

생태 분야 데이터 리포지터리 운영 현황 분석 및 EcoBank 서비스 제안

Analysis of Ecological Data Repository Operation Status and EcoBank Service Proposal

김 주 섭 (Juseop Kim)*

강 효 숙 (Hyosuk Kang)**

김 선 태 (Suntae Kim)***

목 차

- | | |
|-----------------------------|-------------------|
| 1. 서 론 | 4. EcoBank 서비스 제안 |
| 2. 이론적 배경 | 5. 논의 및 결론 |
| 3. 해외 생태 분야 데이터 리포지터리 현황 분석 | |

초 록

데이터 공유 및 재사용은 필수가 되었다. 데이터 리포지터리는 이러한 데이터 공유 및 재사용을 위한 핵심 도구이다. 본 연구의 목적은 국립생태원이 구축 및 운영 중인 EcoBank의 서비스를 제안하기 위함이다. 연구 목적을 달성하기 위하여 re3data.org에 등록된 생태 분야 해외 데이터 리포지터리 123개 중 10개를 선정하여 조사 및 분석하였다. 분석 결과 3가지 서비스가 공통으로 도출되었다. 3가지 서비스는 첫째, 연구데이터 정책, 둘째, 연구데이터 품질 검토 그리고 연구데이터 관리 교육 및 워크숍으로 구성된다. 여기에 EcoBank의 글로벌한 데이터 공유를 위해서는 re3data.org와 같은 데이터 리포지터리 레지스트리에 등록을 해야 하며 리포지터리의 신뢰성 및 품질 확보를 위한 인증을 추진할 것을 제안해 본다.

ABSTRACT

Sharing and reusing data has become essential. Data repositories are a key tool for sharing and reusing this data. The purpose of this study is to propose the service of EcoBank, which is being built and operated by the National Institute of Ecology. To achieve the research purpose, 10 out of 123 foreign data repositories in the field of ecology registered on re3data.org were selected, investigated, and analyzed. As a result of the analysis, three services were derived in common. The three services consist of first, research data policy, second, research data quality review, and research data management training and workshops. Here, in order to share EcoBank's global data, it is necessary to register with a data repository registry such as re3data.org, and it is proposed that certification be promoted to ensure the reliability and quality of the repository.

키워드: 데이터 리포지터리, 연구데이터 관리, 국립생태원, Ecobank, re3data

Data repository, research data management, National Institute of Ecology, Ecobank, re3data

* 전북대학교 문헌정보학과 강사, 연구데이터융복합연구소 전임연구원
(kimjuseop@jbnu.ac.kr / ISNI 0000 0004 7492 1806) (제1저자)

** 전북대학교 교육대학원(사서교육전공) 석사과정 재학, 연구데이터융복합연구소 연구원
(hyosuk79@jbnu.ac.kr / ISNI 0000 0005 1319 792X) (공동저자)

*** 전북대학교 문헌정보학과 부교수, 연구데이터융복합연구소장
(kim.suntae@jbnu.ac.kr / ISNI 0000 0004 6492 6355) (교신저자)

논문접수일자: 2023년 10월 30일 최초심사일자: 2023년 11월 3일 게재확정일자: 2023년 11월 11일
한국문헌정보학회지, 57(4): 289-310, 2023. <http://dx.doi.org/10.4275/KSLIS.2023.57.4.289>

※ Copyright © 2023 Korean Society for Library and Information Science

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. 서론

1.1 연구 배경 및 연구 질문

데이터의 공유 및 재사용은 이제 필수가 되어 가고 있다. 이러한 점에서 많은 비용과 노력이 투입된 데이터 관리를 위하여 최근에는 해당 커뮤니티, 학문 분야 그리고 연구기관 등을 중심으로 RDM(Research Data Management, 이하 RDM)이라는 서비스 형태로 제안되고 있다. 데이터의 공유 및 재사용에 있어 해당 데이터를 담을 수 있는 그릇 즉, 데이터 리포지터리는 데이터 관리에 있어 반드시 필요한 도구이다. 이러한 데이터 리포지터리는 연구데이터의 저장, 관리 및 보존을 위한 체계적인 플랫폼을 제공하고 있으며 글로벌 연구 커뮤니티를 위한 필수 인프라 역할을 하며 다양한 분야에 걸친 연구의 투명성, 재현성 및 협업을 가능하게 하고 있다.

2023년 9월 27일자로 과학기술정보통신부에서는 「국가연구데이터 관리 및 활용 촉진에 관한 법률」 제정안을 입법 예고하였다. 해당 법률의 주요 내용 중 '연구데이터 통합플랫폼(안 제13조)'를 통해 저장소 구축·연계 등을 지원을 위한 근거를 마련하도록 법 조항에 제시되어 있다. 이것은 법령으로 리포지터리의 구축이 강제되어 가고 있음을 알 수 있는 사례라고 판단되며 따라서 앞으로 연구기관에서는 데이터 공유 및 재사용을 위한 데이터 리포지터리는 필수 불가결한 도구가 될 것이다(과학기술정보통신부, 2023).

본 연구의 목적은 생태 분야 데이터 리포지터

리 현황을 살펴보고, 국립생태원의 EcoBank¹⁾와 비교하여 서비스 개선방안을 도출함에 있다. 연구의 목적을 달성하기 위하여 다음과 같이 연구 질문을 설정하였다.

- 연구질문 1: 생태 분야 데이터 리포지터리의 운영 현황은 어떠한가?
- 연구질문 2: 국립생태원 EcoBank의 운영 현황은 어떠한가?
- 연구질문 3: EcoBank의 서비스는 어떻게 구성될 수 있는가?

1.2 연구 방법

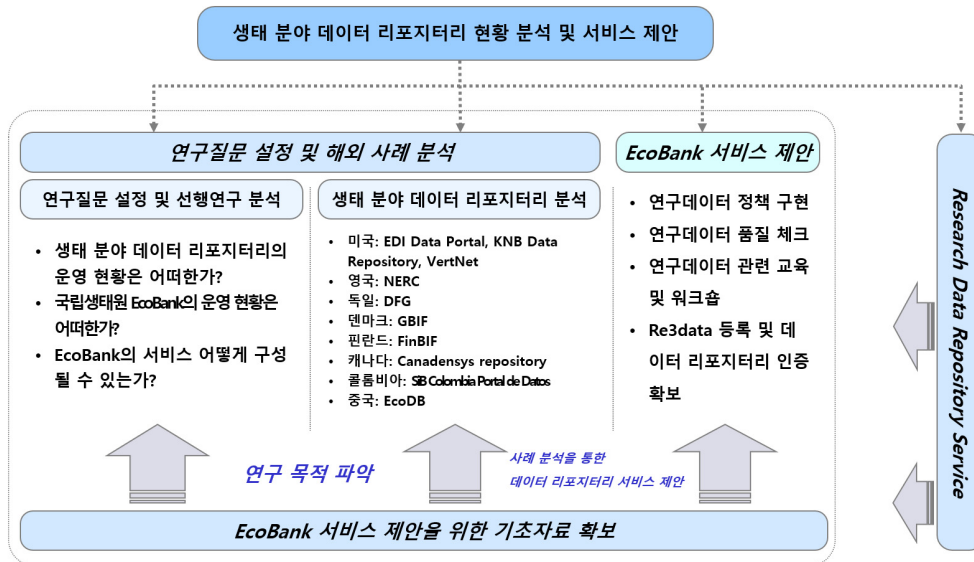
본 연구의 목적은 생태 분야 글로벌한 데이터 리포지터리 현황을 조사 및 분석하여 국립생태원이 운영하고 있는 데이터 리포지터리인 EcoBank 서비스를 제안하기 위함이다. 1장 1절에서 제시된 연구질문에 답을 하기 위하여 다음의 <표 1>과 같이 연구 프로세스를 설정하였다.

먼저 1단계는 연구배경과 연구질문을 도출하였으며 2단계에서는 EcoBank와 유사한 데이터 리포지터리를 선정하여 조사 및 분석하였다. 선정 기준에는 re3data의 속성에 나타난 메타데이터 표준, 영구식별자, 데이터 현황 및 리포지터리 인증 등을 포함하였다. 다음으로 3단계에서는 2단계에서 조사 및 분석된 내용과 국립생태원 EcoBank와 비교하였다. 마지막으로 EcoBank의 서비스를 제시하는 것으로 마무리하였다. 다음의 <그림 1>은 본 연구에서 설계한 연구 방법 및 연구 전략을 도식화한 것이다.

1) EcoBank는 국립생태원에서 개발한 데이터 리포지터리로서 2018년에 서비스를 시작하였다.

〈표 1〉 연구 프로세스

| 단계 | 내용 |
|-----|--|
| 1단계 | • 연구범위 설정 및 연구질문 도출 |
| 2단계 | • re3data.org 내 생태 분야 데이터 리포지터리 탐색 • EcoBank와 유사한 생태 분야 데이터 리포지터리 도출 • 생태 분야 데이터 리포지터리 조사 및 분석 |
| 3단계 | • 해외 데이터 리포지터리 및 EcoBank 비교 분석 |
| 4단계 | • EcoBank 서비스 제안 |



〈그림 1〉 연구방법 및 전략

본 연구는 해외 생태분야 데이터 리포지터리를 분석하여 국내 EcoBank의 서비스 개선 방안을 제시하는 것이다. 해외 현황을 분석하여 국내에 적용하는 것은 전문가 검토 및 현장 적용 가능성을 검증해야 하지만 본 연구에서는 그러한 절차가 빠져 있다는 것이 연구의 제한점이라고 볼 수 있다. 하지만 데이터 리포지터리가 법적으로 강제되어 가고 있는 시점에서 해외 데이터 리포지터리의 서비스가 어떻게 구성되어 있는지 살펴보고 국내 리포지터리에 서

비스를 제안할 수 있는 기초연구로서는 타당하다고 판단된다.

2. 이론적 배경

2.1 데이터 리포지터리 현황

데이터 리포지터리는 연구데이터를 수집, 저장 및 보존을 통해 재사용할 수 있는 수단이다. 연

구데이터 리포지터리 레지스트리인 re3data는 전세계의 데이터 리포지터리를 색인화한 사이트로서 2012년에 구축되었다. re3data는 연구자, 펀딩 기관, 도서관 및 출판사에 연구데이터 리포지터리 환경과 유형에 대한 개요를 제공하고 있어 세계 연구데이터 인프라의 가장 대표적인 참고자료로 활용되고 있다. 또한, 리포지터리 인증 여부를 포함하고 있어 리포지터리 신뢰성을 확인할 수 있으며 2023년 9월 현재, 3,155개의 리포지터리가 등록되어 있다. 다음의 <표 2>는 국가별 데이터 리포지터리 현황을 나타낸 것이다(re3data, 2023).

국가별 데이터 리포지터리를 확인한 결과, 미국이 1,171개로 가장 많았고 그 다음으로 독일, 캐나다, 영국, 유럽 연합 순으로 나타났다. 한국은 28위로 폴란드와 동일한 13개의 리포지터리를 보유하고 있다. 한국이 보유한 13개의 데이터 리포지터리의 경우, 글로벌 컨소시엄 리포지터리 3개, 더 이상 운영되지 않은 리포지터리 3개를 제외한 순수 리포지터리 개수는 7개로 나타났다. 다음은 7개의 데이터 리포지터리를 나타낸 것이다.

- Korean Labor & Income Panel Study (KLIPS)
- Jurisdictional Ocean Information Sharing System(JOISS)
- DataON Repository
- International Space Environment Service(ISES)
- Conservation Genome Resource Bank for Korean Wildlife(CGRB)
- Geo Big Data Open Platform
- Korea Polar Data Center(KPDC)

본 연구의 대상인 생태분야 데이터 리포지터리를 분석하기 위하여 re3data의 주제 분류에 생태학이 포함된 리포지터리를 도출하였다. 도출하기 위해 사용한 주제 구분은 다음과 같다.

- 생명 과학 - 생물학 - 식물과학 - 식물 생태 및 생태계 분석
- 생명 과학 - 생물학 - 동물학 - 동물생태학, 생물다양성 및 생태계 연구

<표 2> re3data의 국가별 데이터 리포지터리 현황

| 순위 | 국가 | 개수 | 순위 | 국가 | 개수 | 순위 | 국가 | 개수 |
|----|---------------|-------|----|-------|----|----|-------------|-----------|
| 1 | 미국 | 1,171 | 11 | 네덜란드 | 77 | 21 | 러시아 연방 | 23 |
| 2 | 독일 | 505 | 12 | 일본 | 66 | 22 | 폴란드 | 21 |
| 3 | 캐나다 | 395 | 13 | 스페인 | 55 | 23 | 브라질 | 19 |
| 4 | 영국 | 320 | 14 | 인도 | 51 | 24 | 멕시코 | 18 |
| 5 | 유럽 연합 | 287 | 15 | 이탈리아 | 50 | | 남아프리카공화국 | 18 |
| 6 | International | 283 | 16 | 오스트리아 | 45 | 26 | 포르투갈 | 16 |
| 7 | 프랑스 | 128 | 17 | 벨기에 | 38 | 27 | 뉴질랜드 | 14 |
| 8 | 오스트레일리아 | 103 | 18 | 노르웨이 | 31 | 28 | 핀란드 | 13 |
| 9 | 스위스 | 84 | 19 | 스웨덴 | 29 | | 대한민국 | 13 |
| 10 | 중국 | 81 | 20 | 덴마크 | 27 | 30 | 체코, 그리스, 대만 | 12 |

‘식물 생태 및 생태계 분석’에 해당되는 데이터 리포지터리는 72개 그리고 ‘동물생태학, 생물다양성 및 생태계 연구’에는 93개의 데이터 리포지터리가 포함되어 있다. re3data의 주제 구분은 독립적이지 않고 중복을 허용하기 때문에 이 수치에는 중복된 리포지터리가 포함되어 있다.

2.2 EcoBank

EcoBank는 국립생태원에서 조사·연구한 생태정보 및 국내외 생물·생태계 관련 정보를 통합, 연계, 관리, 서비스하는 생태정보 포털서비스이다. 개방형 웹 기반 생태 데이터베이스 플랫폼을 지향하는 EcoBank는 웹 GIS, 애플리케이션 프로그래밍 인터페이스(API), 반응형 웹, 오픈소스 소프트웨어(OSS)의 핵심기술로 구성된다. EcoBank의 생태 데이터셋은 국립생태원 뿐만 아니라 국내외 생태 관련 플랫폼 및 리포지터리를 포함하고 있으며, 연구자, 정책 입안자, 일반 사용자를 대상으로 서비스를 제공하고 있다(Kim et al., 2021).

EcoBank의 목표는 국립생태원(National Institute of Ecology, NIE), 국내 및 국제 데이

터베이스의 주요 접근점 역할을 하는 것으로, 국립생물자원관 등 국내 환경기관과 생태 데이터를 교환하고 정부 공공데이터포털과의 연계를 통해 더 많은 이용자가 국립생태원의 생태 데이터에 접근할 수 있는 기회를 제공하고 있다. 또한 국내외 생태계 분석에 중요한 역할을 수행할 수 있도록 공간정보산업진흥원에서 운영하는 공간정보 오픈플랫폼 서비스인 브이월드(VI)의 API 서비스를 이용하여 높은 해상도와 최신 공간정보를 바탕으로 한 GIS(Geographic Information System, 지리정보시스템) 서비스를 제공하고 있다. EcoBank에서 보유하고 있는 데이터에 대한 현황은 다음의 <표 3>과 같다(국립생태원, 2023).

EcoBank의 서비스는 크게 생태공간정보 서비스, 데이터개방 서비스, 참여서비스, 통계서비스 그리고 출판서비스로 구성되어 있다. 먼저, EcoBank의 생태공간정보 및 생태·자연도 서비스에서는 주소/자료검색, 조사자료 검색, 지표종 정보, 습지 공간정보, 로드킬 발생빈도, 지자체 생태정보로 나누어진 다양한 검색 서비스를 제공하고 있다. 두 번째 데이터 개방 서비스의 경우, 다음과 같은 내용을 제공하고 있다.

<표 3> EcoBank 보유 데이터 현황

| 종류 | 건수 | 세부항목 | |
|----------|---------|--|--|
| 데이터 | 1,496만건 | <ul style="list-style-type: none"> • 전국자연환경조사 • 생태계 정밀조사 • 장기생태 | <ul style="list-style-type: none"> • 생태·자연도 • 생태통로 • 외래생물 |
| 국내외 연계 | 113건 | <ul style="list-style-type: none"> • 환경부 • GBIF 국토교통부 | <ul style="list-style-type: none"> • 지질자원연구원 • 환경정책평가연구원 |
| OPEN API | 86건 | <ul style="list-style-type: none"> • 생태 자연도 • 전국자연환경조사 | <ul style="list-style-type: none"> • 생태계 정밀조사 |

- 연구자료: 국립생태원과 외부기관 또는 등록 회원이 생태분야 연구자료를 등록하여 연구자와 공유할 수 있으며, 등록된 자료는 DOI 발행을 통해 이력을 추적관리 할 수 있음.
- 공개데이터: 행정안전부 공공데이터포털(data.go.kr)과 연계하여 국가중점 개방데이터(생태·자연도, 식생평가, 습지평가, 지형평가 4종) 및 생태정보에 대한 데이터 다운로드 서비스를 제공하고 있음.
- 오픈 API: EcoBank는 DB에 탑재된 데이터를 Open API 방식으로 제공한다. Open API 서비스를 통해 각종 홈페이지나 프로그램에서 생태계 관련 지도가 필요할 때 활용할 수 있음.
- 유관기관 조사정보: 환경부 소속·산하 기관(국립생태원, 국립생물자원관, 국립공원공단, 국립낙동강생물자원관)별로 구축된 자연환경조사 결과를 연계하여 자연환경 정보 통합 서비스 및 링크 연계를 제공하고 있음.

다음으로 참여서비스의 경우 EcoBank 등록 회원(연구자, 일반회원 등)이 생태조사 시 취득한 생물정보, GPS 및 지도기반 위치정보, 사진 정보를 등록하고 공유할 수 있도록 기능을 제공하고 있다. 시민참여 영역으로는 전국자연환경조사, 내륙습지조사가 해당되며, 직원참여 영역으로는 생태원 생물상 조사가 있다. 이 외에도 회원들 간 생태연구와 EcoBank 이용에 대한 자유토론 공간인 생태포럼 서비스를 제공하고 있다. 네 번째 통계서비스에서는 전국자연환경조사, 생태·자연도, 장기생태 및 외래생물 조

사정보를 그래프 형태의 통계자료로 제공하고 있다. 분석 및 분포정보에서는 생물종의 위치정보와 래스터 형태의 환경변수 정보를 활용하여 종분포모델(Maxent) 분석서비스를 제공하고 있다. 또한 국립생물자원관 링크 연계를 통해 기후변화 지표종 정보를 조회할 수 있으며, 세계 생물 다양성 정보 기구(Global Biodiversity Information Facility, 이하 GBIF)를 통해 학명 정보 및 전세계 동식물, 미생물에 대한 정보를 찾을 수 있다. 마지막으로 출판서비스의 경우 EcoBank에서 제공하는 학술지로서 여기에는 국립생태원 국제 학술지인 'PNIE', 지구과학 데이터저널 'GEO DATA', 생물분자정보 등록 및 유전체지도 변환 서비스를 제공하는 '생물분자정보관리'를 포함하고 있다.

2.3 선행연구

본 연구와 선행연구의 차이점을 나타내기 위해 국내의 선행연구 10건을 분석하였다. 다음의 <표 4>는 선행연구 목록을 정리한 것이다.

Gries et al.(2023)은 Environmental Data Initiative(EDI)라는 데이터 리포지터리에 대해 분석하였다. 해당 리포지터리의 서비스에는 환경 및 생태 연구 커뮤니티를 위해 설계된 도메인별 데이터 리포지터리 운영, 데이터 관리 및 출판 서비스, 데이터 패키지 DOI 발급, ezEML, 데이터 큐레이터 그리고 FAIR 기준의 준수 여부 평가 등과 관련한 내용을 포함하고 있다. 특히, 저자는 재사용 가능한 데이터 플랫폼을 제공하고 출판 워크플로우와 검토 프로세스를 간소화하는 도구 및 지원을 제공하는 EDI의 특성이 환경 연구분야 연구자들로 하여금 EDI를 더

〈표 4〉 국내외 선행연구 목록

| 저자(연도) | 연구내용 | 비고 |
|-------------------------|---|--------------------|
| Gries et al. (2023) | Environmental Data Initiative(EDI) 분석 | 리포지터리 사례 분석 |
| Heinz et al. (2023) | 오픈 사이언스 서비스 운영과 관련된 10가지 주요 이슈를 식별하고 각 이슈에 대한 정의를 제공 | 리포지터리 사례 분석 |
| 김예현 (2023) | 과학기술분야 기관 연구데이터 리포지터리 현황 사례조사와 실무자 관점의 분석을 통해 향후 연구기관에서 연구데이터 수집과 공유를 활성화할 수 있는 방안을 제시 | 리포지터리 사례 분석 |
| 김선태 (2021) | re3data.org의 생태 분야 354개 리포지터리 현황 분석 데이터 리포지터리의 메타데이터 형식과 버전관리 수준에 대해 조사 및 분석 | 리포지터리 사례 분석 |
| 김우중 외 (2021) | Re3data.org의 자연과학 분야 연구데이터 리포지터리(402개)의 세부속성을 대상으로 기술통계분석, 계층적 군집분석, 대응일치분석, 네트워크 분석 수행 | 리포지터리 사례 분석 |
| 김주섭 외 (2020) | 생태 분야 연구데이터 관리를 위한 메타데이터 요소 도출(51개: 필수 6개, 권고 23개, 선택 22개) | 메타데이터 도출 |
| Alves et al. (2018) | 시간 경과에 따른 생태학 분야의 메타데이터 도구 사용에 대한 포괄적인 개요를 제시 | 메타데이터 동향 분석 |
| Easterday et al. (2018) | 현장에서 클라우드로: 데이터 과학의 원리를 사용하여 현장의 기록 데이터를 공유하는 세 가지 접근 방식에 대한 검토 | 생태 데이터의 FAIR 원칙 적용 |
| 안병근 외 (2018) | 한국(KRM), 미국(ICPSR), 영국(GiR), 호주(RDA), 네덜란드(DANS) 대표 리포지터리들의 연구과제 및 연구성과물 유형별 데이터 관리사례를 조사 | 리포지터리 사례 분석 |
| 허태상 (2016) | 데이터 통합 및 국제연동을 위한 한국형 생태정보시스템 구축을 위한 데이터 관리 요건 도출 | 리포지터리 구축 |

많이 사용하는 요인이 되고 있음을 밝혔다.

Heinz et al.(2023)은 re3data의 설계, 거버넌스 및 목표를 개괄적으로 설명하여 글로벌 RDR(Research Data Registry, 이하 RDR) 환경의 중심 데이터 소스로 발전한 re3data 서비스의 배경정보와 오픈 사이언스 서비스 운영과 관련된 10가지 주요 이슈를 식별하고 각 이슈에 대한 정의를 제공하였다. 다음은 10가지 주요 이슈를 정리한 것이다.

- 개방성: 메타데이터 및 스키마에 대한 무제한 액세스, 인용 투명성, RDR 커뮤니티와의 소통
- 품질보증: 검증 가능한 고품질 정보를 보

장하기 위한 모든 프로세스

- 커뮤니티 참여: 참여형 프로세스에서 글로벌 RDR 커뮤니티와의 상호작용을 보장하는 모든 활동
- 상호운용성: 확립된 표준에 따라 글로벌 RDR 커뮤니티와의 상호 작용 및 메타데이터 교환
- 개발: 글로벌 RDR 커뮤니티의 변화하는 요구사항에 동적으로 대응하기 위한 지속적인 개발
- 지속가능성: 장기적이고 안정적인 서비스 보장
- 정책: 안정적인 서비스 운영과 고품질 메타데이터 유지에 대한 책임을 수반하는 것

- 데이터 재사용: 제3자가 re3data의 메타데이터를 기반으로 글로벌 RDR 커뮤니티를 지원하는 서비스를 구축할 수 있도록 보장
- 연구용 메타데이터: RDR 설명 제공
- 커뮤니케이션: RDR 이해관계자들과의 소통

김예현(2023)은 과학기술분야 기관 연구데이터 리포지터리 현황 사례조사와 실무자 관점의 분석을 통해 향후 연구기관에서 연구데이터 수집과 공유를 활성화할 수 있는 방안을 제시하였다. 기관 연구데이터 리포지터리 활성화 방안을 제시하기 위하여 문헌 연구, 사례 분석, 국내외 기관 리포지터리 담당자와의 인터뷰를 수행하였다. 이러한 인터뷰를 바탕으로 다음과 같이 데이터 리포지터리 운영 활성화 방안을 제시하였다.

- 리포지터리 규정 및 정책 수립: 현재 연구데이터와 관련한 규정인 국가연구개발정보처리기준의 지위 향상과 리포지터리 근거 규정의 명시가 필요
- 연구데이터 공유 인식 개선: 전반적인 연구데이터 교육과 우수 사례 발굴 필요
- 연구데이터 품질 관리 강화: 연구자-담당자-위원회의 상호작용과 표준화 작업, 장기 보존을 위한 준비 필요

김선태(2021)는 생태 연구 분야 데이터 리포지터리 현황을 분석하기 위하여 re3data에 등록된 데이터 리포지터리 중 생태 분야 키워드를 활용하여 선별한 354개 저장소(전체 저장소의 약 14%)를 대상으로 현황을 조사하였다. 우선, 생태 연구데이터 리포지터리의 데이터

기술에 사용되는 주요 메타데이터 형식을 조사한 결과 FGDC/CSDGM, 더블린 코어, ISO 19115, EML(Ecological Metadata Language), DIF(Directory Interchange Format), 다윈 코어, DDI 등을 활용하는 것으로 나타났다. 다음으로 데이터 리포지터리의 데이터 버전관리 수행 비율은 전체 비율(83.9%)에 비해 다소 높은 수준(86.6%)인 것으로 분석되었다.

김우중 외 (2021)는 re3data에 등록된 자연과학 분야 연구데이터 리포지터리를 분석한 결과 연구주제에 따라 콘텐츠 타입에 차이가 있으며 북미 및 유럽 지역의 데이터 리포지터리의 활용은 활발한 반면 아시아 국가는 활용도가 낮은 것으로 확인되었다. 또한 방문빈도가 높은 데이터 리포지터리일수록 표준화된 연구문서 형태로 데이터가 공유되는 것을 확인하였다. 해당 연구에서는 국내 데이터 리포지터리 활용률 저조 원인으로 연구데이터 공유와 관련된 인프라, 역량 및 인식 부족임을 확인하면서 연구데이터 공유 및 활용 지침에 있어서 기관의 고유 연구 분야에 특화된 기준이 필요함을 지적하였다.

김주섭 외(2020)는 생태분야 메타데이터(ABCD 스키마, DarwinCore, EML, DataCite)를 분석하고 크로스워크를 진행, 메타데이터 요소검증을 위한 인터뷰 및 설문조사를 수행하였다. 연구 결과 생태 분야 연구데이터 관리를 위한 메타데이터 요소를 도출하였으며 해당 결과는 생태분야의 데이터 카탈로그에 대한 공유 및 재사용이 가능하도록 DCAT 국제 표준에 기반한 메타데이터 스키마로 설계되었다.

Alves et al.(2018)은 시간 경과에 따른 생태학 분야의 메타데이터 도구 사용에 대한 포

괄적인 개요를 제시하고자 하였다. 연구 목적을 달성하기 위해 93개 참고문헌을 대상으로 메타 분석 수행하여 생태학 분야의 메타데이터 도구 사용 현황을 시간대 별로 살펴보았다. 그 결과 생물학 분야 메타데이터 도구는 ABCD, FGDC, CIF(Crystallographic Information Framework), Humboldt Core, DIF, GML, ISO 19115 등 다양한 것으로 나타났다. 또한, 1990년대 이후 생태학 연구자들이 사용하는 메타데이터 도구 수는 증가하였으며 공통 표준에 따른 데이터 공유를 위한 글로벌 이니셔티브의 개발이 추진되었다. 저자는 이에 따라 생태학 분야에 적합한 메타데이터 모델과 표준 개발이 필요함을 기술하였다.

Easterday et al.(2018)은 현재의 디지털 생태 데이터 환경에 기여할 수 있는 과거 다크 데이터의 잠재력에 대해 논의하였다. 그 결과 전체 데이터 공정 프로세스에 걸쳐 다학제간 작업이 필요하며 데이터셋을 검색 가능하고 접근 가능하도록 만들어야 함을 언급하면서 데이터 용도 변경 기술의 발달은 다크데이터의 활용 가능성을 높일 수 있다고 덧붙였다. 또한, 저자는 데이터를 지속적으로 수집하고 접근할 수 있도록 적절한 인센티브, 지원 또는 표준이 마련되어야 함을 지적하였다.

안병근 외(2018)는 국내외 대표 연구데이터 리포지터리 홈페이지를 통해 연구데이터 관리를 위한 메타데이터들을 비교·분석함으로써 학문 지원 시스템으로서의 연구데이터 관리의 문제점과 표준 개발의 개선점을 제시하였다. 제시된 개선점에는 과도한 연구성과물 유형 분류

를 간략화 하여 이용자가 구축된 연구성과물을 직관적으로 이해할 수 있게 해야 하며 메타데이터 필수 입력항목에 대한 표준을 설정하여 이용자 혼란을 줄여야 하고 마지막으로 메타데이터 요소 설명서를 이용자에게도 제공하여 메타데이터에 대한 이용자의 이해도를 높여야 함을 포함하고 있다.

허태상(2016)은 국가 장기생태 연구정보시스템 및 구축 데이터를 분석하여 데이터 관리의 문제점을 파악하고 데이터 정제 및 구축방안에 대하여 연구를 수행하였다. 해당 연구 결과 국제적으로 데이터 연동이 가능하고 체계적인 장기생태 연구데이터 관리가 가능한 데이터 플랫폼 모델을 제시하였다. 데이터 통합 및 국제 연동을 위한 한국형 생태정보시스템 구축을 위한 데이터 관리 요건은 다음과 같다.

- EML 기반의 데이터 입력 관리 및 데이터 연계
- 외부 연동데이터의 데이터 포맷 변환 프로세스
- 국제적 데이터 공유 및 데이터 무결성
- 데이터 관리 및 검색향상을 위한 통제어휘 구축 등 데이터 특성 및 속성에 따른 요건
- 사용자 기반 데이터 접근

이상으로 국내외 논문 10편을 분석한 결과, 10편 중 6편은 데이터 리포지터리 사례 분석과 관련한 논문이며 2편은 메타데이터 1편은 데이터의 FAIR 원칙²⁾ 적용 그리고 나머지 1편은

2) FAIR 원칙은 Findable(검색가능), Accessible(접근가능), Interoperable(상호운용가능), Reusable(재사용가능)을 말한다.

데이터 플랫폼 즉 리포지터리 구축과 관련한 논문인 것으로 나타났다. 분석된 선행연구와 본 연구의 차이점은 글로벌한 생태분야 데이터 리포지터리의 서비스 현황을 파악하였다는 점이다. 각 데이터 리포지터리의 서비스 파악을 통해 분석된 결과는 국립생태원이 운영하고 있는 EcoBank와 비교하여 해당 데이터 리포지터리에 필요한 서비스 개선 방향에 활용하고자 한다.

3. 해외 생태 분야 데이터 리포지터리 현황 분석

이번 장에서는 연구데이터 리포지터리의 글로벌 레지스트리인 re3data에 등록된 리포지터리 중 EcoBank와 유사한 해외 데이터 리포지터리 현황을 살펴보고자 한다. ‘생태학 및 생물다양성’과 관련된 주제를 다루는 해외 데이터 리포지터리 수는 123개로 ‘식물 생태 및 생태계 분석(Plant Ecology and Ecosystem Analysis)’을 다루는 리포지터리 72개와 ‘동물생태학,

생물다양성 및 생태계 연구(Animal Ecology, Biodiversity and Ecosystem Research)’를 다루는 리포지터리가 93개이다. 이 중 두 주제가 모두 중복되는 리포지터리 수는 42개이며, 중복된 리포지터리 수를 제외한 ‘생물 생태 및 생태계 분석’을 다루는 리포지터리 수는 30개이며, ‘동물생태학, 생물다양성 및 생태계 연구’를 다루는 리포지터리 수는 51개이다. 총 123개의 데이터 리포지터리 중 메타데이터 표준, 영구식별자, 데이터 현황을 기준으로 10개의 리포지터리를 선택하였다. 다음 <표 5>는 분석대상 리포지터리 목록을 정리한 것이다.

먼저 GBIF는 1999년 OECD의 생물다양성 정보학 하위그룹(Biodiversity Informatics Subgroup)에서 “세계 생물다양성 정보 시설” 설립 권고에 따라 2001년 참여 정부 간 양해각서를 통해 국제기구로서 덴마크에서 공식적으로 설립되었다. 전 세계 정부가 예산을 지원하는 국제 네트워크 및 데이터 인프라로, 지구상의 모든 유형의 생명체에 대한 오픈 액세스를 목표로 하고 있으며, 데이터를 모바일화하고 검색 메커니즘, 데이터 및 메타데이터 표준, 웹 서비스, 생물 다양성을

<표 5> 분석대상 데이터 리포지터리 목록

| 리포지터리명 | 국가 | 펀딩기관 | 비고 |
|------------------------------|----------|---------------------|-----------------|
| GBIF | 덴마크 | | 글로벌 네트워크/WDS 인증 |
| BExIS | 독일 | DFG | |
| EDI Data Portal | 미국 | NSF | CTS 인증 |
| KNB Data Repository | 미국 | NCEAS/NSF | |
| VertNet | 미국 | NSF | |
| EIDC | 영국 | NERC | WDS/CTS 인증 |
| EcoDB | 중국 | CAS | |
| Canadensys repository | 캐나다, 덴마크 | CFI | |
| SiB Colombia Portal de Datos | 콜롬비아 | | 글로벌 네트워크 |
| FinBIF | 핀란드 | Partners of LAJI,FI | |

위한 인터넷 기반 정보 인프라의 기타 구성 요소를 개선하기 위한 역할을 하고 있다. 메타데이터 표준은 Repository-Developed Metadata Schemas, EML, Darwin Core를 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(GBIF, 2023).

BExIS(Biodiversity Exploratories Information System, 이하 BExIS)은 2007년부터 BE(Biodiversity Exploratories, 생물다양성탐사단) 프로젝트(BExIS 1세대)에서 개발되고 지속적으로 확장되어 2012년 DFG(Deutsche Forschungsgemeinschaft, 독일연구재단) 프로젝트가 시작되면서 무료로 사용할 수 있는 서버 기반 연구데이터 관리 애플리케이션 BEXIS2를 개발함으로써 독일에서 구축되었다. BExIS라는 웹 애플리케이션을 사용하여 연구데이터의 저장 및 교환과 관련된 주요 목표를 달성하고, BExIS에 데이터를 업로드하는 데 도움을 주며, 높은 데이터 품질(데이터 큐레이션)을 보장하면서 연구자들이 데이터를 출판할 수 있도록 조언하고 지원한다. 또한 프로젝트에서 중요한 문서 및 출판물의 교환을 촉진하고 기후 데이터와 같은 기본 데이터를 쿼리하기 위한 소프트웨어 도구를 제공한다. 메타데이터 표준은 Darwin Core, EML, ABCD를 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(BExIS, 2023).

EDI Data Portal(Environmental Data Initiative Repository, 이하 EDI)은 모든 사람이 접근할 수 있는 데이터 공유를 통해 과학적 발견의 미래를 구상하며, 공개적이고 재현 가능한 과학을 위해 환경 데이터를 보존하고, 공간과 시간에 걸친 종합을 촉진하고, 환경 변화

와 그 결과에 대한 평가를 지원한다. 또한 환경 및 생태학적 데이터가 잘 관리되고 미래에도 검색 및 재사용이 가능하도록 과학계에 핵심 서비스와 기술 전문 지식을 제공하며, 현장 스테이션, 개별 실험실 및 모든 규모의 연구 프로젝트의 연구원들이 환경 데이터를 보관하고 출판하도록 지원한다. EDI는 FAIR 데이터를 만들기 위해 최선을 다하며, 고품질 데이터와 메타데이터를 보관하고 출판하는 데 도움이 되는 지원, 교육 및 리소스를 제공한다. 그리고 안전한 데이터 리포지터리를 운영하고 LTER Network Communications Office 및 DataONE과 같은 정보 관리 분야의 다른 리더들과 긴밀히 협력하여 데이터 관리 모범 사례 및 관리를 장려한다. 이외에도 많은 수의 데이터 실무자, 소프트웨어 개발자, 연구 과학자로 구성되어 있다. EDI는 CTS(CoreTrustSeal) 인증을 받은 신뢰할 수 있는 데이터 리포지터리이다. 메타데이터 표준은 Dublin Core, OAI-ORE, EML을 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(EDI Data Portal, 2023).

KNB Data Repository(Knowledge Network for Biocomplexity, 이하 KNB)는 생태 및 환경 연구를 촉진하기 위한 국제 리포지터리로 국립생태분석종합센터(NCEAS)에서 생성된 데이터셋과 연구 제품을 장기 아카이브를 목적으로 NSF(National Science Foundation, 미국 국립과학재단/이하, NSF)의 펀딩을 받아 1998년에 미국에서 출범하였다. DataONE과 제휴하여 환경/생태학적 관련 데이터를 수집하고 이러한 데이터셋에 대한 장기적인 접근을 보장하며 디지털 개체 식별자가 포함된 데이터셋의 게시 및 연구자들이 생태 데이터를 공유, 검색,

접근 및 분석할 수 있는 풍부한 메타데이터 제공하며, 데이터 공유 및 보존에 대한 FAIR 원칙을 준수한다. 메타데이터 표준은 Dublin Core, EML을 활용하고 있으며, 영구식별자의 경우 DOI, PURL, URN, ARK 및 ORCIDID를 부여하고 있다(KNB Data Repository, 2023).

VertNet은 NSF가 펀딩을 지원하는 협업 프로젝트로 생물 다양성 데이터를 웹에서 무료로 제공하며, 생물 다양성 데이터를 검색, 캡처 및 게시할 수 있도록 설계된 도구이다. 또한 생물 다양성 데이터를 제공하고 이를 개선하기 위해 협력하는 수백 개의 생물 컬렉션 간의 협력의 핵심으로 기능하는 훈련용 엔진으로 현재 및 미래의 전문가가 데이터 품질, 큐레이션, 연구 및 데이터 출판 분야의 모범사례를 사용하고 구축할 수 있다. 메타데이터 표준은 Darwin Core를 활용하고 있으며, 영구식별자의 경우 URN을 부여하고 있다(VertNet, 2023).

EIDC(Environmental Information Data Centre, 이하 EIDC)는 육상 및 담수 과학과 관련된 국가적으로 중요한 데이터셋을 관리하는 영국의 국가 데이터 센터로 NERC(Natural Environment Research Council, 이하 NERC) 환경 데이터 서비스 중 하나이며, UKCEH(영국 생태학 및 수문학 센터)가 운영하고 있다. NERC의 펀딩을 통해 생성된 데이터 등 장기 저장, 공개 배포, 큐레이션을 위해 재사용 가능한 형식의 환경 데이터를 수집하며, 연구자들이 생성하는 데이터가 장기적으로 적절하게 선별되어 이용자 커뮤니티에 대한 보안, 접근성 및 재사용성을 보장하고, 프로젝트 기간에 생성된 데이터 관리에 대한 모범사례 조언 및 장기적 큐레이션을 위해 EIDC에 데이터 저장 지

원을 목적으로 한다. 메타데이터 표준은 Dublin Core, ISO19115를 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(EIDC, 2023).

EcoDB(National Ecosystem Data Bank, 이하 EcoDB)는 CAS(Computer Network Information Center, 중국과학원)의 컴퓨터 네트워크 정보 센터(CNIC)에서 구축하고 유지 관리하며, 글로벌 연구자, 팀, 학술 저널 및 출판사, 연구기관 및 대학에 데이터 저장, 장기 보존, 출판, 공유 및 접근 서비스를 제공한다. 현재는 ScienceDB로 리포지터리명이 변경되었으며, 다양한 데이터 공유 방법과 다양한 데이터 라이선스를 제공하여 데이터 소유자의 권리와 이익을 보호하고 있다. 또한 글로벌 오픈사이언스 데이터 공유 서비스를 목표로 하는 데이터 저장 및 출판을 위한 신뢰할 수 있는 오픈 플랫폼으로 FAIR 원칙 준수 및 연구 커뮤니티에서 데이터 공유 문화의 조성을 위해 데이터 표준을 준수하고 과학 데이터를 출판하고 있다. 메타데이터 표준은 Dublin Core를 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(EcoDB, 2023).

Canadensys repository(Database of Vascular Plants of Canada, VASCAN)는 캐나다 몬트리올 생물 다양성 센터(Université de Montréal Biodiversity Centre)에 의해 운영되며, 생물학적 컬렉션에 보유된 생물 다양성 정보를 모든 사람이 이용할 수 있도록 하고 있다. 캐나다 생물다양성 정보 시설(CBIF) 및 세계 생물다양성 정보 시설(GBIF)과 호환되는 분산 데이터 베이스 네트워크를 통해 3만 개의 표본(20%)을 디지털화, 출판 및 지리 참조하는 것이 초기 목표였으며, 현재는 11개 이상의 참여 대학, 5개의 식물원 및 2개의 박물관이 포함되어 있고,

이들은 1,300만 개 이상의 표본을 집합적으로 보관하고 있다. 또한 데이터를 공개함으로써 지리공간적 모델과 환경적 모델 내의 종 발생 데이터에 대한 교차 분석을 용이하게 하고, 이를 통해 글로벌 환경 이슈에 대한 이해를 높이고 국가 전체의 건전한 생물 다양성 정책을 개발할 수 있도록 지원하고 있다. 메타데이터 표준은 Darwin Core를 활용하고 있으며, 영구식별자의 경우 DOI, PURL(Persistent Uniform Resource Locators)을 부여하고 있다(Canadensys repository, 2023).

SiB Colombia Portal de Datos(GBIF Colombia)는 콜롬비아의 생물다양성에 관한 오픈 데이터 국가 네트워크로 1993년에 설립된 국가 환경 시스템(Sina) 창설 과정의 일환으로 1994년에 구축되었으며 GBIF에서 해당 국가의 공식 노드이다. 지속가능한 사회 건설을 위해 국가의 생물다양성에 관한 정보에 대한 공개적인 접근을 제공하며, 생물다양성 데이터와 정보의 온라인 공개를 촉진하고 다양한 이용자의 활용을 촉진하여 적시에 효율적인 방식으로 생물다양성의 포괄적인 관리를 지원한다. 또한 생물다양성과 생태계 서비스의 지식, 보존, 지속가능한 이용과 관련된 연구, 교육 또는 의사결정 과정을 지원하기 위해 정보의 이용자, 출판자, 생산자를 한자리에 모아 생물다양성에 관한 데이터와 정보의 관리를 촉진하는 역할을 한다. 메타데이터 표준은 Darwin Core, Plinian Core를 활용하고 있으며, 영구식별자의 경우 DOI를 부여하고 있다(SiB Colombia Portal de Datos, 2023).

FinBIF(Finnish Biodiversity Information Facility; Suomen Lajitietokeskus Finlands Artdatacenter, 이하 FinBIF)는 핀란드 생물

다양성 정보 시설로 생물다양성 데이터를 수집, 보관 및 배포하는 국가 서비스이다. 이는 데이터 관리를 위한 디지털 인터넷 기반 서비스 인 프라이며, 생물다양성 데이터는 주로 파트너 조직, 핀란드 환경 및 천연자원청 연구 기관, 생물다양성 데이터를 관리 및 생산하는 기타 정부 기관, 자연사 박물관에서 관리하는 컬렉션 및 시민 과학 자료에서 수집된다. 또한 종 관찰 데이터 컴파일 및 무료 사용을 위한 데이터 공유, 오픈 데이터 및 과학 촉진, 국가 분류학적 명명법 유지, 생물 다양성 데이터의 흐름과 관련된 프로세스를 강화하고 명확히 하며, 데이터 아카이빙 및 국가 생물다양성 데이터 관련 데이터 정책을 유지 및 업데이트한다. 이와 함께 공유 데이터에 주석을 달아 안정성과 품질을 확인하며, 시민 과학 진흥을 위해 노력하고 있다. 메타데이터 표준은 Darwin Core, FGDC/CSDGM을 활용하고 있으며, 영구식별자의 경우 PURL을 부여하고 있다(FinBIF, 2023).

다음의 <표 6>은 re3data에 등록된 EcoBank와 유사한 생태 분야의 해외 데이터 리포지터리 10개를 분석하여 서비스 현황을 나타낸 것이다. 카테고리로는 국가, 펀딩기관, 메타데이터 표준, 영구식별자, 운영현황(서비스 및 데이터 현황)과 기타로 나누어 정리하였다.

4. EcoBank 서비스 제안

다음의 <표 7>은 3장의 <표 6>에서 제시한 10개의 데이터 리포지터리와 EcoBank를 비교한 것이다. 비교 기준은 <표 6>의 서비스 운영현황의 세부 내용을 포함하였다. 세부 내용은

〈표 6〉 생태분야 데이터 리포지터리 현황 분석

| 구분 리포지터리 | 국가 | 원딩기관 | 메타데이터 표준 | 영구 식별자 | 운영현황 | | 기타 |
|---------------------|-----|--------------|--|--|--|---|--------|
| | | | | | 서비스 현황 | 데이터 현황 | |
| GBIF | 덴마크 | 글로벌 네트워크 | Dublin Core EML ABCD | DOI | <ul style="list-style-type: none"> • 데이터 paper 및 해버나 제공 • GBIF API • 데이터 정책: 개인정보보호 정책, 인용지침 등 • 도구: 출판 톨킷(IPT 등)/데이터 접근 및 사용/GBIF LABS • 데이터셋(품질) 지원 서비스 • 데이터 출판 서비스 • 랠리즈 노트 | <ul style="list-style-type: none"> - 레코드: 1,900,000,000 - 데이터셋: 60,000 | WDS 인증 |
| BExIS | 독일 | DFG | Repository-Developed Metadata Schemas EML Darwin Core | DOI | <ul style="list-style-type: none"> • 시스템 지원 <ul style="list-style-type: none"> - 자원 예약, 문서 업로드, 이벤트 등록 시스템 - 포도 갤러리/그래픽 연구 영역 관리 - 기후 데이터 임베딩 시스템/토지 이용의 중요 매개 변수 계산 • 공개 자료: 출판물, 용어집, 공공데이터 • 도구: 기후 도구(ClimateTool)/LUI 계산 도구 | <ul style="list-style-type: none"> - 데이터셋: 1,500 이상 - 출판물: 900개 이상 | |
| EDI Data Portal | 미국 | NSF | Dublin Core OAI-OR-EML | DOI | <ul style="list-style-type: none"> • 리소스: 리포지터리를 이용하는 이용자 특성에 따른 정책 안내 • 데이터 정책 <ul style="list-style-type: none"> - 용어 정의/데이터 패키지 접근성/민감 데이터 - 오픈라인 데이터/데이터 기여자의 지적 권리 • 언어 국제화 정책 • 환경 및 생태 연구 커뮤니티를 위해 설계된 도메인별 데이터 리포지터리 운영 • 데이터 관리 및 출판 서비스 • 데이터 패키지 DOI 발급 • 데이터 큐레이터, FAIR 기준의 준수 여부 평가 • 교육: 웨비나, 워크숍, EDI 정보 관리자의 개별 지원을 통해 데이터 보관 프로세스와 데이터 관리 모범 사례에 대한 교육 제공 • 도구 지원: ezEML(메타데이터 편집기) 등 | <ul style="list-style-type: none"> - 디지털 데이터 파일: 약 405,000개 | CTS 인증 |
| KNB Data Repository | 미국 | NCEAS NSF | Dublin Core EML | PURL URN ARK DOI ORCID id | <ul style="list-style-type: none"> • 데이터 정책 <ul style="list-style-type: none"> - 데이터 기탁 및 라이선스, 데이터 보존 정책 • 도구 <ul style="list-style-type: none"> - MetacatUI: 웹 인터페이스/DataONE: KNB 및 기타 - Metacat: 메타데이터 DB - EML: 생태학 메타데이터 표준 - KNB 개발자 도구: shell, R, Matlab, Python 등 • 메타데이터 품질 확인: MetaDig(품질엔진) 활용 | <ul style="list-style-type: none"> - 데이터셋: 28,983 - 메타데이터: 28,983 - 데이터: 65,393 - 복제본(replicas): 15,779 | |

| 구분 리포지터리 | 국가 | 편입기관 | 메타데이터 표준 | 영구 식별자 | 운영현황 | | 기타 | |
|--|------------|--------------------------|-----------------------------|-------------|--|---|---|----------------|
| | | | | | 서비스 현황 | 데이터 현황 | | |
| VertNet | 미국 | NSF | Darwin Core | URN | <ul style="list-style-type: none"> VertNet IPT(Integrated Publishing Toolkit) 출판된 생물다양성 데이터 호스팅 리소스 저작권 및 라이선스 지침 등 데이터셋, 도구 및 코드 교육 워크샵, 지오레퍼런스/퍼블리케이션 & 비디오 | <ul style="list-style-type: none"> - 레코드: 21,600,262 - 데이터 리소스: 316 - 컬렉션: 402 | <ul style="list-style-type: none"> - 전체 기록: 1,900 - 데이터 컬렉션: 141 - 데이터셋: 1,700 - 지도: 43 | CTS /WDS 인증 |
| EIDC | 영국 | NERC | Dublin Core ISO 19115 | DOI | <ul style="list-style-type: none"> 데이터 정책: 기탁 보존, DOI 정책 및 NERC 데이터 정책 교육: 'Data tree'(무료 온라인 과정) 연구데이터 관리와 비즈니스 정책 입안자, 미디어 및 더 많은 대중과 데이터를 참여시키고 공유하는 방법을 다룸 | <ul style="list-style-type: none"> - 데이터셋: 321 - 데이터 파일: 12,726 | | |
| EcoDB (Science Data Bank) | 중국 | CAS | Dublin Core | DOI | <ul style="list-style-type: none"> 출판 프로세스 데이터 정책 신호하는 파일 형식/저널 데이터 정책에 대한 맞춤 서비스 데이터 유형, 데이터 검토 기준 및 라이선스 계약 등 | <ul style="list-style-type: none"> - 레코드: 6,349,902 - 컬렉션: 87 - 데이터셋: 106 - 이미지: 186,144 | | |
| Canadaensis repository | 캐나다 맨마크 | CFI | Darwin Core | DOI PURL | <ul style="list-style-type: none"> 디지털화: 이미징 데이터, 이미징 표본/지리참조/데이터 정리 데이터 사용 및 출판 규범 더워이어(DwC)/통합 톨킷(IPT) VASCAN: 카탈로그 데이터 품질 요구사항 안내 도구: 좌표변환, 날짜 구분 분석/GBIF API | <ul style="list-style-type: none"> - 레코드: 25,211,059 - 데이터셋: 2,676 - 종 시트: 5,771 - 이미지: 857,859 | | |
| SiB Colombia Portal de Datos | 콜롬비아 | 글로벌 네트워크 | Darwin Core Plinian Core | DOI | <ul style="list-style-type: none"> 공유 데이터 유형/표준/출판 가이드/기록 리소스 출판 도구/오픈 액세스/인용 가이드/메타데이터 표준 데이터 품질: GBIF 데이터 유효성 검사 도구 활용 등 교육: 생물다양성에 관한 데이터의 출판, 접근 및 사용 | <ul style="list-style-type: none"> - 관찰 데이터: 48,024,630 - 종 데이터: 46,433 - 컬렉션: 610 - 물리적 표본: 5,222,013 - 디지털 표본: 359,736 | | |
| Finnish Biodiversity Information Facility | 핀란드 | Partners of LAJIFI | Darwin Core FGDC/CSDGM | PURL | <ul style="list-style-type: none"> 종에 관한 검색 및 목록 관측 데이터 품질관리 도구 iNaturalist Suomi: 모바일 앱으로 사진 찍기, 즉시 신원 확인 기능, 관찰 내용 추적 등의 기능을 제공하는 어플리케이션 마이 노트(my note): 관찰 내용 추적, 활성 관찰자를 위한 도구 Data Bank(데이터 저장) 데이터 관리 데이터셋 및 사용/데이터 품질/민감 데이터/체크리스트 FinBIF 출판물 | | | |

〈표 7〉 생태 분야 데이터 리포지터리 비교 분석

| 카테고리 | 리포지터리 | GBIF | BExIS | EDI | KNB | VertNet | EIDC | EcoDB (ScienceDB) | Canadensys repository | SiB Colombia | FimBIF | EcoBank | Total |
|-----------|--------------|------------|-------|------------|-----|---------|----------------|----------------------|--------------------------|-----------------|--------|---------|-------|
| 서비스 현황 | 데이터 정책 | 0 | | 0 | 0 | 0 | 0 | 0 | | 0 | 0 | | 8 |
| | 데이터 품질 체크 | 0 | | 0 | 0 | | | | 0 | 0 | 0 | | 6 |
| | 데이터 교육 및 워크숍 | 0 | | 0 | | 0 | 0 | | | 0 | | | 5 |
| | 용어집 | | 0 | | | | | | | | | | 1 |
| | 출판 | 0 | 0 | 0 | | | | 0 | 0 | 0 | 0 | 0 | 9 |
| | 도구 및 API | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 |
| 리포지터리 인증 | 지도 서비스 | 0 | | 0 | 0 | 0 | | | 0 | 0 | 0 | 0 | 8 |
| | | 0 (WDS) | | 0 (CTS) | | | 0 (CTS/WDS) | | | | | | 3 |

데이터 정책, 데이터 품질 체크, 데이터 교육 및 워크숍, 용어집, 출판, 도구 및 API, 지도 서비스 그리고 리포지터리 인증 여부를 포함한다.

첫 번째, 서비스 현황 중 '데이터 정책'을 웹 페이지에 게시한 리포지터리는 8개(GBIF, EDI, KNB, VertNet, EIDC, EcoDB, SiB Colombia, FinBIF)로 나타났다. GBIF에서는 데이터 인용 정책이 안내되어 있으며, EDI는 관리자 지원에 있어서 데이터 정책을 게시하고 있는데, DMP(데이터관리계획) 정책, 데이터 품질 정책, 데이터 패키지 관련 정책 및 민감 데이터 정책 등을 웹페이지에 게시하고 있다. KNB의 경우 데이터 기탁 및 라이선스 정책, 데이터 보존 정책 등이 안내되어 있다. VertNet에는 저작권 및 라이선스 지침, 데이터 사용 및 공개에 대한 VertNet 규범 등이 있다. EIDC에서는 기탁 정책, 보존 정책, DOI 정책, 데이터 관리 계획 및 템플릿, NERC 데이터 정책 등이 웹페이지에 게시되어 있다. EcoDB의 데이터 정책에는 데이터 유형 범위, 데이터 커뮤니티, 데이터 제출자가 준수해야 할 행동강령, 데이터 검토 기준, 데이터 라이선스 계약, 데이터 참조 사양, 데이터 철회 등에 대한 정책들이 게시되어 있다. SiB Colombia에는 출판 정책, 데이터 인용 정책 등이 게시되어 있다. FinBIF의 데이터 정책에는 데이터 인용 정책, 데이터 품질 정책, 민감 데이터 정책 등이 웹페이지에 게시되어 있다.

두 번째 서비스는 '데이터 품질 체크' 서비스이다. 이는 총 6개 리포지터리(GBIF, EDI, KNB, Canadensys repository, SiB Colombia, FinBIF)에서 시행되는 서비스로 데이터 기탁 시, 데이터 품질을 체크함으로써 고품질의 데이터를 수

집하는데 유용하다. GBIF는 새로운 데이터셋을 '색인화'하거나 '수집'하여 사용자가 공통 검색 및 다운로드 서비스를 통해 모든 데이터를 검색할 수 있는 공통 접근 시스템에 이를 통합하며, 데이터셋이 색인화되면 추가 검사, 해석 및 변환 루틴을 수행하여 데이터가 상호 운용 가능하고 데이터 형식, 데이터 품질 및 사용 적합성에 대한 최소 표준을 준수하는지 확인한다. EDI는 정보관리자 지원에서 데이터 품질에 대한 가이드를 통해 데이터 품질 보장 및 데이터 품질 관리를 실행하고 있다. KNB의 경우, 데이터 기탁 시, MetaDig를 활용하여 메타데이터 품질을 평가한다. 메타데이터 레코드의 완전성을 평가하려면 데이터셋을 제출한 후 데이터셋 랜딩 페이지의 데이터셋 인용 바로 아래에 있는 "품질 보고서" 버튼을 클릭하여 MetaDig 품질 엔진이 데이터셋에 대해 자동으로 수행하는 일련의 검사 결과를 표시하게 한다. 또한 데이터셋 문서를 개선하는 방법으로 이용자가 이 자동화된 피드백을 활용하도록 권장한다. Canadensys repository는 데이터 게시를 위한 7단계 가이드에서 다양한 데이터 세트 클래스에 따른 데이터 품질 요구사항 및 권장사항을 안내하고 있다. 이러한 데이터 품질 요구사항 및 권장사항을 따르면 데이터 게시자는 데이터셋의 품질, 완전성 및 가치를 향상시킬 수 있다. SiB Colombia는 데이터 품질 향상을 위해서 생물다양성 데이터의 검증, 구조화, 정리를 위한 유용한 문서와 도구를 사용하고 있다. 데이터 정리와 구조화에는 OpenRefine, 생물 다양성 데이터 품질을 위한 스크립트(Scripts for biodiversity Data Quality), 데이터 유효성 검사(Data Validator) 등의 도구를 제공하고 있으며, 지리정보 검증을 위해서

는 Georeferencing Calculator Manual와 같은 도구를 안내하고 있다. FinBIF의 경우, FinBIF가 편집한 데이터에는 품질에 대한 추가 표시를 추가하는 주석이 포함되는 경우가 있으며, 주석은 원본 데이터에 추가되지만 원래 제공된 정보를 대체하지는 않는다. FinBIF는 주석을 기반으로 한 데이터를 제거하거나 숨기지 않지만 데이터 사용자가 의도한 목적에 가장 적합한 데이터를 찾을 수 있도록 주석을 기반으로 데이터를 필터링할 수 있는 수단을 제공한다. 또한 품질 필터를 통해 “품질” 하위 제목 아래에는 결과를 필터링할 수 있는 여러 가지 품질 기반 필터를 제공하여 데이터의 품질을 체크한다. FinBIF에는 모든 새로운 기록에 적용되는 자동화된 데이터 검증이 포함되어 있으며, 자동화된 검증 중에 오류가 발견되면 기록에 ‘품질 문제’가 있는 것으로 표시되어 데이터 품질을 체크한다.

세 번째 서비스는 ‘데이터 관리 교육 및 워크숍’에 대한 지원으로 5개의 리포지터리에서 시행되고 있다. GBIF는 강사 및 멘토 커뮤니티와 협력하여 생물다양성 데이터 동원 및 사용에 관한 커리큘럼을 개발하였으며, 이러한 과정은 스스로 학습할 수 있도록 온라인으로 제공되거나 현장 또는 가상 워크숍에 맞게 조정될 수 있다. 또한 ‘GBIF Data Use Club’은 DataCamp Donates 프로그램을 통해 DataCamp와 협력하여 제공되는 무료 라이선스를 통해 숙련된 데이터 사용자, 온라인 교육 리소스 및 데이터 활용 능력 구축 역량을 갖춘 세미나를 통해 전 세계적으로 GBIF 매개 데이터(GBIF-mediated data) 사용을 지원하기 위한 프로그램으로 교육적 역할을 한다. EDI는 데이터 출판 모범 사

례, 도구 및 새로운 개발에 대한 고유한 주제와 반복되는 주제를 제시하며 정기적인 웹 세미나를 통해 데이터 관리 등에 대한 교육을 실시하고 있다. VertNet에서는 지리 참조 및 생물 다양성 정보학 분야의 교육 워크숍을 개발하였다. 이러한 워크숍의 목표는 컬렉션과 연구데이터의 품질과 완전성을 향상시키는 데 필요한 모범 사례, 도구, 작업 흐름 및 프로토콜을 사용하여 생물다양성 과학 및 자연사 커뮤니티 구성원을 교육하는 것으로, 워크숍에는 흥미로운 교육, 실습, 심층 토론이 포함된다. EIDC는 NERC에서 예산을 지원하는 무료 온라인 과정인 ‘Data tree’를 제공하는데, 이는 연구데이터 관리와 비즈니스, 정책 입안자, 미디어 및 더 많은 대중과 데이터를 참여시키고 공유하는 방법을 다룬다. SiB Colombia Portal de Datos에서는 생물다양성 데이터의 출판 및 관리를 위한 오프라인 워크숍을 진행하며, 워크숍에 대한 큰 수요와 지역사회의 높은 관심으로 가상 구성 요소를 갖춘 교육 전략을 수립하게 되었다. 또한 더 많은 사람과 지역, 교육 주제 등을 네트워크의 관심 주제로 다루도록 확장하였다.

이외에도 용어집을 제공하는 리포지터리가 1개(BExIS) 있었으며, WDS(World Data System) 및 CTS(CoreTrustSeal) 인증을 받은 리포지터리(GBIF, EDI, EIDC)가 3개로 도출되었다.

이를 종합해 볼 때, EcoBank에서 필요한 서비스는 ‘데이터 정책’, ‘데이터 품질 체크’, ‘데이터 관리 교육 및 워크숍’ 등이다. ‘데이터 정책’을 게시하여 이용자들에게 서비스한다면 EcoBank에 대한 수집, 관리, 보존, 공유 및 활용에 대한 데이터 관리에 대한 신뢰도가 높아질 것으로 판단된다. 또한 ‘데이터 품질 체크’의 경우, 데

이터 기탁 시 데이터에 대한 품질 체크가 이루어져서 데이터 오류를 발견하고, 고품질의 데이터를 수집 및 제공할 수 있게 된다. 이와 함께 '데이터 관리 교육' 등 교육 프로그램을 운영하고 '용어집'을 제공함으로써 이용자들이 데이터 관리에 대한 올바른 이해와 데이터를 다루는데 도움이 될 것이다.

5. 논의 및 결론

데이터 리포지터리는 데이터의 수집, 저장, 보존 및 재사용을 위한 기본 도구이다. 오픈 사이언스 시대에서 데이터 공유 및 재사용은 이미 필수가 되었다. 이러한 시점에서 데이터 리포지터리의 활용은 학문 커뮤니티, 연구기관 그리고 국가를 중심으로 활발하게 구축 및 운영되고 있다.

본 연구의 목적은 국립생태원이 운영하고 있는 EcoBank의 현황을 파악하고 해외 생태 분야 데이터 리포지터리를 분석하여 EcoBank에 필요한 서비스 방안을 제안하는 것이다. 연구 목적을 달성하기 위하여 re3data 내의 10개의 생태 분야 데이터 리포지터리를 선정하였으며 각각을 분석하여 EcoBank와 비교하였다. 비교한 결과 다음의 3가지 내용에 대한 서비스 제안 사항이 도출되었다.

- (1) 데이터 정책: 연구데이터 관리를 위한 데이터 정책 및 가이드라인이 해외 데이터 리포지터리의 경우 해당 사이트에 게시하여 이용자들에게 서비스하고 있다. 이러한 데이터 정책은 연구자와 일

반 이용자를 구분하여 구성될 필요가 있으며 또한 데이터 생명 주기에 맞춘 정책 및 가이드라인으로 구현하여 서비스 안내가 이루어져야 할 것이다.

- (2) 데이터 품질 검토: 데이터 형식, 품질 및 적합성 확보를 위한 최소 표준 준수 여부 체크, 메타데이터 품질, 데이터 정리 및 구조화를 위한 도구 제공 그리고 데이터를 필터링하기 위한 도구 지원 등이 EcoBank를 통해 향후 추가되어야 할 것으로 판단된다.
- (3) 데이터 관리 교육 및 워크숍: 학문 커뮤니티별로 데이터 관리가 차별화되어 이용자에게 제공되어야 한다. 특히, 생태학 분야는 타 학문분야보다 복잡성과 다양성을 지니고 있어 엄밀한 데이터 관리 교육 등이 제공되어야 할 것이다. 서비스 제안 사항으로 제시된 품질 검토도 오프라인 및 온라인 교육 과정을 개설하여 연구자 또는 일반 이용자를 대상으로 제공된다면 데이터 품질 확보도 어느 정도 달성할 수 있을 것이다.

이외 현재 EcoBank는 re3data.org에 등록되지 않았다. 글로벌한 데이터 유통을 위한 또는 기관의 리포지터리 홍보를 위해서는 re3data와 같은 레지스트리의 등록은 필수이며 해당 사이트를 통한 국립생태원의 위상도 높일 수 있을 것으로 파악된다. 또한, 데이터 리포지터리의 인증 획득도 EcoBank가 데이터의 신뢰성과 품질을 확보할 수 있는 수단이 될 수 있다. 현재, re3data에 등록된 생태분야(식물생태학 및 생태계 분석 및 동물생태학, 생물다양성 및 생

태계 연구)를 기준으로 데이터 리포지터리 인증을 받은 리포지터리는 단 3곳으로 파악되었다. 추후 EcoBank가 데이터 리포지터리 인증을 준비하는 것도 서비스 개선을 확보할 수 있는 중요한 핵심 도구가 될 것이다.

제안된 서비스의 경우 단순 해외 데이터 리포지터리 현황을 분석하여 제시한 것으로 전문가 검토 및 현장 적용 가능성에 대한 논의가 이루어지지 않았다는 점은 본 연구의 한계라고 볼 수 있다. 하지만 생태 분야 데이터 리포지터리를 조사하고 분석하여 도출된 내용은 향후 EcoBank의 서비스를 개선하기 위한 기초자료

로 활용될 수 있다고 판단된다.

추후 본 연구에서 제안된 서비스 제안 사항이 EcoBank에 적용되기 위한 전략 방안에 대해 논의될 필요성이 있다고 본다. 즉, 데이터 정책의 경우 생태분야의 데이터 정책 및 가이드라인 사례를 파악하여 EcoBank에 적용할 수 있는 연구와 데이터 품질을 개선할 수 있는 시스템적인 도구를 어떻게 구현해야 할지에 대해서 연구가 필요할 것이다. 본 연구를 통해 국립생태원의 EcoBank 서비스를 개선하기 위한 전략 방안 도출에 도움이 되길 희망해 본다.

참 고 문 헌

- 과학기술정보통신부 (2023. 9. 27.). '국가연구데이터 관리 및 활용 촉진에 관한 법률' 제정안 입법예고. 출처: <https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&pageIndex=&bbsSeqNo=94&nttSeqNo=3183524&searchOpt=ALL&searchTxt=>
- 국립생태원 (2023. 9. 18.). EcoBank 소개. 출처: <https://www.nie-ecobank.kr/cmmn/about.do>
- 국립생태원 (2023. 9. 18.). EcoBank 활용 가이드북. 출처: <https://www.nie-ecobank.kr/participation/edu/EduListDetailInquire.do>
- 김예현 (2023). 과학기술분야 기관 연구데이터 리포지터리 운영 활성화 방안 연구. 한국비블리아학회지, 34(3), 109-134. <https://doi.org/10.14699/KBIBLIA.2023.34.3.109>
- 김우중, 김형준, 박희진, 최상욱 (2021). 자연과학 분야 연구데이터 리포지터리 속성 분석: 운영수준, 주제, 유형, 활용도를 중심으로. 기술혁신학회지, 24(4), 577-595. <https://doi.org/10.35978/jktis.2021.2.24.1.577>
- 김주섭, 윤희남, 권용수, 김선태 (2020). 생태 분야 연구데이터를 위한 메타데이터 설계: DCAT을 중심으로. 한국도서관·정보학회지, 51(4), 249-278. <https://doi.org/10.16981/KLISS.51.4.202012.249>
- 안병균, 박규리, 심지우, 조형민 (2018). 연구데이터 국가 리포지터리 관리에 관한 연구: 국내외 대표 연구데이터 리포지터리 홈페이지 메타데이터를 중심으로. 한국정보관리학회 2018년도 제25회

- 학술대회 논문집, 117-126.
- 허태상 (2016). 장기생태연구정보의 전주기 관리를 위한 데이터 플랫폼. 박사학위논문, 배재대학교 대학원 컴퓨터공학과.
- Alves, C., Castro, J. A., Ribeiro, C., Honrado, J. P., & Lomba, Ângela (2018). Research data management in the field of Ecology: an overview. International Conference on Dublin Core and Metadata Applications, 87-94. Available: <https://dcpapers.dublincore.org/pubs/article/view/3965>
- Bex IS. Available: <https://www.bexis.uni-jena.de/>
- Canadensys repository. Available: <https://data.canadensys.net/ipt/>.
- Easterday, K., Paulson, T., DasMohapatra, P., Alagona, P., Feirer, S., & Kelly, M. (2018). From the field to the cloud: a review of three approaches to sharing historical data from field stations using principles from data science. *Frontiers in Environmental Science*, 6, 88.
- EDI Data Portal. Available: <https://portal.edirepository.org/nis/home.jsp>
- Environmental Information Data Centre. Available: <https://eidc.ac.uk/>.
- Finnish Biodiversity Information Facility. Available: <https://laji.fi/en>.
- GBIF. Available: <https://www.gbif.org/>.
- Gries, C., Hanson, P. C., O'Brien, M., Servilla, M., Vanderbilt, K., & Waide, R. (2023). The environmental data initiative: connecting the past to the future through data reuse. *Ecology and Evolution*, 13(1), e9592. <https://doi.org/10.1002/ece3.9592>
- Kim, Hyun Woo, Yoon, Sungsoo, Kim, Mokyoung, Shin, Manseok, Yoon, Heenam, & Kim, Kidong (2021). EcoBank: A flexible database platform for sharing ecological data. *Biodiversity Data Journal* 9: e61866. Available: <https://doi.org/10.3897/BDJ.9.e61866>
- Kim, Suntae (2021). Analysis of the current status of data repositories in the field of ecological research. *Proceedings of the National Institute of Ecology of the Republic of Korea*, 2(2), 139-143. <https://doi.org/10.22920/PNIE.2021.2.2.139>
- KNB Data Repository. Available: <https://knb.ecoinformatics.org/>.
- National Ecosystem Data Bank. Available: <https://ecodb.scidb.cn/en>.
- Pampel, H., Weisweiler, N. L., Strecker D., Witt, M., Vierkant, P., Elger, K., Bertelmann, R., Buys, M., Ferguson, L. M., Kindling, M., Kotarski, R., & Petras, V. (2023). Re3data-indexing the global research data repository landscape since 2012. *Scientific Data*, 10(1), 571.
- SiB Colombia Portal de Datos. Available: <http://biodiversidad.co/>.
- VertNet. Available: <http://vertnet.org/>.

• 국문 참고자료의 영어 표기

(English translation / romanization of references originally written in Korean)

- An, Byoung-Goon, Park, Kyu-Ri, Shim, Jiwoo, & Cho, Hyungmin (2018). A study on national research data repository management. Proceedings of the 69th the Korean Society for Information Management Conference, 117-126.
- Heo, Taesang (2016). Data Platform for Full Life-cycle Management of Long-term Ecological Research Information. Doctoral dissertation, Department of Computer Engineering, Pai Chai University Graduate School.
- Kim, Juseop, Kwon, Yongsu, Kim, Sun-Tae, & Yoo, Heenam (2020). Metadata design for ecological research data: focused on DCAT. Journal of Korean Library and Information Science Society, 51(4), 249-278. <https://doi.org/10.16981/KLISS.51.4.202012.249>
- Kim, Woojoong, Kim, Hyungjoon, Park, Heejin, & Choi, Sangok (2021). An analysis on operational level, subjects, content types and utilization of Research Data Repositories in Natural Science. Journal of Korea Technology Innovation Society, 24(4), 577-595. <https://doi.org/10.35978/jktis.2021.2.24.1.577>
- Kim, Ye Hyeon & Kim, Jihyun (2023). A study on strategies to promote the activation of institutional research data repositories in the field of science and technology. Journal of the Korean BIBLIA Society for Library and Information Science, 34(3), 109-134. <https://doi.org/10.14699/KBIBLIA.2023.34.3.109>