re3data.org를 활용한 Geoscience 분야 데이터 리포지터리 평가 및 사례 연구*

Evaluation and Case Study of Geoscience Data Repositories Using re3data.org

김 주 섭 (Juseop Kim)**
김 선 태 (Suntae Kim)***

목 차

초 록

연구데이터 공유 및 재사용을 위하여 데이터 리포지터리가 연구기관, 커뮤니티 그리고 국가를 중심으로 운영되고 있다. 현재, re3data.org에는 데이터 리포지터리가 3,236개가 등록되어 있으며 해당 리포지터리는 각 주제별로 운영되고 있다. 본 연구의 목적은 Geoscience 분야 데이터 리포지터리의 운영 현황을 파악하여 해당 분야의 리포지터리 서비스를 제공하기 위함이다. 연구 결과, Geoscience 분야 데이터 리포지터리 890개 중 634개가 9개 속성 중 4개 이상을 만족하고 있는 것을 나타나 약 74%의 리포지터리가 보편적으로 잘 운영되고 있다는 것을 확인하였다. 또한 해당 리포지터리에서 서비스를 확인한 결과 데이터 정책, 큐레이션, 도구 및 API 그리고 품질관리 측면에서 다양한 시사점을 제공하는 것으로 파악되었다. 이러한 연구 결과는 앞으로 Geoscience 분야 데이터 리포지터리 서비스 구성 시 참고할 수 있는 기초가 될 것이다.

ABSTRACT

In order to share and reuse research data, data repositories are operated mainly by research institutes, communities, and countries. Currently, there are 3,236 data repositories registered on re3data.org, and the repositories are operated by each subject. The purpose of this study is to identify the operational status of data repositories in the Geoscience field and provide repository services in the field. As a result of the study, 634 out of 890 data repositories in the Geoscience field satisfied more than 4 out of 9 properties, confirming that approximately 74% of the repositories are generally well operated. In addition, as a result of checking the services in the repositories, it was found that they provide various implications in terms of data policy, curation, tools and APIs, and quality management. These research results will serve as a basis for reference when configuring data repository services in the Geoscience field in the future.

키워드: re3data.org, 데이터 리포지터리, Coretrustseal, Geoscience, Master Data Repository
re3data.org, Data Repository, Coretrustseal, Geoscience, Master Data Repository

* 이 논문은 2024년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.
** 전북대학교 문헌정보학과 강사, 연구데이터융복합연구소 전임연구원
(kimjuseop@jbnu.ac.kr / ISNI 0000 0004 7492 1806) (제1저자)
*** 전북대학교 문헌정보학과 부교수, 연구데이터융복합연구소장
(kim.suntae@jbnu.ac.kr / ISNI 0000 0004 6492 6355) (교신저자)
논문접수일자: 2024년 7월 26일 최초심사일자: 2024년 8월 3일 게재확정일자: 2024년 8월 16일
한국문헌정보학회지, 58(3): 169-191, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.3.169

1. 서 론

1.1 연구 목적 및 연구 질문

데이터 공유 및 재사용을 위해 각 커뮤니티 및 학문 분야에서는 도메인 분야의 특성을 반영한 데이터 리포지터리를 활용하고 있다. 이러한 데이터 리포지터리를 등록하는 사이트로 re3data.org가 활용되고 있다. 또한 데이터 인용 관리를 위하여 특정 출판사에서는 데이터 리포지터리 목록도 만들어 이용자에게 제공하고 있다(re3data.org, 2024; Clarivate, 2024).

re3data.org는 데이터 리포지터리의 주제, 국가, 운영 및 펀딩 기관, 식별자, 메타데이터 그리고 리포지터리 인증 여부까지 확인할 수 있도록 구성되어 있다. 이러한 속성을 통해 주제 분야의 데이터 리포지터리 운영 현황을 평가할 수 있다(김우중 외, 2021). 특히 특정 주제 분야의 데이터 리포지터리에 접근할 수 있으며 해당 리포지터리를 통해 데이터 공유 및 재사용이 가능하다는 점에서 re3data.org는 데이터 리포지터리의 개요 및 현황을 확인하는 데 있어 효율적이라고 볼 수 있다.

따라서 본 연구에서는 re3data.org를 통해 Geoscience 분야의 데이터 리포지터리의 운영 현황을 평운영을 평가하고 우수하다고 판단되는 데이터 리포지터리 서비스를 파악하여 Geoscience 분야의 데이터 리포지터리 서비스를 제안하고자 한다. 본 연구의 목적을 달성하기 위하여 다음과 같이 세 가지의 연구질문을 설정하였다.

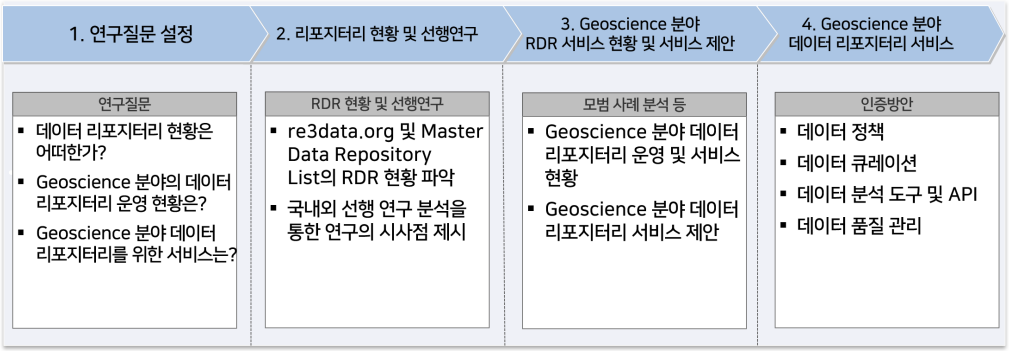
- 연구질문 1: 연구데이터 리포지터리 현황은 어떠한가?
- 연구질문 2: Geoscience 분야 데이터 리포지터리의 운영 현황은 어떠한가?
- 연구질문 3: Geoscience 분야 데이터 리포지터리를 위한 서비스는 어떻게 구성될 수 있는가?

1.2 연구 방법

본 연구의 핵심 목적은 re3data.org에 등록된 Geoscience 분야 데이터 리포지터리의 운영 현황을 분석하고 각 플랫폼의 기능을 살펴본 후 해당 도메인 분야의 리포지터리에 적용가능한 서비스를 제안하기 위함이다. 따라서 본 연구 목적을 달성하기 위하여 다음의 〈표 1〉과 〈그림 1〉과 같은 연구범위와 프로세스를 수립하였다.

〈표 1〉 연구 범위

단계	내용
1단계	• 연구범위, 연구질문 및 연구 수행 전략 도출 • 데이터 리포지터리 현황 파악(re3data.org, 마스터 데이터 리포지터리 목록)
2단계	• re3data.org에 등록된 Geoscience 분야 데이터 리포지터리 운영 현황 분석 • 분석 대상 12개의 데이터 리포지터리 도출
3단계	• Geoscience 분야 데이터 리포지터리 서비스 분석 • 데이터 리포지터리 서비스 분석을 통한 공통 서비스 도출
4단계	• Geoscience 분야를 위한 데이터 리포지터리 서비스 제안



〈그림 1〉 연구 프로세스

1단계에서는 본 연구의 목적을 달성하기 위한 범위, 구체적인 연구 질문을 제시하였다. 2단계에서는 리포지터리 현황을 살펴보기 위하여 re3data.org와 Clarivate사의 마스터 데이터 리포지터리 목록에서 주제별 및 국가별 현황을 살펴보았다. 또한 해당 단계에서 선행연구를 파악하여 본 연구와의 차이점을 제시하였다. 3단계에서는 Geoscience 분야 데이터 리포지터리 운영 및 서비스 현황을 파악하였다. 리포지터리 운영 현황을 파악하기 위하여 re3data.org의 속성 9가지를 대상으로 평가기준을 삼았다. 이러한 평가기준으로 890개의 Geoscience 분야의 데이터 리포지터리 운영 현황을 평가하였다. 평가한 결과 중 Geoscience분야 데이터 리포지터리 12개를 선정하여 해당 리포지터리의 서비스를 분석하였다. 분석한 결과를 바탕으로 4가지의 Geoscience 분야 데이터 리포지터리 서비스를 제안하였다. 해당 서비스에는 데이터 정책, 데이터 큐레이션, 데이터 분석 도구 및 API 그리고 데이터 품질 관리가 포함되었다. 본 연구를 통해 도출된 Geoscience 분야 데이터 리포지터리 서비스는 해당 도메인 분야의 데이터 공유 및 관리를 위한 핵심 내용으로 활용될 것으로 판단된다.

2. 이론적 배경

이번 장에서는 데이터 리포지터리 현황을 re3data.org와 Clarivate사의 마스터 데이터 리포지터리 목록을 통해 살펴보고 관련된 선행연구를 확인하여 본 연구와의 차이점을 살펴보고자 한다.

2.1 re3data.org의 연구데이터 리포지터리 현황

현재, re3data.org에 등록된 연구데이터 리포지터리는 3,236개이다. 해당 사이트에서는 인문학 및 사회과학, 생명과학, 자연과학 그리고 공학 4개의 카테고리 하위에 14의 중분류를 통해 주제별 접근을 제공하고 있다. 또한 14개의 중분류는 48개의 소분류로 구성되어 있으며 해당 분류는 또한 204개의 세부분류로 구성되어 있다. 리포지터리마다 1개씩 주제가 할당되는 것이 아닌 복수의 주제를 해당 운영 주체에서

선택해서 등록한다. re3data.org는 일반적으로 레지스트리 사이트이므로 리포지터리 운영주체가 속성 값을 입력하도록 되어 있다 해당 속성은 오픈액세스, 라이선스, 인증, 정책, 영구식별자, 저자식별자, 품질관리, 버전관리 및 메타데이터 등 9개 이상의 속성 값을 등록할 수 있도록 되어 있으며 리포지터리 품질과 관련된 인증을 선택하여 등록할 수 있도록 되어 있어 리포지터리에 대한 신뢰성 여부도 확인할 수 있도록 제공하고 있다(re3data.org, 2024).

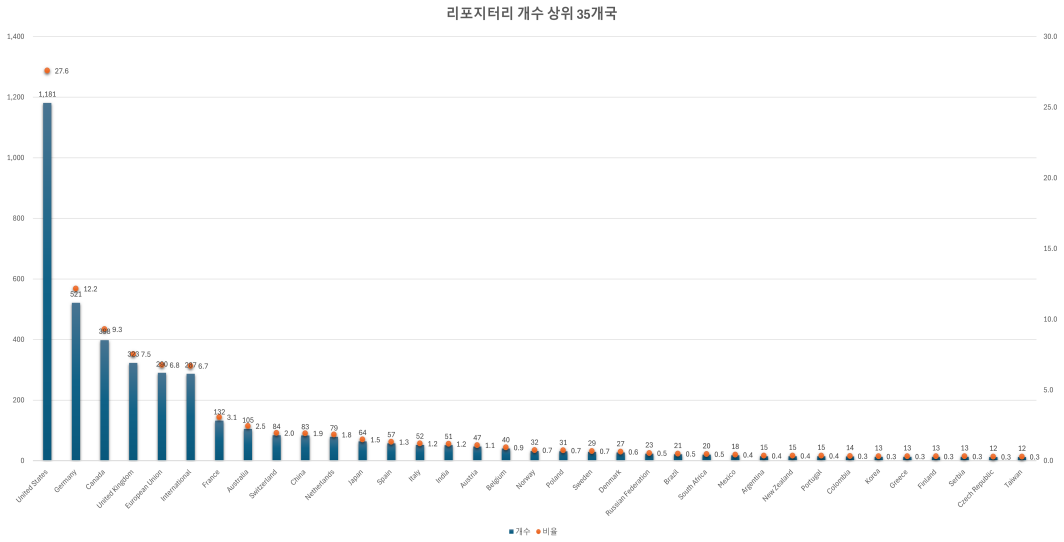
다음의 〈표 2〉는 re3data.org에 등록된 데이터 리포지터리 주제 분류 현황을 나타낸 것이다.

중복으로 인하여 주제별 총 데이터 리포지터리 개수(5,452개)는 등록된 리포지터리 수(3,236개)보다 많음을 감안하더라도 가장 많은 비중을 차지하고 있는 것은 생명과학이며 그 뒤를 이어 자연과학이 그리고 인문학 및 사회과학 마지막으로 공학으로 확인되었다. 생명과학과 자연과학 2개의 대분류를 합하면 약 63%로 전체 데이터 리포지터리의 약 2/3를 차지하고 있다. 다음의 〈그림 2〉는 데이터 리포지터리의 국가별 현황을 그래프로 표현한 것이다.

국가별로 현황을 살펴보면 미국이 전체 1,181개로 36.5%를 차지하고 있으며 다음으로 독일이 521개로 16.1% 그리고 캐나다가 398개로 12.3%, 영국이 323개로 10%로 확인되었다. 미국과 독일이 전체 약 53%로 과반 이상을 차지하고 있음을 알 수 있다. 한국과 중국 그리고

〈표 2〉 데이터 리포지터리 주제별 현황(대-중분류)

대분류	중분류	개수	전체 점유율(%)
인문학 및 사회과학		1,225	22.47
	인문학	365	
	사회 및 행동과학	507	
생명과학		1,850	33.93
	생물학	1,024	
	의학	758	
	농업, 임업, 원예 및 수의학	239	
자연과학		1,664	30.52
	화학	264	
	물리학	346	
	수학	49	
	지주과학(지리학 포함)	891	
공학		713	13.08
	기계 및 산업 공학	14	
	열 공학/공정 공학	24	
	재료 과학 및 공학	55	
	컴퓨터 과학, 전기 및 시스템 공학	179	
	건설 공학 및 건축	58	
Total		5,452	



〈그림 2〉 상위 35개국 데이터 리포지터리 개수 및 비율

일본을 기준으로 비교하면 중국이 83개로 전체 약 2.6% 일본이 64개로 2% 마지막으로 한국이 13개로 파악되었다. 한국이 운영 중인 데이터 리포지터리는 전체 13개로 여기에 국제 컨소시엄과 운영이 중단된 6개의 리포지터리를 제외하면 7개로 확인되었다. 7개 중 한국노동연구원의 Korean Labor & Income Panel Study (KLIPS)와 극지연구소의 Korea Polar Data Center(KPDC)는 마스터 데이터 리포지터리 목록에 등록되어 있다. 여기에 한국과학기술정보연구원의 DataON이 2024년 3월 28일자로 Coretrustseal 인증을 획득하였다.

2.2 Clarivate의 마스터 데이터 리포지터리 목록

re3data.org 이외에 Clarivate사에서 제공하는 마스터 데이터 리포지터리 목록이 있다. 해당 목록에는 총 456개의 데이터 리포지터리가 등록되어 있다. 리포지터리의 등록 기준에는 리포지터리의 지속성과 안정성, 펀딩, 동료 검토, 데이터 보존 그리고 연구 문헌에 대한 링크가 포함된다. 첫째, 지속성과 안정성은 리포지터리의 새로 저장된 데이터가 꾸준히 유입되면 리소스가 활성화되어 있다는 지표로 간주된다. 이건 결국 데이터 인용을 보장하며 등록된 리포지터리의 데이터는 Data Citation Index에 포함된다. 둘째 펀딩의 경우 데이터 출처와 펀딩 정보가 함께 제공되는 리포지터리를 고려 대상으로 삼는다. 여기에는 메타데이터로 설명될 수 있으며 해당 메타데이터와 인용은 영어로 표기되어야 한다. 셋째 동료 검토에서는 보편적인 프로세스는 아니지만 해당 지표가 데이터의 품질과 인용된 참고문헌의 완전성을 나타내는 것으로 파악하고 있다. 넷째 데이터 보존에서는 기탁된 데이터의 연령에 대한 제한은 없는 것으로 제시하며 프로젝트가 종료되더라도 데이터는 계속 인용되고 재사용될

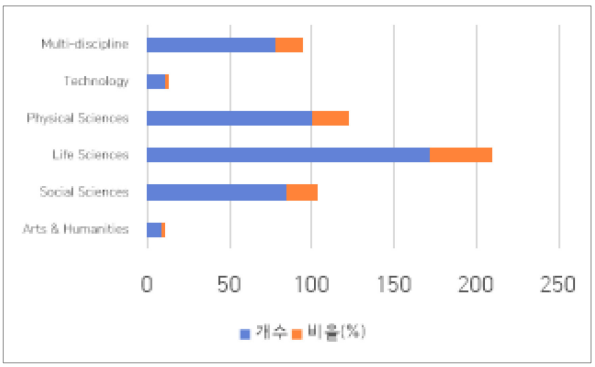
수 있다는 점에 주목하고 있다. 마지막으로 연구 문헌과의 링크이다. 데이터 인용 표준을 홍보하고 학문의 영향을 측정하기 위해 데이터 세트와 재사용한 연구 문헌의 출처를 보여주는 데이터 리포지터리에 우선순위가 주어진다. 해당 목록은 매주 업데이트되며 자체 리포지터리에서 Clarivate사에 등록 요청을 할 수도 있다(Clarivate, 2024).

다음의 〈표 3〉과 〈그림 3〉은 마스터 데이터 리포지터리 목록의 주제별 현황을 나타낸 것이다.

마스터 데이터 리포지터리 목록의 주제별 현황을 살펴본 결과 생명과학과 물리학의 비중이 약 58%로 절반 이상을 차지하고 있으며 그 뒤로 사회과학, 다학제, 기술 등으로 뒤를 잇고 있는 것으로 확인되었다. 다음의 〈그림 4〉와 〈그림 5〉는 마스터 데이터 리포지터리 목록에 포함된 국가별 현황을 나타낸 것이다.

〈표 3〉 주제별 리포지터리 현황(1)

주제	개수	비율(%)
예술 & 인문학 (Arts & Humanities)	9	2
사회과학 (Social Sciences)	85	18.7
생명과학(Life Sciences)	172	37.8
물리학 (Physical Sciences)	100	22
기술(Technology)	11	2.4
다학제(Multi-discipline)	78	17.1
Total	455	100



〈그림 3〉 주제별 리포지터리 현황(2)

<대륙별 합계>

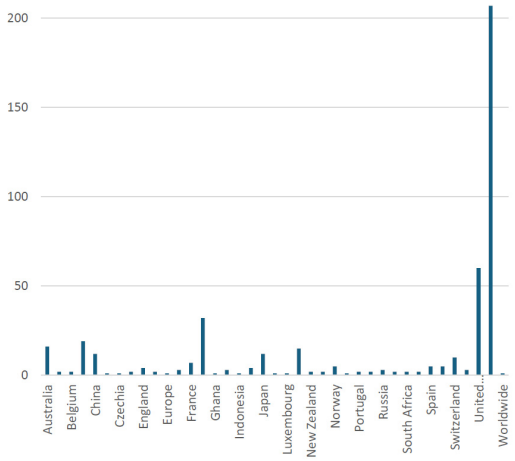
유럽	미주	아시아	오세아니아	아프리카	기타	총 합계
169	226	38	18	3	2	456

* 기타 : 유럽 전체, 세계 전체의 2건

<국가/지역별 개수>

Australia	Austria	Belgium	Canada	China	Croatia	Czechia
16	2	2	19	12	1	1
Denmark	England	Estonia	Europe	Finland	France	Germany
2	4	2	1	3	7	32
Ghana	India	Indonesia	Italy	Japan	Lithuania	Luxembourg
1	3	1	4	12	1	1
Netherlands	New Zealand	Northern Ireland	Norway	Poland	Portugal	Republic of Ireland
15	2	2	5	1	2	2
Russia	Singapore	South Africa	South Korea	Spain	Sweden	Switzerland
3	2	2	2	5	5	10
Taiwan	United Kingdom	USA	Worldwide			
3	60	207	1			

〈그림 4〉 국가/지역별 현황(1)



〈그림 5〉 국가/지역별 현황(2)

마스터 데이터 리포지터리의 국가별 현황에서는 미국이 207개로 가장 많이 확인되었고 다음으로 영국이 60개, 독일이 32개, 캐나다가 19개 순으로 나타났다. 대륙별로 보면 미주가 226개, 우럽이 169개, 아시아가 38개 순으로 확인되었다. 한국의 경우 2개가 해당 목록에 포함되었다.

2.3 선행연구

본 연구의 목적을 달성하기 위해 우선 선행연구에서의 시사점과 본 연구와의 차이점에 대해서 본 절에서 기술하고자 한다.

Khan et al.(2024)은 연구데이터 리포지터리(Research Data Repository, 이하 RDR)의 현황을 조사하고, 국가별 효율적인 구축 및 활용을 조명하기 위해, re3data.org의 데이터를 수집하여 국가별 기여도, 콘텐츠 유형, 언어, 소프트웨어 사용 등을 분석하였다. 연구 결과, 영어가 주 언어로 사용되며, 가장 많이 사용되는 소프트웨어는 DataVerse으로 나타났다. 또한 대부분의 리포지터리는 개방형이며, 절반 이상이 특정 분야에 특화되어 있음을 언급하였다. 이 연구는 데이터 리포지터리가 적거나 없는 국가에 유용한 도구와 기법을 제공하고, RDR의 전반적인 특징을 제시하였다.

Lin et al.(2024)은 생물의학 연구와 오픈 사이언스에서 데이터 리포지터리의 중요성을 강조하고, 연구자와 이해관계자에게 데이터 관리 및 공유를 위한 적절한 리포지터리 선택에 대한 지침을 제공하기 위하여, 데이터 리포지터리의 기능, 서비스 기대치, 평가 기준을 탐구하고, 관련 커뮤니티에 가이드라인을 제시하였다. 연구 결과, 데이터 리포지터리는 연구데이터의 관리, 보존, 공유를 촉진하며, 투명성과 재현성을 높이는 데 기여한다는 것을 확인하였다. 또한 다양한 이해관계자들이 데이터 리포지터리 선택 시 고려해야 할 요소들을 명확히 하였다.

Stvilia, Lee(2024)는 RDR의 데이터 품질 이해와 데이터 품질 보증(Data Quality Assurance, 이하 DQA) 활동 구조에 대한 이론 기반의 탐색을 제공하기 위하여, 122개의 RDR 신청서, 32명의 큐레이터와 리포지터리 관리자 인터뷰, 및 관련 웹페이지를 분석하였다. 총 146개의 독특한 RDR을 포함한 데이터 세트를 활용하였다. 연구 결과로, DQA 활동이 평가, 개입, 커뮤니케이션의 세 가지 주요 활동으로 구분되었고, 각 활동의 구조가 나타났다. 참가자들은 데이터 품질을 전통적인 정의를 넘어 7개의 윤리적 및 효과적인 정보 시스템의 측면으로 확장하였으며, DQA 활동은 데이터 가치, 품질 수준, 전문가 가용성, 비용 및 자금 유인에 따라 우선순위가 정해졌음을 언급하였다.

Mosha, Ngulube(2023a)은 탄자니아의 고등교육 기관에서 연구데이터를 저장하고 공유하기 위한 오픈 연구데이터 리포지터리(RDR)의 활용을 조사하기 위해, 설문조사를 통해 넬슨 만델라 아프리카 과학기술 대학교(NM-AIST)에서 대학원생을 대상으로 데이터 수집 및 분석 관련 설문조사를 실시하였다. 응답자의 절반 미만이 오픈 RDR에 대한 인식과 사용 경험이 있었으며, 데이터 공유에 대한 저항감이 나타났다. 주요 이슈로는 데이터 소유권 상실과 데이터 보안 문제가 지적되었다. 연구 결과, 고등교육 기관에서는 신뢰할 수 있는 리포지터리 사용에 대한 교육과 동기 부여가 필요함을 언

급하였다.

Mosha, Ngulube(2023b)는 고등 교육 기관의 데이터 리포지터리에서 연구 데이터를 보존, 발견 및 재사용하기 위한 메타데이터 표준에 대한 기존 연구를 종합하기 위해, 2003년 1월부터 2023년 4월까지 발표된 1,597편의 논문을 5개 데이터베이스에서 검색하여 13편의 관련 논문을 선정하였다. 연구 결과, 선정된 논문들은 연구 데이터의 보존을 위한 세 가지 주요 메타데이터 유형(기술적, 구조적, 관리적)을 제시하였다. 한편 메타데이터 표준이 데이터 발견 및 재사용에 미치는 영향에 대한 증거는 제한적으로 나타났다. 또한, 특정 고등 교육 기관에서 메타데이터 표준을 적용한 사례는 보고되지 않았으며, 이 연구를 통해 연구자, 학생, 사서 및 데이터 관리자 등 다양한 이해관계자에게 실천적 통찰을 제공할 수 있는 가능성을 제공하였다.

김주섭(2023)은 Geoscience 분야 데이터 리포지터리 현황을 파악하고 나아가 리포지터리 인증을 획득할 수 있는 방안을 제시하였다. 국가별 데이터 리포지터리의 경우 한국은 13개가 등록되어 있으며 이마저도 국가간 협력 리포지터리를 제외하면 8개에 지나지 않음을 지적하였다. 또한, 리포지터리 인증을 위하여 필요한 전략으로 수집, 관리 그리고 보존 등을 포함한 데이터 관리 가이드라인 여기에 직원, 예산 등 영속성과 전문성을 보장하기 위한 조직 인프라 마지막으로 메타데이터, 품질 보장, 보안 등와 관련된 시스템 인프라 등 3개 영역과 관련한 요구사항을 제시하였다.

이혜림(2023)은 Coretrustseal 인증을 받은 33개의 데이터 리포지터리 보존 정책에 대하여 분석하였다. 분석 결과, 디지털 보존 정책 프레임워크의 필수 구성 요소를 도출하였으며 해당 구성 요소는 Coretrustseal 인증 획득시 필요한 보존 관리 정책 구성의 기초자료로 활용될 것으로 기대하고 있다.

Frank(2022)는 디지털 리포지터리의 신뢰성을 위한 감사 및 인증 과정에서 위험 개념을 이해하기 위해, Trustworthy Repositories Audit & Certification (TRAC) 시스템에 관련된 42명의 이해관계자와 인터뷰를 진행하고 ISO 16363 표준 관련 문서를 분석하였다. 결과적으로 이해관계자들은 위험을 재정, 법적, 조직 거버넌스, 리포지터리 프로세스, 기술 인프라의 다섯 가지 주요 범주로 분류하였다. 표준 개발자와 검토자는 문서화된 위험 식별 및 완화 전략이 신뢰성의 충분한 증거라고 보았지만, 그렇지 않은 의견을 제시한 경우도 있었다. 본 연구를 통해 위험 관리의 복잡성을 강조하며 이해관계자 간의 협력과 소통의 필요성을 제안하였으며, 더 나아가 디지털 리포지터리의 신뢰성을 보장하기 위해 보다 포괄적인 접근 방식을 촉구할 것을 언급하였다.

Khan et al.(2022)은 다양한 데이터 리포지터리의 현황을 조사하기 위해, re3data.org의 리포지터리 관리자들을 대상으로 온라인 설문조사를 실시하였다. 총 189명이 응답하였고, 응답자의 47%는 특정 분야의 리포지터리에서, 34%는 기관 리포지터리에서 활동함을 조사하였다. 결과적으로, 71%의 리포지터리가 맞춤형 기술 프레임워크를 사용하며, 데이터 재사용 및 확인에 관하여 어려움을 겪고 있으며, 주요 도전 과제로는 사용자 참여 부족과 인적 자원 부족이 지적되었고, FAIR 데이터 확보가

가장 높은 우선사항으로 나타났다. 이러한 연구 결과를 통해 데이터 리포지터리의 기능 향상 및 사용자 경험 개선을 위한 유용한 통찰을 제공하였다.

Schöpfel(2022)은 연구데이터 리포지터리의 주요 문제와 향후 전망을 탐구하기 위하여, 관련 문헌 분석과 현황 조사를 통해 도전 과제를 언급하였다. 결과적으로, 데이터 양과 복잡성, 연구 인프라 통합, 데이터 품질, 연구자 수용 등이 주요 과제로 나타났다. 또한, 연구 데이터 리포지터리는 데이터 기록, 설명, 보존, 검색, 배포의 중요한 역할을 수행함을 강조하였다. 또한 리포지터리가 연구 생태계에서 어떻게 기능하는지를 분석하였으며, 궁극적으로, 데이터 리포지터리의 효과적인 운영을 위한 개선 방안이 필요함을 언급하였다.

Downs(2021)는 연구데이터의 재사용 가능성을 높이고 보다 넓은 커뮤니티와 새로운 용도로의 활용을 촉진하기 위하여, 데이터 리포지터리의 개선과 인증 절차를 통해 데이터 재사용의 잠재력을 평가하고 분석하였다. 연구 결과, 데이터 리포지터리의 개선은 연구데이터의 가치를 높이고 새로운 사용을 촉진하는 데 기여한다는 점이 확인되었다. 또한, 효과적인 데이터 관리 및 지속적인 사용을 위한 인증 도구의 발전이 중요하다는 결론을 도출하였다. 이러한 접근은 연구데이터의 큐레이션 및 지속적인 사용에 대한 의미 측면에서 설명하고 논의하였다.

김우중 외(2021)는 과학기술 분야의 연구 데이터 개방과 공유를 위하여 오픈 엑세스와 사회연결망 이론으로 re3data.org에 등록된 자연과학 분야 연구데이터 리포지터리의 세부 속성을 대상으로 다양한 분석을 수행하였다. 그 결과, 데이터 리포지터리의 운영수준에 대하여 우수, 보편, 열악, 비공개 4가지 유형의 클러스터를 도출하였으며 또한 연구데이터 리포지터리의 분야에 따라 컨텐츠 유형에 차이가 있으며, 이용도(방문횟수)가 높은 데이터 리포지터리일수록 표준 연구문서 형태로 데이터가 공유되는 경우가 많다는 점을 파악하였다. 이러한 분석 결과를 통해 개방형 혁신을 위한 데이터 리포지터리 활용을 제고할 필요성이 있음을 언급하였다.

Donaldson(2020)는 TDRs 인증을 받은 디지털 리포지터리 웹사이트에서 인증 관련 정보의 존재와 품질을 평가하기 위해, 91개의 TDRs 인증 리포지터리 웹사이트를 대상으로 콘텐츠 분석을 수행하여 인증 상태, 인증 마크, 추가 정보 링크, 인증 과정 설명, 감사 보고서 공유 여부를 조사하였다. 연구 결과, 약 75%의 웹사이트가 TDRs 상태에 대한 명확한 진술을 제공하며, 60%는 인증 마크에 추가 정보 링크를 포함하고 있다. 또한, 60%의 리포지터리가 검토 보고서를 공유하고, 1/3 이상이 인증 과정을 설명하였다.

조재인, 박종도(2019)는 re3data.org에 등록된 인문사회 분야의 데이터 리포지터리를 분석하여 4개의 군집으로 유형화하였다. 그 결과 약 70% 정도가 보편 수준의 리포지터리로 확인되었으며 특히 운영주체가 독일이거나 분야가 언어학인 경우 우수한 군집임을 제시하였다. 하지만 아시아 국가 및 국내 RDR이 국제 수준에 미치지 못함을 지적하면서 리포지터리 인프라 구축이 필요함을 제언하였다.

이상과 같이 2019년부터 2024년까지 국내외

논문 14편을 살펴본 결과, 리포지터리 활용에 관한 논문이 2편으로 연구데이터 저장, 공유 및 보존에 RDR이 유용한지를 확인하였다. 다음으로 리포지터리 현황 분석에 관한 연구가 5편으로 학문 분야의 RDR 속성을 분석하여 발전 방향을 제시하였다. 또한, 리포지터리 품질과 인증 관련한 선행 연구는 6편으로 인증 현황, 인증 프로세스에서의 위험 인식 그리고 보존 정책을 포함한 인증 전략 등에 대한 내용을 살펴볼 수 있었으며 리포지터리 선택을 위한 고려사항을 제시한 논문도 확인할 수 있었다. 본 연구에서는 리포지터리 운영 현황을 파악하기 위하여 Geoscience 분야의 re3data.org의 9가지 속성으로 운영 현황을 조사하였고 덧붙여 해당 도메인 분야에서 '우수'한 운영 현황을 보인 리포지터리 서비스를 확인하여 Geoscience 분야 데이터 리포지터리 서비스를 제안하고자 한다.

3. Geoscience 분야 데이터 리포지터리 운영 현황 분석

이번 장에서는 re3data.org에 등록된 Geoscience 분야 데이터 리포지터리의 운영 현황을 분석하고 해당 분야 리포지터리 중 운영 현황이 우수하다고 판단되는 리포지터리를 선정하고자 한다. 선정된 데이터 리포지터리는 한국지질자원연구원이 운영 중인 Geo Big Data Open Platform에서 참고할 수 있는 서비스를 제안하기 위한 근거자료로 활용하고자 한다.

3.1 Geoscience 분야 데이터 리포지터리 운영 현황

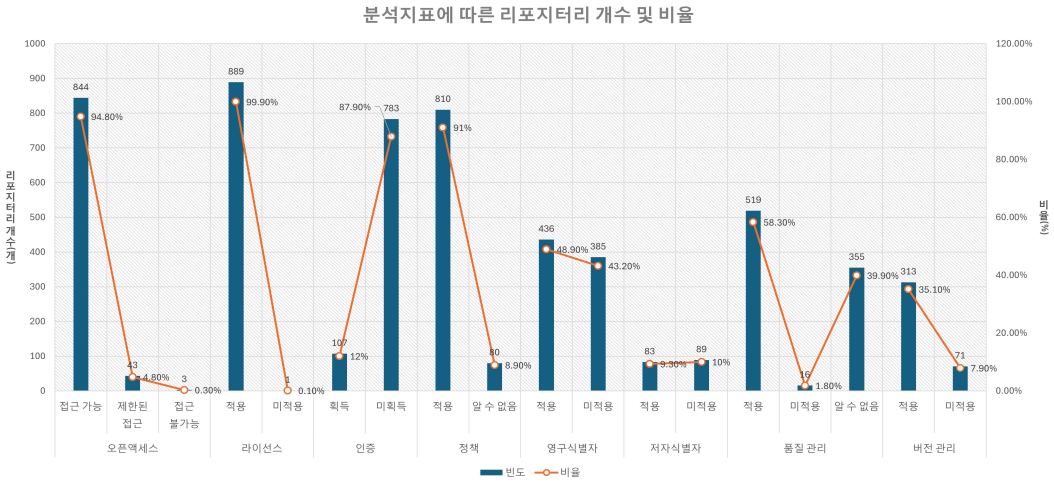
re3data.org에 등록된 Geoscience 분야의 890개의 데이터 리포지터리를 대상으로 re3data에서 제공하는 9개의 속성(오픈액세스, 라이선스, 인증, 정책, 영구식별자, 저자식별자, 품질관리, 버전관리, 메타데이터)으로 운영 현황을 파악하였다. 다음의 〈표 4〉 및 〈그림 6〉은 8개의 속성[1]을 기준으로 890개의 Geoscience 분야 데이터 리포지터리 운영 현황을 정리한 것이다.

먼저 오픈액세스의 경우 접근 가능한 데이터 리포지터리가 844개로 약 95%로 대다수를 차지하고 있다. 라이선스의 경우 CC(Creative Commons)와 같은 다양한 라이선스를 1곳을 제외한 889곳에서 적용하고 있음을 알 수 있다. 데이터 리포지터리 인증의 경우 인증을 획득한 리포지터리가 107개로 약 12%를 차지하고 있으며 이 중에서도 CTS가 60개로 가장 많은 수치를 기록하고 있다. 정책의 경우 데이터 리포지터리가 810개, 91%로 대부분의 리포지터리가 적용하고 있음을 확인하였다. 데이터에 대한 영구식별자의 경우 DOI를 비롯한 식별자를 적용한 리포지터리가 전체 대비 49%로 절반 가까운 곳이 사용하고 있다. 이와 대비적으로 저자식별자의 경우 전체 대비 9.3%로 적용하고 있는 리포지터리가 83개로 파악되었다. 저자식별자의 경우 ORCID가 가장 많은 수치로 확인되었다. 다음으로 품질관리과 버전관리의 경우 각각 58.3% 그리고 35.1%에 해당하는 리포지터리에서 적용하고 있음을 알 수 있다. 마

1) 메타데이터는 사용하고 있는 스키마가 다양하여 표와 그림에서는 현황을 제외하였음.

〈표 4〉 re3data의 속성 기반 리포지터리 운영현황

분석 지표			빈도	비율
오픈액세스	접근 가능		844	94.8%
	제한된 접근		43	4.8%
	접근 불가능		3	0.3%
라이선스	적용		889	99.9%
	미적용		1	0.1%
인증	획득	CoreTrustSeal	60	6.7%
		DIN 31644	1	0.1%
		DSA	5	5.6%
		Trusted Digital Repository	1	0.1%
		WDS	35	3.9%
		기타	5	0.05%
		Total	107	12%
	미획득		783	87.9%
정책	적용		810	91%
	알 수 없음		80	8.9%
영구식별자	적용	ARK	7	0.7%
		DOI	352	39.5%
		hdl	37	4.1%
		IGSN	1	0.1%
		PURL	4	0.4%
		URN	10	1.1%
		기타	25	2.8%
		Total	436	48.9%
	미적용		385	43.2%
저자식별자	적용	AuthorClaim	3	0.3%
		ISNI	2	0.2%
		ORCID	75	8.4%
		기타	3	0.3%
		Total	83	9.3%
	미적용		89	10%
품질 관리	적용		519	58.3%
	미적용		16	1.8%
	알 수 없음		355	39.9%
버전 관리	적용		313	35.1%
	미적용		71	7.9%



〈그림 6〉 분석지표에 따른 리포지터리 개수 및 비율

지막으로 메타데이터의 경우 약 60%의 리포지터리에서 표준 스키마를 채택하고 있었으며 2개 이상의 스키마를 사용하는 리포지터리가 대부분으로 파악되었다. 가장 많이 사용하는 메타데이터를 순서대로 나열하면 ISO 19115, Dublin Core, DataCite 등 분야와 관계없이 적용할 수 있는 스키마로 나타났으며 이외에도 도메인 분야의 특성을 나타내는 ABCD, CF(Climate and Forecast), Darwin Core, EML 등도 사용함을 알 수 있었다.

다음의 〈표 5〉는 re3data.org의 9개 속성값을 이용하여 데이터 리포지터리 운영 현황을 나타낸 것이다.

9개의 속성을 기준으로 속성 7개를 만족한 리포지터리의 유형을 '우수'로 4개이상 6개 이하를 만족한 유형은 '보편'으로 속성 3개 이하를 만족한 유형은 '열악'으로 표시하였으며 마지막으로 접근할 수 없는 리포지터리는 'N/P(Non-Public)'로 구분하였다(김우중 외, 2021).

re3data의 속성을 기반으로 평가한 결과 '우수' 유형으로 확인된 리포지터리는 164개이며 가장 많이 차지하는 '보편' 등급은 470개로 약 56%를 차지하였다. 이외 접근할 수 없는 데이터 리포지터리는 890개 중 6개로 비율로는 1%가

〈표 5〉 Geoscience 분야 데이터 리포지터리 운영 현황

평가 등급	빈도	비율(%)	비고
우수	164	18.4	• 우수: 속성 7개 이상 만족 • 보편: 속성 4개 이상 6개 이하 만족 • 열악: 속성 3개 이하 만족 • N/P: 접근할 수 없음
보편	470	55.8	
열악	223	25.1	
N/P	6	0.6	
Total	890	100	

되지 않았다. 이상으로 확인한 결과 Geoscience 분야의 데이터 리포지터리 운영 현황은 보편적으로 지표를 만족한다고 볼 수 있다.

3.2 Geoscience 분야 데이터 리포지터리 서비스 현황

본 절에서는 Geoscience 분야 데이터 리포지터리 서비스를 제안하기 위하여 분석 대상을 선정하였다. 우선 3장 1절에서 분석한 데이터 리포지터리 운영 현황 중 결과가 '우수'로 판단되는 리포지터리 중 Geoscience 분야와 가장 관련성 있는 11개의 리포지터리와 한국지질자원연구원의 Geo Bigdata Open Platform을 포함하여 다음의 〈표 6〉과 같이 정리하였다.

해당 내용을 간략히 살펴보면 국가별로는 미

〈표 6〉 분석대상 데이터 리포지터리 목록

No.	리포지터리명	국가	운영기관	펀딩기관	리포지터리 인증
1	Australian Ocean Data Network Portal	호주	Australian Government, Australian Institute of Marine Science	Australian Government, Department of Education and Training, National Collaborative Research Infrastructure Strategy	CTS
2	Environmental Information Data Centre	영국	UK Centre for Ecology & Hydrology	Natural Environment Research Council	CTS/WDS
3	Geo Bigdata Open Platform	한국	한국지질자원연구원	과학기술정보통신부	
4	ICTS SOCIB Data Repository	스페인	Govern de les Illes Balears	Spanish National Research Council	CTS
5	IFREMER-SISMER Portail de données marines	프랑스	Institut français de recherche pour l'exploitation de la mer		CTS
6	National Geoscience Data Centre	영국	British Geological Survey	Natural Environment Research Council	CTS
7	Neotoma Paleoecology Database	미국	North Dakota State University	National Science Foundation	WDS
8	NSF Arctic Data Center	미국	DataONE	National Science Foundation	CTS
9	PANGAEA	독일	University of Bremen	Alfred-Wegener-Institut, Helmholtz-Zentrum für Polar- und Meeresforschung (AWI)	CTS/WDS
10	Research Data Archive at NCAR	미국	University Corporation for Atmospheric Research	National Science Foundation	CTS
11	Tethys	오스트리아	GeoSphere Austria		CTS
12	The CEDA Archive	영국	National Center for Atmospheric Science	Natural Environment Research Council	CTS

국, 영국이 각각 3곳, 프랑스, 스페인, 오스트리아, 독일, 호주 그리고 한국이 각각 1곳씩 포함되었으며 인증의 경우 CTS 인증을 받은 국가는 10 그리고 WDS는 3곳으로 나타났으며 CTS와 WDS 둘 다 획득한 곳은 2곳으로 확인되었다.

Australian Ocean Data Network Portal(호주 해양 데이터 네트워크, 이하 AODN)은 호주의 해양 및 기후 데이터에 대한 통합적인 리포지터리로 호주 해양 커뮤니티에서 수집한 데이터를 제공한다. AODN의 데이터는 맵 인터페이스와 메타데이터 카탈로그를 통해 검색할 수 있으며, 2016년 5월에 IMOS(Integrated Marine Observing System, 통합 해양 관측 시스템/ eMarine Information Infrastructure(eMII) 시설과 합병되었다. IMOS는 오픈 데이터 접근에 중점을 둔 다기관 협업체로 AODN을 관리하고 있다. AODN 포털은 데이터를 검색 및 하위 집합으로 나누고 다운로드하는 방법에 대한 시각적 개요를 제공하는 비디오 튜토리얼을 제공한다. AODN은 해양 및 기후 데이터 리소스의 상호 운용 가능한 온라인 네트워크로 IMOS와 6개 호주 연방 기관이 AODN의 주체이며, 많은 대학과 주 정부 기관 또한 AODN에 데이터 리소스를 제공하고 있다. 해양 데이터는 대중에게 무료로 제공된다. 데이터는 바다를 항해하는 선박, 자율 주행 차량, 계류 및 기타 플랫폼에서 수집한 다양한 해양 환경의 광범위한 매개변수를 포함하고, 지리적으로 바다에서 해안까지, 학문 분야(물리적, 생지화학, 생물학적)에 걸친 관찰 범위는 사용자가 필요한 데이터를 효율적으로 얻을 수 있도록 직관적이고 사용하기 쉬운 강력한 정보 인프라 및 지역 사회가 AODN에 참여하는 데 도움이 되는 추가 정보와 지침을 제공한다(Australian Ocean Data Network Portal, 2024).

Environmental Information Data Centre(환경 정보 데이터 센터, 이하 EIDC)는 NERC(Natural Environment Research Council's, 자연환경 연구위원회)의 환경 데이터 서비스의 일부이며, UKCEH(UK Centre for Ecology & Hydrology, 영국 생태 및 수문학 센터)를 중심으로 지상 및 담수 과학과 관련된 국가적으로 중요한 데이터셋을 관리한다(Environmental Information Data Centre, 2024).

한국지질자원연구원이 운영하는 Geo Bigdata Open Platform은 2015년 GDR(Geoscience Data Repository) 파일럿 시스템으로 출발하여 현재는 지질자원 정보를 누구나 접근할 수 있도록 만들어진 오픈 플랫폼 형식의 리포지터리로 운영되고 있다. 국토지질, 광물자원, 지질환경, 석유해저 등 주제별 검색과 조사, 탐사 등 유형별 검색을 제공하고 있다. 데이터 관리 가이드라인의 경우 수집, 관리, 보전 및 윤리 등 관련 가이드라인을 제공하고 있으며 해당 플랫폼은 2023년 8월 기준으로 CTS 인증 신청을 해놓은 상황이며 현재 심사가 진행중이다(Geo Bigdata Open Platform, 2024).

ICTS SOCIB Data Repository(발레아레스 제도 연안 관측 및 예측 시스템(이하 ICTS SOCIB)은 해양학 데이터 제품 스트림과 모델링 서비스를 제공하는 다중 플랫폼 분산 및 통합 시스템이다. 스페인, 유럽 및 국제 프레임워크에서 운영 해양학을 지원하고 글로벌 변화 맥락에서 해양 및 연안 연구의 요구에 기여한다. ICTS SOCIB는 8개 시설의 광범위한 장비 및 모델의 배치, 데이터 관리를 조정하며, 외부 국제 기관의 데이

터를 관리하고 해양 데이터의 보급을 위해 국제적으로 협력한다(ICTS SOCIB Data Repository, 2024).

SISMER(Scientific Information Systems for the Sea)는 Ifremer가 구현을 담당하는 수많은 해양 데이터베이스와 정보 시스템을 관리하는 Ifremer의 서비스이다. SISMER가 관리하는 정보 시스템은 CATDS(SMOS 위성 데이터)부터 지구과학 데이터(수심측량, 지진, 지질 샘플)에 이르기까지 다양하며, 수층 데이터(물리학 및 화학, 운영 해양학 데이터 - Coriolis - Copernicus CMEMS), 어류 데이터(Harmonie), 연안 환경 데이터(Quadrige) 및 심해 환경 데이터(Archimède)도 포함된다. 따라서 SISMER는 Ifremer와 많은 국가, 유럽 및 국제 프로젝트에서 해양 데이터베이스 관리에 핵심적인 역할을 한다(ifremer, 2024).

National Geoscience Data Centre(NGDC)는 BGS의 환경 모니터링 데이터, 디지털 데이터베이스, 물리적 컬렉션(시추공 코어, 암석, 광물 및 화석), 기록 및 아카이브를 포함하여 400개 이상의 데이터셋을 관리한다. 지구과학 데이터와 정보를 수집하고 보존하여 광범위한 사용자와 커뮤니티에 장기적으로 제공하며 지구과학 데이터를 위한 NERC 환경 데이터 센터이며, 지구과학적으로 가치 있는 정보와 데이터셋을 관리하고 있다. NGDC가 보유한 데이터는 지질 시간에서부터 실시간 센서 및 데이터 스트림까지 측정된 지구의 물리적 구조와 이에 작용하는 프로세스를 다루는 많은 지질학 분야를 포괄한다. 기존 과학 기록을 뒷받침하는 이 데이터를 보존하고 광범위한 사용자 커뮤니티에서 향후 재사용할 수 있도록 하는 것이 NGDC 정책이다(National Geoscience Data Centre, 2024).

Neotoma는 현대의 미세화석 샘플을 포함하여 플리오세-제4기를 포괄하는 다중 프록시 고생태 데이터베이스로 19개 기관이 국제적으로 협력하고 있다. Neotoma 고생태 데이터베이스에는 꽃가루 미세화석, 식물 거대화석, 척추동물군, 규조류, 목탄, 바이오마커, 조개, 물리적 퇴적학 및 수화학을 포함하여 20개 이상의 데이터 유형이 있다. Neotoma는 데이터 수집, 검색, 표시, 분석 및 배포를 위한 공통 소프트웨어 도구 개발을 가능하게 하는 기본 사이버 인프라를 제공하는 동시에 도메인 과학자에게 중요한 분류학 및 기타 데이터 품질 문제를 제어할 수 있는 권한을 제공한다(Neotoma, 2024).

NSF Arctic Data Center(이하, NSF Arctic)는 NSF Polar Programs의 Arctic 섹션을 위한 주요 데이터 및 소프트웨어 리포지터리이다. 이 센터는 연구 커뮤니티가 데이터, 메타데이터, 소프트웨어, 문서 및 이들을 연결하는 출처를 포함하여 NSF에서 펀딩한 모든 북극 연구 제품을 재현 가능하게 보존하고 발견할 수 있도록 한다. 데이터는 오픈 라이선스에 따라 공개되며, NSF Arctic 연구 프로그램에서 지원하는 모든 과학, 공학 및 교육 연구가 포함된다. 여기에는 자연 과학(지구과학, 해양학, 생태학, 대기 과학, 생물학 등) 및 사회 과학(고고학, 인류학, 사회 과학 등)이 있다(NSF Arctic Data Center, 2024).

PANGAEA(지구 및 환경 과학을 위한 데이터 퍼블리셔)는 지구, 환경 및 생물 다양성 과학의 지리 참조 데이터를 보관, 출판 및 배포하는 오픈 액세스 라이브러리로서 30년의 역사

를 가지고 있다. 퇴적물 코어를 위한 데이터베이스에서 발전하여 알프레드 베게너 연구소, 헬름홀츠 극지 및 해양 연구 센터(AWI) 및 브레멘 대학교의 해양 환경 과학 센터(MARUM)의 공동 시설로 운영되고 있다. PANGAEA는 세계 기상 기구(WMO)의 위임을 받았으며 세계 방사선 모니터링 센터(WRMC)로 인증되었다. 2001년에 국제 과학 위원회(ICS)에서 세계 데이터 센터로 추가로 인증되었으며 2019년부터 Core Trust Seal로 인증되었다. PANGAEA와 출판 업계 간의 성공적인 협력과 해당 기술 구현을 통해 출판물의 보충 자료로 보관된 과학 출판물과 데이터셋의 교차 참조가 가능해졌다. PANGAEA는 수많은 국제 과학 저널에서 권장하는 데이터 리포지터리이다(PANGAEA, 2024).

NSF National Center for Atmospheric Research Research Data Archive(이하, NSF NACR RDA)는 Computational and Information Systems Laboratory(계산 및 정보 시스템 연구실, CISL)의 Data Engineering and Curation Section(데이터 엔지니어링 및 큐레이션 섹션, DECS)에서 관리하고 있다. 또한 대기 및 지구 과학 연구를 지원하는 광범위하고 다양한 기상 및 해양학 관측 자료, 운영 및 재분석 모델 출력, 원격 감지 데이터셋을 보관하며, NSF NCAR 고성능 컴퓨팅 리소스와 통합되어 대기 및 지구 과학 연구를 지원한다(Research Data Archive at NCAR, 2024).

Tethys는 GeoSphere 오스트리아의 오픈 액세스 연구데이터 리포지터리로, GeoSphere 오스트리아에서 공동으로 생성한 지리 참조 지구 과학 연구데이터를 출판하고 배포한다. 연구데이터 출판물과 관련 메타데이터는 주로 독일어 또는 영어로 제공된다. Tethys는 공개 데이터로 게시된 데이터셋을 제공하고 FAIR 데이터 원칙에 따라 데이터를 검색 가능, 접근 가능, 상호 운용 가능 및 재사용 가능하도록 하는 것을 목표로 한다(Tethys, 2024).

The CEDA(Centre for Environmental Data Analysis. 환경 데이터 분석 센터) Archives는 대기 과학 및 지구 관측을 위한 NERC의 데이터 리포지터리이다. CEDA는 3개의 데이터 센터, 데이터 분석 환경 및 다양한 관련 연구 프로젝트 참여를 통해 환경 과학 커뮤니티에 서비스를 제공하며, 환경 과학을 지원하고, 환경 데이터 보관 정책을 더욱 발전시키고, 데이터 접근성을 향상시키기 위한 새로운 기술을 개발하고 배포하는 것을 목표로 한다. 또한 대규모 데이터 분석을 지원하는 서비스를 제공한다(The CEDA Archive, 2024).

4. Geoscience 분야 데이터 리포지터리 서비스 제안

이번 장에서는 3장에서 제시한 Geoscience 분야 데이터 리포지터리에서 제공하는 서비스에 대하여 비교 및 분석한 내용을 바탕으로 해당 분야에서 활용이 가능하도록 군집화하여 제공하고자 한다. 먼저, Geoscience 분야의 데이터 리포지터리 서비스의 상세 내용에 대하여 공통 카테고리를 '데이터 정책', '데이터 큐레이션', '품질 관리(메타데이터 포함)' 그리고 '도구', 'API' 그리고 '메타데이터'로 구분하였다. 다음의 〈표 7〉은 제시한 카테고리로 Geoscience 분애 데이터

〈표 7〉 Geoscience 분야 12개 데이터 리포지터리 서비스 비교 분석

카테고리 / 리포지터리	데이터 정책	데이터 큐레이션	도구	API	품질 관리	메타데이터
AODN	• 기탁 및 스토리지 • 보존 및 관리 • 인용	• AODN 포털 검색 및 다운로드 • 데이터 제공: THREDDS 데이터 서버(TDS) • AODN 용어집	• 동물 추적 DB, 음향 및 AUV 이미지 뷰어 • IMOS 툴박스 및 NetCDF 파일 • 글라이더스코프 및 주피터 노트북 • AODN 메타데이터 입력 도구	-	• 품질 보증 및 품질 관리 프레임워크	• ISO 19115
EIDC	• 기탁 및 DMP • 라이선스 및 인용 • 보존 및 DOI 정책	• 식별자 발급 • 카탈로그 및 검색, 다운로드 • 데이터 제공: 카탈로그 및 타 서비스 통해서도 데이터 제공	-	-	-	• DataCite Metadata Schema
ICTS SOCIB	• 데이터 관리, 보존 • 윤리 • 인용	• 카탈로그: 검색, 다운로드, 시각화 • 데이터 관리 • 도큐멘테이션	• Thredds 데이터 서버(표준서비스)	• API: REST	• 해양 데이터 스트림 • 데이터 품질 전략 및 품질 관리 절차	• CF Metadata Conventions
IFREMER-SISMER	• 기탁 • 출판 및 인용 • 보존 • 데이터 관리	• DOI 제공 • 데이터 관리 • 종이기록에 대한 주문형 스캔 서비스	• 예약 접근 • Ifremer.fr	-	• 품질 접근방식	• CF Metadata Conventions
NGDC	• 기탁/수집 • 관리 및 보존 • 인용	• 기탁 • 데이터 분석 및 서비스 • 인용 데이터 카탈로그 • DOI 부여 및 데이터셋 표준 정책 • 용어사전 및 출판물	• 오픈 지오사이언스 • SIGMA(통합지구과학매핑시스템) • 그라운드호그 v2.8.3 및 아키모드 • CLiDE 환경 모델링 플랫폼 • GIS 지하수 및 리소스 프레임 뷰어 • BGS GitHub • 웹서비스 - 웹 맵 서비스(WMS) 및 웹 픽처 서비스(WFS), API - 웹용 OGC 카탈로그 서비스(CSW) - GeoRSS 피드: 영국 및 전 세계 지진 지도 - AGS 다운로드 서비스 - Mashups 및 NGDC 기탁 데이터 - Linked data 및 좌표변환기	-	• 데이터 값 체크리스트	• ISO 19115
Neotoma	• 인용 • 기탁	• 데이터 시각화 및 큐레이션	• 데이터 분석(R패키지) 및 데이터 모델 • Data Viewer: Explorer • 데이터 시각화 및 큐레이션: Tilia	• API 및 소프트웨어 개발 키트(SDK) - DB에 원격접근	• Tilia 소프트웨어	• Repository-Developed Metadata Schemas
NSF Arctic	• 기탁 및 보존 • 윤리	• 검색 및 다운로드	• 공간도구(spatial tools)	• DataONE REST API	-	• CF Metadata Conventions

카테고리 / 리포지터리	데이터 정책	데이터 큐레이션	도구	API	품질 관리	메타데이터
NSF NCAR RDA	• 보안 • 인용	• 데이터 큐레이션 수준 • 데이터 관리 • 수집-배포까지 워크플로우 • 데이터 인용 서비스 • 도큐멘테이션	• THREDDS 데이터 서버 • 웹서비스 - 데이터셋 요청 API 및 CFSR 요청 자동화 - OGC 카탈로그 서비스 및 웹 맵 서비스	• API 및 키워드 뷰어 - REST API 및 OAI-PMH 메타데이터	-	• DIF
PANGAEA	• 기탁 및 접근 • FAIR 및 라이선스 • 품질보증 • 출판 및 장기보존	• 수집 및 기탁 • 관리 • 접근 및 출판	• 고급 검색도구: 데이터 웨어하우스 • 스크립팅 언어를 위한 데이터 검색 도구 • 수심측량 웹 맵 서비스 • BSRN 툴박스	• 데이터 검색 시 프로그래밍 방식 접근을 위한 API	• 품질보증 문서화 - 품질 플래그(flag) 등	• DIF
Tethys	• FAIR • 보존	• 데이터 출판 및 공개 • 일반 데이터 및 정보 수집	-	• 오픈스트리트맵: API를 통해 오픈 소스 매핑 도구인 “OpenStreetMap”(OSM)을 사용	• 완전성 검사	• DataCite Metadata Schema
The CEDA Archive	• 수집 및 기탁 • 접근 • 저작권 • 보존	• 환경데이터 큐레이션 전문성	• 일반 다운로드 서비스 • 특정 데이터셋 서비스 - 지구 시스템 그리드 연합(ESGF) - 항공편 검색기(항공 데이터) 등	• OPeNDAP:http를 통한 아카이브 API	• 표준 데이터 형식으로 환경 데이터 수집 • 데이터 형식 검사기 - CSV / NASA-Ames	• CF Metadata Conventions
Geo Bigdata Open Platform	• 수집, 관리, 보존 • 윤리・저작권・라이선스	• 국제지질시료번호(IGSN) 발급 • 수치지질도 다운로드 • 지질자원주제도 2D/3D가시화 • 독도 드론플랫폼	-	• Open API	-	• DC(Dublin Core)

리포지터리에서 제공하는 서비스에 대하여 비교 분석한 것이다.

다음의 〈표 8〉은 위의 〈표 7〉에서 제시한 내용을 바탕으로 카테고리별 서비스 진행 여부에 대하여 나타낸 것이다.

우선 12개의 Geoscience 분야를 대상으로 정리한 결과 데이터 정책, 및 메타데이터의 경우 모든 리포지터리에서 적용하고 있는 것으로 파악되었다. 다음으로 데이터 큐레이션 및 리포지터리 인증의 경우 Geo Bigdata Open Platform을 제외하고 11곳에서 적용하고 있으며 CTS 또는 WDS 인증을 획득한 것으로 파악되었다. 다음으로 분석 도구의 경우 EIDC와 Geo Bigdata Open Platform을 제외하고 10개의 리포지터리에서 해당 분야 데이터 분석 서비스를 위한 도구를 제시하고 있다. 다음으로 데이터 배포를 위한 API 도구는 9곳에서 마지막으로 품질 관리의 경우 8곳의 리포지터리에서 제시하고 있는 것으로 나타났다.

첫 번째, 데이터 정책 관련하여 12개의 리포지터리에서 제공하는 세부 내용을 빈도순으로 살펴본 결과, 데이터 보존이 10건, 기탁이 8건 인용이 7건, 데이터 관리가 5건, 그리고 라이선스, 윤리, 수집과 관련한 정책이 3건으로 나타났다. 이러한 내용을 보아 Geoscience 분야 데이터 리포지터리에서는 보존, 기탁, 인용 등에 관한 데이터 관리 정책을 제안할 수 있다.

두 번째, 데이터 큐레이션의 경우 데이터 검색, 수집, 접근 및 출판 등 데이터 라이프 사이클과 관련한 기능을 제공하고 있다. 세부 내용으로 DOI와 같은 식별자를 발급하거나 데이터 분석과 관련한 서비스를 제공하고 있으며 용어사전을 통해 관련 검색을 보조하는 도구도 제시하고 있는 것으로 파악되었다. 이외에도 오프라인 자료에 대한 스캔 서비스, 데이터 시각화 등을 제시하거나 데이터 인용 형식을 제시하고 있다. 이러한 내용을 바탕으로 Geoscience 분야 데이터 리포지터리에서는 데이터 관리 기능을 추가해야 할 것이다.

세 번째, 데이터 분석 도구 및 API이다. 데이터 분석을 위하여 데이터 분석 및 데이터 모델링을 위한 도구와 데이터 뷰어 그리고 데이터

〈표 8〉 Geoscience 분야 데이터 리포지터리 비교 분석 종합

카테고리 \ 리포지터리		1	2	3	4	5	6	7	8	9	10	11	12	Total
서비스 현황	데이터 정책	O	O	O	O	O	O	O	O	O	O	O	O	12
	데이터 큐레이션	O	O	O	O	O	O	O	O	O	O	O		11
	분석 도구	O		O	O	O	O	O	O	O	O	O		10
	API			O		O	O	O	O	O	O	O	O	9
	품질 관리	O		O	O	O	O			O	O	O		8
	메타데이터	O	O	O	O	O	O	O	O	O	O	O	O	12
리포지터리 인증		O	O	O	O	O	O	O	O	O	O	O		11
계		6	4	6	5	6	6	5	5	6	6	6	3	-

1. AODN, 2. EIDC, 3. ICTS SOCIB, 4. IFREMER-SISMER, 5. NGDC, 6. Neotoma, 7. NSF Arctic, 8. NSF NCAR RDA, 9. PANGAEA, 10. Tethys, 11.The CEDA Archive, 12. Geo Bigdata Open Platform

시각화 및 큐레이션을 위한 다양한 도구를 제공하고 있다. 도메인 특성상 공간을 분석하기 위한 도구와 특정 데이터 형식을 해석할 수 있는 소프트웨어도 제공하고 있으며 여기에 클라우드에서 데이터 분석을 하기 위한 환경을 제공하고 있으며 메타데이터 입력 도구도 제시하기도 한다. 또한 스크립팅 언어를 위한 데이터 검색 도구와 특정 데이터셋을 위한 다양한 접근 및 다운로드 서비스도 제공하고 있다. API의 경우 데이터 검색 및 배포를 위하여 Open API 형태로 다양한 데이터셋 또는 메타데이터에 대하여 공개하고 있다. 해당 내용을 통해 Geoscience 분야 데치터 리포지터리에서는 데이터 분석을 위한 도구를 제공해야 할 것이다.

네 번째, 데이터 품질 관리에 관하여 데이터 품질 관리 전략 및 절차, 품질 보증 및 관리 프레임워크 그리고 데이터 품질 보증 문서화 등으로 요약될 수 있다. 여기에는 데이터 형식 검사기 또는 완전성을 검사할 수 있는 방법을 제안하고 있으며 표준 데이터 형식으로 해당 분야 데이터 형식을 포함하고 있다. 또한 데이터 품질 여부를 가늠할 수 있는 데이터 값 체크리스트를 제공하고 있다. 데이터 품질관리를 위한 정책 및 전략 그리고 도구를 통해 데이터 품질 제고 기능을 Geoscience 분야에서는 고려해야 할 것이다.

이외에 메타데이터의 경우 CF(Climate and Forecast), ISO 191115, DIF(Directory Interchange Format) 등과 같은 지구과학 도메인의 특성에 적합한 스키마를 적용하고 있는 것으로 나타났다.

5. 논의 및 결론

본 연구의 목적은 Geoscience 분야 데이터 리포지터리 서비스를 제안하기 위함이다. 이러한 목적을 달성하기 위하여 데이터 리포지터리 현황을 파악하였다. 리포지터리 현황은 re3data.org와 마스터 데이터 리포지터리 목록을 통해 주제별 국가별로 확인하였다. 또한 Geoscience 분야 리포지터리 운영 현황을 파악하기 위하여 re3data.org에서 제공하는 속성 값의 만족 여부에 따라 유형을 4가지로 구분하였다. 이 중에서 운영유형이 우수하다고 판단되는 리포지터리 11개와 한국지질자원연구원이 운영하는 Geo Bigdata Open Platform을 포함하여 12개에 대한 리포지터리 서비스를 분석하였다. 분석 결과를 요약하면 '데이터 정책', '데이터 큐레이션', '데이터 분석 도구 및 API' 그리고 '데이터 품질 관리' 4가지로 구분될 수 있다.

우선, '데이터 정책'을 통해 데이터 보존, 기탁, 인용 등을 위한 가이드라인을 제시할 수 있다. 다음으로 '데이터 큐레이션'의 경우 데이터의 생명 주기에 관련한 내용으로 데이터 수집부터 배포까지의 내용에 관한 것으로 데이터 수집, 접근, 다운로드 등과 관련한 기능을 제공하고 있으며 여기에 식별자 부여 그리고 시각화 등을 포함하고 있다. 다음으로 '데이터 분석 도구 및 API'에서는 도메인 분야 특정 데이터를 분석하기 위한 다양한 도구와 소프트웨어와 데이터셋과 메타데이터 배포를 위한 API를 제공하고 있다. 마지막으로 '데이터 품질 관리'에서는 데이터 품질 관리 전략, 절차 및 문서화 그리고 표준 데이터 형식에 관한 내용을 제공하고 있다.

본 연구에서 제공하고 있는 Geoscience 분야의 데이터 리포지터리 서비스는 우수하다고 판단되는 일부 데이터 리포지터리만 파악하였기 때문에 한계가 있을 수밖에 없다. 또한 해당 도메인 분야의 이해관계자들의 요구사항을 담아내지 못했다는 점은 본 연구의 한계점이라고 볼 수 있다. 하지만 특정 도메인 분야의 데이터 리포지터리 운영 현황에 대하여 평가하였다는 점과 12개의 데이터 리포지터리 서비스 내용을 파악하여 해당 분야의 리포지터리 서비스를 제안하였다는 점은 본 연구의 시사점이라고 볼 수 있다. 향후 해당 분야의 이해관계자의 요구사항을 수용할 수 있는 서비스에 대한 연구가 진행되기를 희망해본다.

참 고 문 헌

김우중, 김형준, 박희진, 최상옥 (2021). 자연과학 분야 연구 데이터 리포지터리 속성분석: 운영수준, 주제, 유형, 활용도를 중심으로. 기술혁신학회지, 24(4), 577-595.

김주섭 (2023). Geoscience 분야 데이터 리포지터리 현황과 Coretrustseal 인증 획득 방안에 관한 연구: re3data.org와 Coretrustseal 인증 모범사례를 중심으로. 한국도서관·정보학회지, 54(2), 89-110. https://doi.org/10.16981/kliss.54.2.202306.89

이혜림 (2023). 데이터 리포지토리의 보존 정책 프레임워크에 관한 연구: CoreTrustSeal 인증을 중심으로. 한국문헌정보학회지, 57(4), 119-138. http://doi.org/10.4275/KSLIS.2023.57.4.119

조재인, 박종도 (2019). re3data를 기반으로 한 인문사회 RDR 연구. 한국비블리아학회지, 30(2), 69-87. https://doi.org/10.14699/kbiblia.2019.30.2.069

지오빅데이터 오픈플랫폼. 출처: https://data.kigam.re.kr/

Australian Ocean Data Network Portal. Available: https://portal.aodn.org.au/

Clarivate. Available: https://clarivate.com/webofsciencegroup/master-data-repository-list/

Donaldson, D. R. (2020). Certification information on trustworthy digital repository websites: a content analysis. PLoS One, 15(12), https://doi.org/10.1371/journal.pone.0242525

Downs, R. R. (2021). Improving opportunities for new value of open data: assessing and certifying research data repositories. Data Science Journal, 20(1), 1. https://doi.org/10.5334/dsj-2021-001

Environmental Information Data Centre. Available: https://eidc.ac.uk/

Environmental Information Data Centre. Available: https://portal.edirepository.org/nis/home.jsp

Frank, R. D. (2022). Risk in trustworthy digital repository audit and certification. Archival Science, 22(1), 43-73. https://doi.org/10.1007/s10502-021-09366-z

ICTS SOCIB Data Repository. Available: https://www.socib.es/data/

IFREMER-SISMER Portail de données marines. Available: https://data.ifremer.fr/

Khan, A. M., Loan, F. A., Parray, U. Y., & Rashid, S. (2024). Global overview of research data repositories: an analysis of re3data registry. Information Discovery and Delivery, 52(1), 53-61. http://dx.doi.org/10.1108/IDD-07-2022-0069

Khan, N., Thelwall, M., & Kousha, K. (2022). Are data repositories fettered? A survey of current practices, challenges and future technologies. Online Information Review, 46(3), 483-502. https://doi.org/10.1108/OIR-04-2021-0204

Lin, D., McAuliffe, M., Pruitt, K. D., Gururaj, A., Melchior, C., Schmitt, C., & Wright, S. N. (2024). Biomedical data repository concepts and management principles. Scientific Data, 11(1), 622. https://doi.org/10.1038/s41597-024-03449-z

Mosha, N. F. & Ngulube, P. (2023a). The utilisation of open research data repositories for storing and sharing research data in higher learning institutions in Tanzania. Library Management, 44(8/9), 566-580. https://doi.org/10.1108/LM-05-2023-0042

Mosha, N. F. & Ngulube, P. (2023b). Metadata standard for continuous preservation, discovery, and reuse of research data in repositories by higher education institutions: a systematic review. Information, 14, 427. https://doi.org/10.3390/info14080427

National Geoscience Data Centre. Available: https://www.bgs.ac.uk/geological-data/national-geoscience-data-centre/

Neotoma Paleoecology Database. Available: https://www.neotomadb.org/

NSF Arctic Data Center. Available: https://arcticdata.io/

PANGAEA. Available: https://www.pangaea.de/

re3data.org. Available: https://www.re3data.org/

Research Data Archive at NCAR. Available: https://rda.ucar.edu/

Schöpfel, J. (2022). Issues and Prospects for Research Data Repositories. Wiley Online Library. https://doi.org/10.1002/9781394163410.ch12

Stvilia, B. & Lee, DJ. (2024). Data quality assurance in research data repositories: a theory-guided exploration and model. Journal of Documentation, 80(4), 793-812.

Tethys. Available: https://www.tethys.at/

The CEDA Archive. Available: https://archive.ceda.ac.uk/

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

Cho, Jane & Park, Jong Do (2019). A study on analysis of research data repository in Humanities

and Social Sciences. Journal of the Korean Biblia Society for Library and Information Science, 30(2), 69-87. https://doi.org/10.14699/kbiblia.2019.30.2.069

Geo Big Data Open Platform. Available: https://data.kigam.re.kr/?lang=en

Kim, Juseop (2023). A study on the status of data repositories in the field of Geoscience and ways to obtain CoreTrustSeal certification: focusing on re3data.org and CoreTrustSeal best practices. Journal of Korean Library and Information Science Society, 54(2), 89-110. https://doi.org/10.16981/kliss.54.2.202306.89

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.

Rhee, Hea Lim (2023). A study on the preservation policy framework of data repository: focusing on CoreTrustSeal certification. Journal of the Korean Society for Library and Information Science, 57(4), 119-138. http://doi.org/10.4275/KSLIS.2023.57.4.119