔 공간 지역 또는 명명된 장소	R
FundingReference	등록된 자원의 재정 지원 (펀딩)에 대한 정보	O

적용방식으로 요소를 구분한다면, 'Identifier', 'Creator', 'Title', 'Publisher', 'PublicationYear' 그리고 'ResourceType' 등 6개는 필수(Mandatory)에 'Subject', 'Contributor', 'Date', 'RelatedIdentifier', 'Description' 그리고 'GeoLocation' 등 6개는 권고(Rcommended)에 'Language', 'AlternateIdentifier', 'Size', 'Format', 'Version', 'Rights' 그리고 'FundingReference' 등 7개는 선택(Optional) 요소에 속한다.

Ⅳ. 생태분야 연구데이터 관리를 위한 메타데이터 요소 선정

이번 장에서는 3장에서 분석한 메타데이터 표준을 기준으로 생태분야를 위한 메타데이터 항목을 도출하였다.

1. 생태분야를 위한 메타데이터 요소 도출

생태분야의 메타데이터 요소를 도출하기 위하여 2장과 3장을 통해 분석한 DCAT, ABCD, DarwinCore, EML 마지막으로 DataCite 4.3을 대상으로 크로스워크 작업을 진행하였다. 다음의 〈표 10〉은 생태분야 메타데이터 요소를 도출하기 위해 사용된 표준을 정리한 것이다.

〈표 10〉 생태분야 메다데이터 요소 도출을 위한 표준

메타데이터	내용
DCAT V2	공개되는 데이터를 검색 활용하기 위해 개방 유통되는 데이터에 대한 정보를 서술하기 위한 어휘 등을 규격화한 기술표준(W3C 표준)
DataCite Schema 4.3	범용적 연구 데이터 기술을 위한 표준 메타데이터(국외 사실적 표준)
ABCD	자연 생태 분야 메타데이터 표준
Darwin Core	
EML	

이번 연구에서는 생태분야의 연구데이터를 기술하기 위한 메타데이터 요소를 도출하는 것이 목적이므로 본 연구 목적에 맞게 데이터 자체를 기술하기 위한 메타데이터 요소를 매핑 대상으로 삼게 되었다. 즉, DCAT은 Resource, Dataset, Distribution 그리고 Dataservice 클래스의 하위 요소를 DataCite는 최상위요소를 ABCD의 경우, Dataset의 하위요소를 DarwinCore는 Record-level Terms의 하위 요소를 마지막으로 EML은 ResourceGroup 모듈의 하위 요소를 매핑 대상으로 하였다.[4] 다음의 〈표 11〉은 DCAT을 기준으로 4개의 메타데이터를 매핑한 것이다.

4) ABCD 매핑 사례를 보면, OECD Minimum Data Set과 ABCD를 매핑하였는데 여기에서 ABCD 전체 요소를

〈표 11〉 생태 분야 메타데이터 요소 도출을 위한 크로스워크

DCAT V2		DataCite	EML	Darwin Core	ABCD
Resource	access rights			accessRights	
	conforms to				
	contact point				
	resource creator	Creator	creator		contentcontact
	description	Description	abstract/additionalInfo/ annotation	informationWithheld / dynamicProperties	description / InformationWithheld
	title	Title	title	DatasetName	
	release date	PublicationYear	pubDate		
	update/modification date			modified	RevisionData
	language	Language	language	language	
	publisher	Publisher			
	identifier	Identifier		identifier Identification	ID
	theme/category	Subject			
	type/genre	ResourceType		type	Unit
	resource relation	RelatedIdentifier		ResourceRelationship	
	qualified relation				
	keyword/tag		keywordSet		Scope
	landing page				
	qualified attribution				DataCenter / Otherprovider / Owner / TechnicalContent
	license		licensed	license	
	rights	Rights	intellectualRights		LegalStatement
	has policy				
	is referenced by			references	
Dataset	dataset distribution				
	frequency				
	spatial/geographical coverage	GeoLocation	coverage	dataGeneralizations / location / Event	
	spatial resolution				
	temporal coverage	Date		GeologicalContext	
	temporal resolution				
	was generated by				
Distribution	access URL		distribution		
	access service				
	download URL				DirectAccessURL
	byte size	Size			
	conforms to				
	media type				
	format	Format			
	compression format				
	packaging format				

대상으로 하지 않고 Dataset의 하위 항목을 대상으로 매핑을 시도하였다.
〈https://www.bgbm.org/tdwg/codata/Schema/ABCD_1.30/OECD-MDS-2-ABCD130.pdf〉

DCAT V2		DataCite	EML	Darwin Core	ABCD
Data Service	endpoint URL				
	endpoint description				
	serves dataset				
		AlternateIdentifier	alternateIdentifier		GUID / ResourceURI
			metadataProvider		
			series		
			shortName		
		contributor	associatedParty*		
		Version			Version
		FundingReference			
				basisOfRecord	
				MeasurementOrFact	
				bibliographicCitation	
					LogoURL

크로스워크를 진행한 결과 DCAT을 기준으로 5개 메타데이터에 모두 포함되는 요소는 'description' 1개이고, 4개 메타데이터에 포함되는 요소는 'resource creator', 'title', 'language', 'identifier', 'type/genre' 그리고 'rights' 등 6개이다. 3개 메타데이터에 포함하는 요소는 'update/modification date', 'resource relation', 'keyword/tag' 그리고 'license' 등 4개이다. 다음으로 2개 메타데이터에 포함되는 요소는 'access rights', 'qualified attribution', 'is referenced by', 'download URL', 'byte size' 그리고 'format' 등 6개이다. 마지막으로 1대 1 매핑은 아니지만[5] 4개의 메타데이터에 포함되어 있는 요소는 'spatial/geographical coverage', 'spatial resolution', 'temporal coverage' 그리고 'temporal resolution' 등 위치정보를 기술하기 위한 항목이다.

다음의 〈표 12〉는 크로스워크를 통해 도출된 생태분야 메타데이터 요소와 해당 의미를 정리한 것이다. 51개 요소 중 DataCite에서 가져온 요소는 'Alternate Identifier', 'Contributor', 'Version' 그리고 'FundingReference' 등 4개이다. 이 중 'Contributor'는 EML의 'associatedParty'와 동일한 의미를 지니고 있어 EML의 'associatedParty'로 대체하였다. 다음 EML 출처를 가진 요소는 'metadata Provider', 'associated Party', 'short Name' 그리고 'series' 등 4개이다. 다음으로 Darwin Core 메타데이터에서 가져온 요소는 'basis Of Record', 'Measurement Or Fact' 그리고 'bibliographic Citation' 등 3개요소이다. 마지막으로 ABCD에서는 'GUID / ResourceURI', 'Version' 그리고 'LogoURL' 등 3개요소를 가져왔으나 'GUID / ResourceURI'는 'Alternate Identifier'와 동일한 요소라 해당 요소로 대체하였으며, 'LogoURL'은 데이터를 기술하는 요소로 적당치 않아 생태분야의 메타데이터 요소에서 제외하기로 하였다.

5) DCAT의 'spatial/geographical coverage'과 'spatial resolution'은 DataCIte의 최상위요소인 'GeoLocation'과 매핑된다.

〈표 12〉 생태정보 메타데이터 요소

ID	요소명(영/한)		출처
1	title	이름, 제목	DCAT
2	release date	발행일	DCAT
3	publisher	출판사	DCAT
4	license	라이선스	DCAT
5	type/genre	유형 / 장르	DCAT
6	Measurement Or Fact	측정 또는 사실	Darwin Core
7	update/modification date	갱신 / 수정일	DCAT
8	temporal coverage	시간 범위	DCAT
9	spatial/geographical coverage	공간/지리적 범위	DCAT
10	description	설명	DCAT
11	spatial resolution	공간 해상도	DCAT
12	temporal resolution	시간 해상도	DCAT
13	basis Of Record	데이터셋 종류 및 특성	Darwin Core
14	keyword/tag	키워드 / 태그	DCAT
15	identifier	식별자	DCAT
16	relationship	관련 식별자	DCAT
17	resource creator	생성자	DCAT
18	theme/category	카테고리 / 주제영역	DCAT
19	access rights	접근 권한	DCAT
20	download URL	다운로드 URL	DCAT
21	conforms to	표준	DCAT
22	metadata Provider	메타데이터 제공자	EML
23	access URL	접속URL	DCAT
24	media type	매체 유형	DCAT
25	bibliographic Citation	레코드 인용 방법	Darwin Core
26	format	포맷	DCAT
27	contact point	연락처	DCAT
28	was generated by	활동명	DCAT
29	associated Party	관련자	DataCite(contributor) / EML
30	rights	권한	DCAT
31	language	언어	DCAT
32	access service	접속 서비스	DCAT
33	compression format	압축 포맷	DCAT
34	short Name	약칭 / 약어	EML
35	byte size	바이트 크기	DCAT
36	frequency	빈도	DCAT
37	Funding Reference	펀딩 참조 기관	DateCIte
38	series	연작	EML
39	dataset distribution	데이터셋 배포	DCAT
40	Version	버전	DataCite / ABCD
41	packaging format	패키지 포맷	DCAT
42	endpoint URL	엔드포인트 URL	DCAT
43	has policy	ODRL 준수 정책	DCAT
44	landing page	랜딩 페이지	DCAT
45	is referenced by	참조 자원	DCAT
46	qualified attribution	한정 속성	DCAT
47	alternateIdentifier	대체 식별자	DataCIte / EML
48	endpoint description	엔드포인트 설명	DCAT
49	serves dataset	데이터 컬렉션	DCAT
50	qualified relation	한정 관계	DCAT
51	resource relation	자원 관계	DCAT

2. 생태분야를 위한 메타데이터 요소 검증

이번 절에서는 이전 절에서 도출된 메타데이터 요소 검증과 요소의 적용방식을 결정하기 위해 메타데이터 전문가를 대상으로 인터뷰를 실시하였고 또한 생태 연구기관인 국립생태원의 연구자를 대상으로 설문조사를 실시하였다. 다음의 〈표 13〉은 인터뷰와 설문조사에 대한 개요를 나타낸 것이다.

〈표 13〉 인터뷰 및 설문조사 개요

검증 방법	개요
설문조사	• 조사 목적: 생태분야 메타데이터의 적용방식 결정을 위한 설문조사 • 조사 대상: 국립생태원 연구자 (32명) • 조사 방법: 웹 설문조사 • 조사 기간: 2020년 01월 09일 ~ 2020년 01월 15일 • 응답률: 100% • 설문지 응답 URL: http://ksurv.kr/?a=29083
인터뷰	• 인터뷰 목적: 생태분야 주요 메타데이터 스키마 검증 • 인터뷰 대상: TTA PG 606 위원 (3명) • 인터뷰 방법: 서면과 대면 인터뷰 • 인터뷰 일시: 2020년 6월 15일 / 2020년 6월 30일

먼저, 설문조사는 국내 생태분야 연구기관인 국립생태원의 연구자를 대상으로 온라인 설문조사를 실시하였다. 설문조사의 목적은 기 도출된 메타데이터 항목을 대상으로 적용방식을 결정하기 위함이다. 설문조사 실시하기 전 생태분야의 전문가인 국립생태원 연구자를 대상으로 도출된 생태분야 메타데이터에 대한 교육을 실시하였다. 다음으로 도출된 생태분야 메타데이터 스키마를 한국정보통신협회(TTA) 산하 위원회인 PG 606 위원을 대상으로 서면과 대면 인터뷰를 2차례 실시하였다. 1차 인터뷰는 메타데이터 스키마를 메일로 보낸 후 이메일로 수정사항을 전달받았다. 2차 인터뷰는 대면 인터뷰로 1시간 동안 메타데이터 전문가와 면담을 진행하였다. 다음의 〈표 14〉는 설문 응답을 바탕으로 메타데이터 요소에 대하여 점수를 부여하여 순위화한 것이다. 점수는 5점 척도를 사용하여 매우 불필요요소 (1점), 불필요요소 (2점), 선택요소 (3점), 필수요소 (4점), 매우 필수요소 (5점) 등으로 가중치를 적용하였다.

〈표 14〉 가중치를 적용한 메타데이터 항목

No.	요소명	매우 불필요 요소	불필요 요소	선택 요소	필수 요소	매우 필수 요소	점수 합계
1	title(이름, 제목)	0	0	1	13	18	145
2	rights(권한)	0	0	0	19	13	141
3	update/modification date(갱신/수정일)	0	0	3	13	16	141
4	release date(발행일)	0	0	4	14	14	138

No.	요소명	매우 불필요 요소	불필요 요소	선택 요소	필수 요소	매우 필수 요소	점수 합계
5	temporal coverage(시간 범위)	0	1	4	12	15	137
6	spatial / geographical coverage (공간/지리적 범위)	0	2	3	13	14	135
7	description(설명)	0	1	2	21	8	132
8	spatial resolution(공간 해상도)	0	3	2	15	12	132
9	publisher(출판사)	0	1	5	16	10	131
10	temporal resolution(시간 해상도)	0	3	4	13	12	130
11	basisOfRecord(데이터셋 종류 및 특성)	0	0	8	15	9	129
12	keyword/tag(키워드/태그)	0	1	5	18	8	129
13	identifier(식별자)	0	2	4	18	8	128
14	license(라이선스)	0	3	4	15	10	128
15	resource creator(생성자)	0	1	5	19	7	128
16	theme/category(카테고리/주제영역)	0	0	8	16	8	128
17	access rights(접근 권한)	0	4	3	15	10	127
18	downloadURL(다운로드 URL)	0	2	5	17	8	127
19	MeasurementOrFact(측정 또는 사실)	0	1	8	15	8	126
20	conforms to(표준)	1	1	7	13	10	126
21	language(언어)	0	3	6	14	9	125
22	metadataProvider(메타데이터 제공자)	0	0	10	16	6	124
23	accessURL(접속URL)	0	2	9	13	8	123
24	media type(매체 유형)	0	2	8	15	7	123
25	bibliographicCitation(레코드 인용 방법)	0	2	6	20	4	122
26	type/genre(유형)	0	0	13	12	7	122
27	format(포맷)	0	4	6	15	7	121
28	contact point(연락처)	0	3	10	12	7	119
29	was generated by(활동명)	0	2	9	18	3	118
30	associatedParty(관련자)	0	1	11	18	2	117
31	access service(접속 서비스)	0	0	13	18	1	116
32	compression format(압축 포맷)	1	3	11	9	8	116
33	shortName(약칭/약어)	0	3	11	14	4	115
34	byte size(바이트 크기)	1	4	8	14	5	114
35	frequency(빈도)	0	3	12	13	4	114
36	FundingReference(펀딩 참조 기관)	0	5	9	14	4	113
37	series(연작)	0	3	12	15	2	112
38	dataset distribution(데이터셋 배포)	0	4	12	13	3	111
39	Version(버전)	0	3	12	16	1	111
40	packaging format(패키지 포맷)	1	4	13	9	5	109
41	endpointURL(엔드포인트URL)	0	6	12	10	4	108
42	has policy(ODRL 준수 정책)	0	4	13	14	1	108
43	landing page(랜딩 페이지)	0	3	15	13	1	108
44	is referenced by(참조 자원)	0	3	15	14	0	107
45	RelatedIdentifier(관련 식별자)	1	2	17	11	1	105
46	qualified attribution(한정 속성)	0	3	20	7	2	104
47	alternateIdentifier(대체 식별자)	1	4	16	11	0	101
48	endpoint description(엔드포인트 설명)	0	7	16	6	3	101
49	serves dataset(데이터 컬렉션)	0	4	20	8	0	100
50	qualified relation(한정 관계)	0	5	19	8	0	99
51	resource relation(자원 관계)	0	7	16	8	1	99

위의 〈표 14〉와 같이 가장 많은 점수가 부여된 요소는 'title'이다. 뒤이어 'rights', 'update/modification date' 등 51개의 요소가 순서대로 나열되었다. 다음으로 메타데이터 전문가를 대상으로 한 인터뷰를 통하여 일부 스키마에 대한 수정사항을 반영하였다. 전문가 의견을 토대로 수정된 스키마는 부록을 통하여 기술하였다.

3. 생태분야를 위한 메타데이터 요소 재선정

이번 절에서는 이전 절에서 수행된 설문조사를 바탕으로 메타데이터 요소 적용방식을 선정 기준에 따라 해당 요소를 필수, 권고 그리고 선택항목으로 결정하였다. 적용방식 선정 기준은 다음과 같다.

- 기준 1: DataCite Schema 4.3에서 제시한 필수 요소, 권고 요소 및 선택요소는 본 연구에서 동일하게 필수 요소, 권고 요소 그리고 선택요소로 한다.
- 기준 2: 기준 1의 R에서 최소 점수와 동일하거나 높은 경우 해당 항목을 권고 요소로 한다.
  - 단, 이미 기준1에서 제시한 요소의 적용수준은 변하지 않는다.
  - 기준 1에서 제시한 R의 최소 점수는 '105'점으로 'RelatedIdentifier'이지만 본 기준에서는 이보다 높은 점수가 부여된 'associatedParty'(117점)를 기준항목으로 정하였다.
- 기준 3: 그 외의 요소는 선택 요소로 규정함

위의 적용방식 선정 기준을 적영한 결과 다음의 〈표 15〉와 같이 도출되었다. 다음의 〈표 15〉는 요소명을 기준으로 해당 요소명의 정의, 적용방식 그리고 출처를 나타낸 것이다.

〈표 15〉 생태분야 메타데이터 요소

ID	요소명(영/한)		정의	필수 여부	출처
1	title	이름, 제목	리소스의 이름 또는 명칭	M	DCAT
2	release date	발행일	데이터셋 발행일	M	DCAT
3	publisher	출판사	데이터셋을 웹에 공개 또는 출판하는 주체	M	DCAT
4	license	라이선스	데이터셋에 제공될 수 있는 라이선스와 권리 정보	M	DCAT
5	type/genre	유형 / 장르	데이터셋의 유형	M	DCAT
6	Measurement Or Fact	측정 또는 사실	자원을 식별할 수 있는 측정치	M	Darwin Core
7	update/modification date	갱신 / 수정일	카탈로그 항목이 변경, 갱신, 수정된 가장 최근 날짜	R	DCAT
8	temporal coverage	시간 범위	데이터셋이 적용되는 기간(시간, 날짜 등)	R	DCAT
9	spatial/geographical coverage	공간/지리적 범위	데이터셋이 적용되는 지리적 영역	R	DCAT
10	description	설명	데이터셋의 내용에 대한 설명	R	DCAT
11	spatial resolution	공간 해상도	데이터셋에 적용된 공간해상도	R	DCAT
12	temporal resolution	시간 해상도	데이터셋에서 확인할 수 있는 최소 기간(시간)	R	DCAT

ID	요소명(영/한)		정의	필수 여부	출처
13	basis Of Record	데이터셋 종류 및 특성	데이터 레코드의 근거	R	Darwin Core
14	keyword/tag	키워드 / 태그	데이터셋을 설명하는 키워드/태그	R	DCAT
15	identifier	식별자	데이터셋의 식별자 또는 식별값	R	DCAT
16	relationship	관련 식별자	관련 자원의 국제적 고유 식별자	R	DCAT
17	resource creator	생성자	데이터셋을 생성 또는 묶은 사람, 조직 등을 기술함	R	DCAT
18	theme/category	카테고리 / 주제영역	데이터셋의 내용이 지닌 주제 정보를 기술함	R	DCAT
19	access rights	접근 권한	데이터셋의 접근제한 또는 이용제한 유형	R	DCAT
20	download URL	다운로드 URL	지정된 형식의 다운로드 가능한 파일의 URL	R	DCAT
21	conforms to	표준	기술된 자원이 준수하는 표준	R	DCAT
22	metadata Provider	메타데이터 제공자	데이터셋의 설명 및 기타 메타데이터를 제공한 사람	R	EML
23	access URL	접속URL	데이터셋의 배포에 접근할 수 있는 리소스의 URL	R	DCAT
24	media type	매체 유형	IANA에서 정의한 배포의 매체 유형	R	DCAT
25	bibliographic Citation	레코드 인용 방법	레코드를 인용하는 방법	R	Darwin Core
26	format	포맷	IANA에서 정의하지 않은 배포의 파일 형식	R	DCAT
27	contact point	연락처	데이터셋에 대한 문의 사항에 대응할 수 있는 담당자의 연락처 정보	R	DCAT
28	was generated by	활동명	데이터셋 생성을 위한 비즈니스 컨텍스트를 생성하거나 제공하는 활동	R	DCAT
29	associated Party	관련자	자원과 연관되어야 하는 다른 개인/단체/직위의 전체 이름	R	EML
30	rights	권한	리소스의 저작권 정보	O	DCAT
31	language	언어	정보자원의 기본 언어	O	DCAT
32	access service	접속 서비스	데이터셋의 배포에 접근할 수 있는 데이터 서비스	O	DCAT
33	compression format	압축 포맷	다운로드 파일의 크기를 줄이기 위해 데이터가 압축된 형태로 되어 있는 배포의 압축 포맷	O	DCAT
34	short Name	약칭 / 약어	기술된 자원의 약칭 또는 지칭명	O	EML
35	byte size	바이트 크기	데이터셋 배포의 크기	O	DCAT
36	frequency	빈도	데이터셋이 출판되는 빈도	O	DCAT
37	Funding Reference	펀딩 참조 기관	연구비 정보로서 연구비 지원기관 정보, 연구비, 프로젝트명, 프로젝트 코드 등을 기술	O	DataCite
38	series	연작	대상 자원의 연속된 데이터셋에 대한 정보	O	EML
39	dataset distribution	데이터셋 배포	데이터셋의 사용가능한 배포, 구체적인 표현	O	DCAT
40	Version	버전	데이터셋의 버전 번호	O	DataCite
41	packaging format	패키지 포맷	하나 이상의 데이터 파일이 함께 그룹화 되어있는 패키지 포맷	O	DCAT
42	endpoint URL	엔드포인트 URL	서비스의 루트 위치 또는 기본 엔드 포인트	O	DCAT
43	has policy	ODRL 준수 정책	자원과 관련된 권한을 나타내는 ODRL 준수 정책	O	DCAT
44	landing page	랜딩 페이지	목록(카탈로그), 데이터셋, 배포 및 추가정보로 접근할 수 있는 웹페이지 브라우저	O	DCAT
45	is referenced by	참조 자원	참조, 인용 또는 카탈로그화된 데이터셋을 가리키는 관련 리소스	O	DCAT
46	qualified attribution	한정 속성	데이터셋에 대해 어떠한 책임이 있는 에이전트	O	DCAT
47	alternateIdentifier	대체 식별자	데이터에 접근하는 또 다른 식별자 또는 해당 엔티티의 추가적인 식별자	O	EML
48	endpoint description	엔드포인트 설명	엔드 포인트를 통해 사용 가능한 서비스에 대한 기계가 읽을 수 있는 설명	O	DCAT
49	serves dataset	데이터 컬렉션	이 데이터 서비스가 배포하는 데이터셋	O	DCAT
50	qualified relation	한정 관계	다른 데이터셋과의 관계를 설명한 링크	O	DCAT
51	resource relation	자원 관계	목록화된 항목과 불특정한 관계가 있는 자원	O	DCAT

위의 〈표 15〉에서 나타난 바와 같이 생태 분야 연구데이터를 관리하기 위한 메타데이터 요소는 전체 51개 중 필수 요소는 6개, 권고 요소 23개 그리고 선택요소 22개로 요소가 도출되었다.

Ⅴ. 결 론

본 연구의 목적은 생태분야 연구데이터를 관리 및 공유를 위한 메타데이터를 설계하기 위함이다. 해당 메타데이터는 웹에서 발행된 데이터 카탈로그들 간 상호운용성 향상을 위해 설계된 RDF 어휘인 DCAT을 기반으로 설계되었다. 또한, 생태 분야의 특성을 반영하기 위하여 ABCD, Darwin Core 그리고 EML 메타데이터를 분석하였다. 여기에 범용적인 연구데이터를 기술하기 위한 표준 메타데이터인 DataCite 4.3을 추가하여 매핑 작업을 진행하였다. 매핑된 결과는 검증을 위하여 메타데이터 전문가와 생태 분야 연구자를 대상으로 인터뷰 미 설문조사를 수행하였다. 이렇게 도출된 생태 분야의 연구데이터를 관리 및 공유하기 위한 메타데이터 요소는 51개로서 필수 요소 6개, 권고 요소 23개 마지막으로 선택 요소 22개를 포함하고 있다. 다음은 적용방식에 따라 해당 요소를 정리한 것이다.

- 필수(Mandatory) : title, release date, publisher, license, type/genre, Measurement Or Fact
- 권고(Recommended) : update/modification date, temporal coverage, spatial/geographical coverage, description, spatial resolution, temporal resolution, basis Of Record, keyword/tag, identifier, relationship, resource creator, theme/category, access rights, download URL, conforms to, metadata Provider, access URL, media type, bibliographic Citation, format, contact point, was generated by, associated Party
- 선택(Optional) : rights, language, access service, compression format, short Name, byte size, frequency, Funding Reference, series, dataset distribution, Version, packaging format, endpoint URL, has policy, landing page, is referenced by, qualified attribution, alternateIdentifier, endpoint description, serves dataset, qualified relation, resource relation

본 연구는 DCAT 기반으로 DataCite 4.3, ABCD, Darwin Core 그리고 EML 등 생태 분야의 메타데이터를 분석하여 메타데이터를 설계하였다는 데 의의를 가질 것이다. 앞으로 많은 분야에서 DCAT을 이용한 응용 프로파일 설계가 시도될 것이라고 판단된다. 따라서 본 연구의 결과는 생태 분야에서 생산되는 연구데이터를 관리 및 재사용하기 위한 플랫폼 설계가 가능할 것이며, 제안되는 메타데이터 요소를 기반으로 연구데이터 관리 및 공유시스템 구축 시 DB를 설계가 가능할 것이다. 또한 DCAT의 장점인 타 시스템과 연계 시 메타데이터 교환 형식을 제시하는데도 사용할 수 있을 것이다. 앞으로 타 분야에서도 이러한 DCAT을 이용한 메타데이터 설계 시 본 연구가 중요한 사례로서 이용될 수 있기를 기대해 본다.

참 고 문 헌

박경현, 원희선, 류근호. 2018. 오픈데이터 플랫폼의 상호운용성을 위한 DCAT 기반 메타데이터 변환도구 설계 및 구현.『한국디지털콘텐츠학회논문지』, 19(1): 59-65.

박옥남. 2018. 연구데이터 관리를 위한 온톨로지 설계에 대한 연구.『한국기록관리학회지』, 18(1): 101-127.

박진호. 2019. DCAT을 활용한 디지털도서관 데이터셋 관리와 서비스 설계.『한국문헌정보학회지』, 53(2): 247-266.

신도겸 외. 2019. 교통 분야 Data 관리와 공유를 위한 DCAT 기반 Data Catalogue 표준 연구: DCAT-Trans.『대한교통학회』, 37(5): 430-444.

Access to Biological Collection Data Task Group. 2007. Access to Biological Collection Data (ABCD), Version 2.06. Biodiversity Information Standards (TDWG).

Aleksejs, Kontijevskis. 2007. Scientific Databases Biological Data Management. The Linnaeus Centre for Bioinformatics and Dept. of Pharmaceutical of Biosciences. 〈https://pdfs.semanticscholar.org/0128/a20c5fa77e142c5aeb86c350ecc778981641.pdf〉 [cited 2020. 1. 5].

Data Center Frontier. 2017. Data Center First: Intel's Vision For A Data-Driven World. 〈https://datacenterfrontier.com/data-center-first-intels-vision-for-a-data-driven-world/〉 [cited 2020. 8. 5].

Dekkers, M. et al. 2016. "StatDCAT-AP, A Common Layer for the Exchange of Statistical Metadata in Open Data Portals." In SemStats@ ISWC.

Jones, Matthew et al. 2019. Ecological Metadata Language Version 2.2.0. KNB Data Repository. doi:10.5063/F11834T2

Max, W. 2010. Datacite: The International Data Citation Initiative: Datasets Programme.

Nuffelen, Bert Van. 2020. DCAT Application Profile for Data Portals in Europe Version 2.0.1. 〈https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/distribution/dcat-ap-201-pdf〉 [cited 2020. 9. 23].

Paolo, Cappello, Marco Comerio and Irene Celino. 2017. "BotDCAT-AP: An Extension of the DCAT Application Profile for Describing Datasets for Chatbot Systems." PROFILES Workshop @ ISWC. 2017. Vienna.

Perego, A. et al. 2017. "GeoDCAT-AP: Representing Geographic Metadata by Using the

DCAT Application Profile for Data Portals in Europe." In Joint UNECE/UNGGIM Europe Workshop on Integrating Geospatial and Statistical Standards, Stockholm, Sweden.

TDWG. 2009. Simple Darwin Core. 〈https://dwc.tdwg.org/simple/〉 [cited 2020. 1. 12].

The Ecological Society of America. 2013. Long-Term Studies Section. 〈https://www.esa.org/about/awards/chapter-and-section-awards/long-term-studies-section/〉 [cited 2020. 1. 2].

Wieczorek, J. et al. 2012. "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard." *PLoS ONE*, 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

William, K. et al. 1997. "Nongeospatial Metadata for the Ecological Sciences." *Ecological Applications*, 7(1) (Feb., 1997): 330-342.

W3C. 2019. Data Catalog Vocabulary (DCAT) - Version 2. 〈https://www.w3.org/TR/vocab-dcat-2/〉 [cited 2020. 5. 7].

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

Park, Jin-Ho. 2019. "Designing Dataset Management and Service System for Digital Libraries Using DCAT." *Journal of the Korea Society for Library and Information Science*, 53(2): 247-266.

Park, Kyounghyun, Hee Sun Wonk, and Keun Ho Ryu. 2018. "A Design and Implementation of a DCAT-based Metadata Transformation Tool for Interoperability in Open Data Platforms." *Digital Contents Society*, 19(1): 59-65.

Park, Ok-Nam. 2018. "A Study on Ontology Design for Research Data Management." *Journal of Korea Society of Archives and Records Management*, 18(1): 101-127.

SHIN, Doh Kyoum et al. 2019. "Data Catalogue Standards Based on DCAT for Transportation Data: DCAT-Trans." *Korea Society of Transportation*, 37(5): 430-444.

[부록] 생태 연구데이터 관리 및 공유를 위한 메타데이터 요소 및 스키마

〈부록 1〉 생태 분야를 위한 메타데이터 요소와 스키마

ID	메타데이터 요소	표시 상수	빈도	적용방식	정의, 이용, 값, 예제, 제한 조건
1	access rights	접근 권한	0-1	R	• 리소스에 액세스하는 사람 또는 보안 상태 표시에 대한 정보 • 액세스 권한에는 개인 정보, 보안 또는 기타 정책에 따른 액세스 또는 제한에 관한 정보가 포함될 수 있음 • 활용예시: Contact the Administrator
2	conforms to	표준	0-n	R	• 기술된 자원이 준수하는 표준 • 서술된 자원이 준수하고 있는 확립된 표준으로, 카탈로그 레코드 메타데이터가 준수하는 모델, 스키마, 온톨로지, 뷰, 프로파일 등을 나타냄 • 활용예시: ISO, W3C, ECOmark
3	contact point	연락처	0-n	R	• 데이터셋에 대한 문의 사항에 대응할 수 있는 담당자의 연락처 정보를 기술함 • 이메일, 전화, 웹주소 등이 값으로 기술할 수 있음 • 활용예시: 홍길동 (041-111-1111, metadata@nie.re.kr)
4	resource creator	생성자	1-n	M	• 데이터셋을 생성 또는 묶은 사람, 조직 등을 기술함 • 기술되는 값의 언어 정보가 함께 기술되는 것을 권고함 • 데이터셋을 웹에 공개 또는 출판하기 위해 반드시 기술되어야 함 • 활용예시: 국립생태원
5	associated Party	관련자	0-n	R	• 자원과 연관되어야 하는 다른 개인/단체/직위의 전체 이름 • 자원을 작성하거나 또는 유지/보수할 수 있는 개인/단체/직위 • 활용예시: 홍길동, 국립생태원 생태정보연구실 에코뱅크팀 등
6	Funding Reference	펀딩 참조 기관	0-n	O	• 연구비 정보로서 연구비 지원기관 정보, 연구비, 프로젝트명, 프로젝트 코드 등을 기술 • 활용예시: 연구재단 (프로젝트 코드 - GBMF3859.01)
7	description	설명	0-n	R	• 데이터셋의 내용에 대한 설명 • 기술방법에는 제한이 없으나 주로 맥락 정보를 상세히 기술 • 활용예시: 이 데이터셋은 OOO에서 1시간 주기로 수집된 데이터셋으로서 재사용을 위해서는 OOO 소프트웨어가 필요함
8	title	이름, 제목	1-n	M	• 데이터셋의 이름 또는 명칭 • 기술되는 값의 언어 정보가 함께 기술되는 것을 권고함 • 데이터셋을 웹에 공개 또는 출판하기 위해 반드시 기술되어야 함 • 활용예시: Microbial Genomic Reference Materials
9	shortName	약칭/약어	0-n	O	• 기술된 자원의 약칭 (파일명을 기재하기도 한다.) • 문서화된 자원을 지칭하는 간략한 이름. 타 저장소 관련 파일명을 저장한다. • 활용 예시: vernal-data-1999, 20200125_000_KST_v01 등
10	release date	발행일	1	M	• 데이터셋 발행일 • 활용 예시: 2020-03-26
11	update/modification date	갱신/ 수정일	0-n	R	• 카탈로그 항목이 변경, 갱신, 수정된 가장 최근 날짜 • 활용 예시: 2020-03-26
12	language	언어	0-1	O	• 목록화된 자원의 문자열 • 메타데이터(제목, 설명 등) 또는 데이터셋 배포 시 값에 쓰이는 자연어 • 자원이 여러 언어로 쓰인 경우, 반복하여 기술할 수 있다. • 활용 예시: Korean, English, Japanese 등
13	publisher	출판사	1	M	• 데이터셋을 웹에 공개 또는 출판하는 주체 • 연구자, 기관, 서비스명 등을 기술함 • 데이터셋을 웹에 공개 또는 출판하기 위해 반드시 기술되어야 함 • 활용예시: 국립생태원
14	identifier	식별자	1	M	• 데이터셋의 식별자 또는 식별값 • 공식적인 식별 체계에 맞는 문자열이나 번호를 사용 • 공식적인 식별 체계에 제한이 없으나, URI(웹주소인 URL을 포함), 디지털 객체 식별기호(DOI), 지털콘텐츠식별체계(UCI) 등이 포함됨 • 데이터셋을 웹에 공개 또는 출판하기 위해 반드시 기술되어야 함 • 활용예시: DOI, URL 등

ID	메타데이터 요소	표시 상수	빈도	적용방식	정의, 이용, 값, 예제, 제한 조건
15	alternate Identifier	대체 식별자	0-n	O	• 해당 엔티티의 추가적인 식별자 • 기관 내 고유/대표성을 가진 기본 식별자는 ID(identifier)이나, 보조 식별자는 엔티티 내에서 구분용으로 쓰일 수 있음 • 주로 쓰이지 않지만 식별성을 가진 보조 항목 • 활용 예시: 가령 프로젝트의 식별자가 A_20200101_001(관내 프로젝트 번호)이라면 추가적인 보조 식별자로 NEI_2020_01_204(타 기관에서 쓰이는 프로젝트 번호)를 둘 수 있다.
16	Related Identifier	관련 자원 식별자	0-n	R	• 관련 자원의 식별자 • 국제적 고유 식별자여야 함 • 활용예시: DOI, ISBN 등
17	theme/category	카테고리 / 주제영역	0-n	R	• 데이터셋의 내용이 지닌 주제 정보를 기술함 • 활용예시: Human Geography
18	basisOfRecord	데이터셋의 종류 및 특성	0-n	R	• 데이터 레코드의 특징 • 활용예시: 종, 표본명, 기계로 관측한 기록이나 현장 노트, 수집 증거물 등
19	type/genre	유형	1	M	• 데이터셋 유형을 자유롭게 기술 • 인구통계 데이터셋인지, 컨퍼런스 초록데이터 인지 등을 자유롭게 기술 • 활용예시: Dataset/Census Data, Text/Conference Abstract
20	resource relationship	자원 관계	0-n	O	• 목록화된 항목과 불특정한 관계가 있는 자원 • 활용예시: 해당 항목의 일부인 자원, 해당 항목이 준수하는 표준, 해당 항목의 이전 버전 등
21	series	연작	0-n	O	• 대상 자원의 연속된 데이터셋에 대한 정보 • 특정 년도 학술지의 특정 권에 있는 저널 한 부 • 활용 예시: Volume 20
22	qualified relation	한정 관계	0-n	O	• 다른 데이터셋과의 관계를 설명한 링크 • 관계의 특성이 잘 알려져 있지만, 주어진 메타데이터 요소와 부합하지 않는 요소 • 활용예시: 보존자, 소유주, 설계자 등
23	keyword/tag	키워드/ 태그	0-n	R	• 데이터셋을 설명하는 키워드/태그 • 활용예시: DNA, nucleosome positioning
24	landing page	랜딩 페이지	0-n	O	• 목록(카탈로그), 데이터셋, 배포 및 추가정보로 접근할 수 있는 웹페이지 브라우저 • 활용예시: https://www.w3.org/TR/vocab-dcat-2/
25	qualified attribution	한정 속성	0-n	O	• 데이터셋에 대해 어떠한 책임이 있는 에이전트 • 활용 예시: 생성자와 출판사를 제외한 다른 형태의 에이전트. • 활용 예시: 보존자, 소유주, 설계자 등
26	license	라이선스	0-1	R	• 데이터셋에 제공될 수 있는 라이선스와 권리 정보 • 라이선스란 무언가를 사용, 수행, 소유할 수 있도록 공식적으로 허가 또는 승인하는 것 • 활용 예시: 크리에이티브 커먼즈(Creative commons), 오픈 데이터 커먼즈(Open Data Commons)
27	rights	권한	0-1	O	• 데이터셋의 저작권 정보를 기술함 • 활용 예시: 해당 데이터셋의 저작권은 연구자에게 귀속됨.
28	has policy	ODRL 준수 정책	0-n	O	• 자원과 관련된 권한을 나타내는 ODRL 준수 정책 • ODRL (Open Digital Rights Language)은 콘텐츠 및 서비스 사용에 대한 진술을 표현하기 위해 유연하고 상호 운용 가능한 정보 모델, 어휘 및 인코딩 메커니즘을 제공하는 정책 표현 언어 • 활용예시: [ODRL-MODEL] 정책
29	is referenced by	참조 자원	0-n	O	• 참조, 인용 또는 카탈로그화된 데이터셋을 가리키는 관련 리소스 • 활용 예시: Microbial Genomic Reference Materials
30	metadata Provider	메타데이터 제공자	0-n	R	• 데이터셋의 설명 및 기타 메타데이터를 제공한 개인/단체/직위 • 활용 예시: 데이터를 수집한 과학자, 데이터 관리자 등
31	MeasurementOrFact	측정 요소 세부 사항 기술 요소	0-n	R	• 자원을 식별할 수 있는 측정치 • 활용 예시: 그램 단위의 유기체 무게, 표면 수온(섭씨), 측정치의 오차 범위 등
32	bibliographic Citation	레코드 인용 방법	0-n	R	• 자원을 식별할 수 있을 정도의 충분한 설명 • 활용 예시(종): Museum of Vertebrate Zoology, UC Berkeley. MVZ Mammal Collection (Arctos). Record ID: http://arctos.database.museum/guid/MVZ:Mamm:165861?seid=101356. Source: http://ipt.vertnet.org:8080/ipt/resource.do?r=mvz_mammal • 활용 예시(분류군): Oliver P. Pearson. 1985. Los tuco-tucos (genera Ctenomys) de los Parques Nacionales Lanin y Nahuel Huapi, Argentina Historia Natural, 5(37):337-343.

ID	메타데이터 요소	표시 상수	빈도	적용방식	정의, 이용, 값, 예제, 제한 조건
33	Version	버전	0-1	O	• 데이터셋의 버전 번호 • 데이터 버전관리 지원 여부를 기입함 • 활용예시: 20200125_000_HKD_v01
34	dataset distribution	데이터셋 배포	0-n	O	• 데이터셋의 사용가능한 배포, 구체적인 표현 • 데이터셋의 자연어, 미디어 유형 및 형식, 시간 및 공간 등을 포함한다. • 활용예시: dataset-001-csv
35	frequency	빈도	0-1	O	• 데이터셋이 출판(게시)되는 빈도 • 데이터셋이 업데이트되는 속도를 말한다. • 활용예시: 매일 업데이트 되는 경우 daily, 막 출간된 경우 continuous
36	spatial / geographical coverage	공간 / 지리적 범위	0-n	R	• 데이터셋에 적용된 지리적 영역 • 범위를 설명하기 위해, 범위의 중심좌표(centroid)나 범위의 꼭짓점 좌표들을 나열한다. • 활용예시: 중심좌표 "점(4.88412 52.37509)"
37	spatial resolution	공간 해상도	0-1	R	• 데이터셋 항목 사이의 가장 작은 거리. • 십진수 미터 단위로 나타낸다. • 활용예시: 수질 검사를 강 100m마다 행했을 경우, 100m
38	temporal coverage	시간 범위	0-n	R	• 데이터셋이 다루는 기간 • 시작 날짜(또는 년도) 와 종료 날짜(또는 년도)로 나타낸다. • 활용예시: 시작 날짜 "2016-03-04", 종료 날짜 "2018-08-05"
39	temporal resolution	시간 해상도	0-1	R	• 데이터셋 항목 사이의 가장 작은 기간 • frequency와 같이 사용된다. • 활용예시: 수온을 15분마다 측정했을 경우 15m, 기온을 1시간 마다 측정했을 경우 1h
40	was generated by	활동명	0-n	R	• 데이터셋 생성을 위한 비즈니스 컨텍스트를 생성하거나 제공하는 활동 • 이 데이터셋을 생성한 프로젝트 이름 등을 말한다. • 활용예시: 전주시수질조사
41	accessURL	접속URL	0-n	R	• 데이터셋의 배포에 접근할 수 있는 리소스의 URL • 활용예시: http://example.org/geopedia/sparql
42	access service	접속 서비스	0-n	O	• 데이터셋의 배포에 접근할 수 있는 데이터 서비스 • 데이터서비스의 표준 및 매체 유형 등이 포함된다. • 활용예시: 〈http://www.example.org/files/001.csv〉
43	downloadURL	다운로드 URL	0-n	R	• 지정된 형식의 다운로드 가능한 파일의 URL • CSV 파일 또는 RDF 파일. 형식은 배포판의 format 또는 mediaType으로 표시됨. • 활용예시: 〈http://www.example.org/files/001.csv〉
44	byte size	바이트 크기	0-1	O	• 데이터셋 배포의 크기 (십진수 바이트 단위) • 활용예시: 24byte
45	media type	매체 유형	0-n	R	• IANA에서 정의한 배포의 매체 유형 • 활용예시: png형태의 이미지 파일의 경우, image/png
46	format	포맷	0-1	R	• IANA에서 정의하지 않은 배포의 파일 형식 • 활용예시: png형태의 이미지 파일의 경우, text/csv
47	compression format	압축 포맷	0-n	O	• 다운로드 파일의 크기를 줄이기 위해 데이터가 압축된 형태로 되어 있는 배포의 압축 포맷 • 가능한 IAVA에서 정의한 매체 유형을 사용해야 한다. • 활용예시: zip 파일
48	packaging format	패키지 포맷	0-n	O	• 하나 이상의 데이터 파일이 함께 그룹화 되어있는 패키지 포맷 • 가능한 IAVA에서 정의한 매체 유형을 사용해야 한다. • 활용예시: TAR 파일, bargit 파일 등
49	endpointURL	엔드 포인트 URL	0-n	O	• 서비스의 루트 위치 또는 기본 엔드 포인트 (웹 분석 가능 IRI) • 엔드 포인트란 API가 기능을 수행하는 데 필요한 자원에 접근할 수 있는 위치이다. • 활용예시: http://example.org/api/table-005/capability
50	endpoint description	엔드 포인트 설명	0-n	O	• 오퍼레이션, 파라미터 등 엔드 포인트를 통해 사용 가능한 서비스에 대한 설명 • 엔드포인트 인스턴스에 대한 구체적인 세부 정보를 제공한다. 기계가 읽을 수 있는 형식으로 표현된다. • 활용예시: http://example.org/api/table-005/capability
51	serves dataset	데이터 컬렉션	0-n	O	• 이 데이터 서비스가 배포하는 데이터셋 • 활용예시: dataset-003, dataset-004