

학문분야별 연구자들의 연구데이터 관리 현황에 관한 연구*

- 문헌정보학, 통계학, 생태학 및 한국음악학을 중심으로 -

A Study on the Current Status of Research Data Management by Researchers in Each Academic Field: Focusing on Library and Information Science, Statistics, Ecology, and Korean Musicology

김 주 섭 (Juseop Kim)** , 김 선 태 (Suntae Kim)***
한 연 중 (Yeonjung Han)**** , 유 원 재 (Won-Jae Youe)*****
전 바 울 (Paul Jeon)***** , 양 성 준 (Seong Jun Yang)*****

초 록

오픈 사이언스의 일환으로 정부는 국가 연구데이터 공유 및 활용 전략을 채택하고, 국가 차원의 거버넌스 체계를 구성하여 연구기관이 이행하도록 정책을 추진하고 있다. 연구데이터 관리를 위한 정책이 학계를 중심으로 수행되고 있지만 선진국에 비해 미흡한 실정이며 이마저도 현장에 도입하기에는 연구자의 인식도 부족한 현실이다. 본 연구의 목적은 학문분야별 연구자의 연구데이터 관리 현황을 파악하기 위함이다. 학문분야는 문헌정보학, 통계학, 생태학 및 한국음악학을 포함한 4개 분야로 구성하였으며 데이터 관리 현황은 설문조사를 통해 파악하였다. 연구데이터 관리 현황은 연구데이터 생산, 공유 및 관리, 저장, 보존 및 재사용 관점에서 분석하였다. 연구 결과, 데이터 저장을 제외하고 데이터 생산, 데이터 공유 및 관리, 데이터 보존 및 데이터 재사용의 경우 학문 분야별로 차이가 있는 것으로 파악되었다.

ABSTRACT

As part of open science, the government adopts a national research data sharing and utilization strategy, forms a governance system at the national level, and promotes policies for research institutes to implement. Although the policy for managing research data is carried out centering on the academic world, it is insufficient compared to developed countries, and even this is a reality in which researchers do not have enough awareness to introduce it to the field. The purpose of this study is to understand the research data management status of researchers in each academic field. Academic fields consisted of four fields including Library and Information Science, Statistics, Ecology, and Korean Musicology, and the current status of data management was identified through a survey. The current status of research data management was analyzed from the perspective of research data production, sharing and management, saving, preservation and reuse. As a result of the study, it was found that there were differences by discipline in terms of data production, data sharing and management, data preservation, and data reuse, except for data savings.

키워드: 오픈 사이언스, 연구데이터, 연구데이터 관리, 데이터 공유 및 활용, 데이터 재사용
Open Science, Research Data, Research Data Management, Data Sharing and Utilization, Data Reuse

* This research was supported by National University Promotion Program at Jeonbuk National University in 2021. 이 논문은 2021년 국립대학 육성사업의 지원을 받아 수행되었음.

** 전북대학교 문헌정보학과 강사, 연구데이터융복합연구소 전임연구원(kimjuseop@jbnu.ac.kr) (제1저자)

*** 전북대학교 문헌정보학과 부교수, 연구데이터융복합연구소장(kim.suntae@jbnu.ac.kr) (교신저자)

**** 국립산림과학원 임산자원이용연구부 임업연구사(yeonjung@korea.kr) (공동저자)

***** 국립산림과학원 임산자원이용연구부 임업연구사(sngkgk@korea.kr) (공동저자)

***** 전북대학교 통계학과 박사과정(vistadesk@naver.com) (공동저자)

***** 전북대학교 통계학과 부교수(응용통계연구소)(sjyang@jbnu.ac.kr) (공동저자)

논문접수일자 : 2022년 11월 21일 논문심사일자 : 2022년 11월 22일 게재확정일자 : 2022년 12월 14일

한국비블리아학회지, 33(4): 229-247, 2022. <http://dx.doi.org/10.14699/kbiblia.2022.33.4.229>

© Copyright © 2022 Korean Biblia 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

공공 기금 연구의 산출물을 법적, 제도적 및 경제적 장벽없이 공유하지는 움직임이 오픈 사이언스라는 운동으로 시작되어 확산되고 있다. 여기에서 산출물은 일반적으로 데이터를 말하며 이 데이터는 학술 영역에서 연구데이터라고 부르고 있다. 해외의 경우 정부를 중심으로 연구결과의 활용 및 재사용성 제고를 위해 데이터 관리 단일화, 공유체계 마련 등 데이터 관리 체계를 개선하고 있다. 또한, 연구개발 과정을 개방하고 연구데이터를 공유하는 오픈 사이언스가 유럽을 중심으로 확산되고 있다. 국내에서도 데이터 공유를 위해 연구데이터 정의를 법령에 추가하였으며, 데이터 관리 계획(Data Management Plan)을 한국연구재단을 중심으로 도입하였다. 또한, 한국과학기술정보연구위원회는 국가 차원의 연구데이터 플랫폼(DataON)을 구축하여 본격적으로 연구데이터를 수집하고 있다. 이러한 추진 배경에는 다음과 같은 법률 제정이 있었다.

- 국가연구개발혁신법 제정('20.5.) 및 시행령 제정령(안) 입법예고를 통해 시행('21.1.1.)
- 과기부, "연구데이터플랫폼"을 통한 공유·활용 체계 구축 전략 수립('18.1.)

우리 정부에서는 국가 연구데이터 공유 및 활용 전략을 채택하고, 국가 차원의 거버넌스 체계를 구성하여 연구관리기관과 연구수행기관이 이행하도록 정책을 추진하고 있다. 특히, 연구데이터 관리를 위한 연구가 학계를 중심으로 다수 수행되고 있지만 선진국에 비해 미흡한 실정

이며 이마저도 현장에 도입하기에는 연구자의 인식이 부족한 현실이다. 여기서 말하는 연구데이터는 연구개발과제 수행 과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실자료로서 연구결과의 검증에 필수적인 데이터를 말한다(국가연구개발정보처리기준). 이러한 연구데이터는 관측, 관찰, 실험, 시뮬레이션 그리고 컴퓨터를 통해 추출되거나 컴파일된다. 특히 우주 연구와 같은 일부 데이터의 경우 고비용 그리고 장시간 동안 연구한 결과로서 생산되기 때문에 보존의 가치가 높을 수밖에 없다. 이러한 연구데이터는 기관 및 국가 차원에서 반드시 관리 및 보존해야 할 필요성이 있다고 볼 수 있다. 또한, 연구데이터 관리를 통해 얻을 수 있는 장점은 다음과 같이 정리될 수 있다(University of Washington, 2022).

- 연구데이터 손실 방지 / 다른 연구자의 연구데이터 공유
- 검색, 접근 그리고 재사용 기능 / 연구에 투입되는 시간과 자원의 절약
- 연구의 오류방지 및 품질 향상 / 연구 결과 검증 및 복제 가능
- 연구데이터 공유로 인한 새로운 발견 가능

본 연구는 학문분야별 연구자들의 연구데이터 관리 현황을 파악하고자 시작되었다. 학문 분야에는 문헌정보학, 통계학, 생태학 및 한국음악학이 포함되었다. 연구대상을 위와 같이 한정된 이유는 우선, 국내외를 중심으로 과학기술 분야에 대한 연구데이터 관리 현황에 대한 연구는 지속적으로 이루어지고 있지만, 사회과학 분야와 예술 분야는 과학기술에 비해

연구가 많지 않았다. 특히, 예술 분야는 국내에서는 시도되지 않은 영역이다. 따라서 본 연구진에서는 사회과학 분야와 예술분야의 설문 대상은 연구진에서 비교적 구하기 쉬운 문헌정보학과 한국음악학으로 제한하였다. 연구데이터 관리 현황은 설문조사를 통해 파악하였으며 설문조사는 온라인을 통해 배포하였다. 설문조사는 문헌정보학, 통계학 및 한국음악학을 중심으로 진행하였다. 연구 대상인 4개 학문분야 중 생태학 분야 연구자들의 설문조사는 2년 전에 이루어진 결과를 본 연구에 활용하였다. 생태학을 대상으로 한 설문조사가 2020년에 진행된 연구라서 본 연구와의 시점이 맞지 않지만, 동일한 설문 문항으로 조사하였기 때문에 충분히 연구가 가능하다고 판단하였다. 하지만, 조사시

점의 차이점은 본 연구에 있어 제한점이라고 볼 수 있다. 제한점은 차치하더라도 4개 분야의 연구데이터 관리 현황을 조사한 연구라는 점에서 본 연구의 의의를 찾을 수 있을 것이다.

2. 이론적 배경

다음의 <표 1>은 국내외 연구데이터 관리와 관련된 선행연구를 요약한 것이다.

Donaldson과 Koepke(2022)는 5개 분야(대기 및 지구 과학, 컴퓨터 과학, 화학, 생태학, 신경과학) 과학자들의 포커스 그룹을 활용하여 데이터 관리에 대한 논의를 진행하였고, 그 결과 데이터 리포지터리 시스템 및 DMP의 데이

<표 1> 국내외 선행연구

저자(연도)	논문명	분야	시사점
Donaldson & Koepke (2022)	데이터 공유 및 연구데이터 관리에 대한 포커스 그룹 연구	대기 및 지구 과학, 컴퓨터 과학, 화학, 생태학, 신경과학	• 메타데이터 품질 관리 및 교육
Milewska et al. (2021)	연구데이터 관리 과정 및 데이터 사서의 역할에 대한 인식 미비를 나타내는 의료 연구자 대상 설문조사	의학	• 데이터 사서를 통한 RDM 지원
Raszewski et al. (2020)	간호학 박사 과정에서 데이터 관리 교육의 최근 사례 조사	간호학	• 데이터 라이프 사이클을 지원하기 위한 가이드라인과 인프라 구축 필요
박윤미, 김지현 (2021)	기계공학분야 연구자들의 연구데이터 생산과 관리에 관한 연구	기계공학	• 메타데이터, 맥락정보의 수집 필요성 그리고 인프라 구축 필요
김성욱, 김성태 (2020)	응집물리분야 연구데이터 관리 방안 연구	물리학	• 연구데이터 관리 개선방안
Wiley & Burnette (2019)	생명공학 및 생명의학 연구진의 데이터 관리 지원 요구사항 평가	생명공학 및 생명의학	• 도서관의 연구데이터 관리 교육 필요
Krahe et al. (2018)	연구데이터 관리 사례: 보건 및 의료 연구자의 평가 결과	보건 및 의학	• 연구데이터 관리 모범사례 적용 필요
Barabucci, Eschweiler, & Speer(2018)	TI-One: 현대 철학과의 능동적 연구데이터 관리	철학	• 연구데이터 관리 개선방안 도출
박미영, 안인자, 남승주 (2018)	과학기술분야 출연연구기관 연구데이터 관리 및 공유 사례 분석 연구	과학기술	• 연구데이터 관리 개선 방안
Wiley & Mischo (2016)	대기 과학자 및 공학 교수진의 데이터 관리 사례 및 관점	대기 과학 및 공학	• 도서관의 데이터관리 지원

터 공유 및 보존 부분을 구현하는 데 도움이 되는 서비스를 도출하였다. 포커스 그룹의 참가자들은 메타데이터 품질 관리 및 교육을 데이터 관리의 주요 문제 영역으로 식별하였으며 또한 메타데이터 제어, 데이터 추적성, 보안, 안정적인 인프라, 데이터 이용 제한 등 연구 커뮤니티가 원하는 리포지터리 기능에 대하여 논의하였다.

Milewska et al.(2021)은 온라인 설문 조사를 배포하고 의학 및 관련 분야를 대표하는 603명의 연구원으로부터 응답을 받았다. 그 결과, 데이터 공유는 광범위하지 않았고 많은 사람들이 연구데이터 공유의 가능한 이점을 언급했지만 표절 등과 같은 단점에 대해 우려한 것을 확인할 수 있었다. 많은 연구자와 마찬가지로 폴란드 의학자들은 공유를 위한 안전한 데이터 준비에 필요한 프로세스를 인식하지 못하고 있으며 도서관은 데이터 사서의 역할을 개발하여 연구자 교육을 지원해야 한다고 지적하였다.

Raszewski et al.(2020)은 간호학 박사 과정에서 데이터 관리 교육의 정도를 파악하기 위해 라이프 사이클 6단계(데이터 수집, 데이터 처리 및 분석, 데이터 출판, 데이터 공유, 데이터 보존 및 데이터 재사용)에 의거한 설문지를 개발하고 설문조사를 실시하였다. 저자는 해당 연구에서 데이터 라이프 사이클의 모든 단계를 다루는 데이터 업무의 전면 개편을 적용한 모범 사례 지침과 대학 또는 간호 데이터 관리 업무 부서의 전체 라이프 사이클을 감독할 데이터 관리자를 임명하는 등의 학생/교직원 지원 인프라 구축이 필요하다는 점을 결론으로 제시하였다.

박윤미, 김지현(2021)은 기계공학 분야 연구

자들의 연구데이터 생산 및 관리에 대한 인식과 경험을 조사한 후 해당 분야의 연구데이터 관리와 서비스를 위한 시사점을 제안하였다. 기계공학 분야 연구데이터 관리와 서비스를 위해 생산과정과 유형, 형태에 대한 데이터 조사를 실시하여 명시적 메타데이터와 암시적 맥락 정보의 수집이 필요하며 데이터 저널에 논문을 출판함으로써 연구실적으로 인정하는 방안을 제안하고 안전한 데이터 관리와 연구자들 간의 소통을 지원하는 클라우드 기반 시스템 등의 인프라 마련이 필요하는 점을 기술하였다. 또한 연구 현장의 다양한 관계자들이 조직적 차원의 연구데이터 관리와 서비스에 대한 역할과 책임을 배분하는 것이 중요함을 제안하였다.

김성욱, 김선태(2020)는 응집물리분야 연구데이터의 체계적 관리를 위한 개선방안을 제시하기 위해 14명의 연구자를 대상으로 설문조사를 실시하였다. 설문조사를 통해 수집된 데이터는 응답한 해당 연구자의 특성 및 기초정보, 데이터 보존 및 관리, 데이터 공유 및 접근에 관한 사항으로 구성되었다. 수집된 설문조사 데이터를 분석한 결과, 응집물리분야의 연구데이터 특성과 데이터 수집 및 생산, 데이터 보존과 관리, 데이터 공유 및 접근에 대한 9가지 문제점을 도출했으며 각 측면에서 도출된 문제점에 대한 개선방안을 제시하였다.

Wiley와 Burnette(2019)는 국립보건원이 편성한 연구 프로젝트에서 생명공학 및 생명의학 연구자의 데이터 관리 지식, 태도 및 업무를 조사하기 위하여 반구조화된 인터뷰를 실시했다. 인터뷰에 따르면 대다수의 연구자는 광범위한 연구 주제, 다양한 파일 저장 솔루션을 확인하고 수많은 양의 데이터를 생성하며 다양한 분

야별 관행을 고수하고 있음을 지적했다. 인터뷰는 파일 구성, 프로젝트 워크플로우 및 도큐멘테이션, 메타데이터 표준 및 데이터 저장 옵션에서 도서관이 교육 기회를 제공해야 한다고 기술하였다.

Krahe(2018)는 호주 학술 기관의 보건 및 의료 연구자를 대상으로 연구데이터 라이프 사이클 6단계에 맞춰 일반적인 RDM 작업과 관련된 온라인 설문 조사를 실시하였다. 해당 연구에서는 다음과 같은 결과 단계를 제시하고 있다. (i) 연구 계획; (ii) 데이터 수집; (iii) 데이터 처리 및 분석. 연구 결과는 계획 연구, 데이터 수집, 데이터 처리 및 분석과 관련된 작업이 보건 및 의료 연구자들 사이에서 크게 다르다는 것을 나타내고 있으며 또한 사용된 처리 및 분석 기술과 함께 연구데이터가 수집된 다양한 장소는 연구자 사이에서 RDM 업무의 복잡한 생태계를 반영하고 있음을 기술하였다.

Barabucci, Eschweiler와 Speer(2018)는 TI-One(켈른 대학 철학 학부의 연구 및 교육 부서인 Thomas-Institut의 데이터 관리 시스템)에 대한 디지털 데이터 관리 현황에 대해 조사했다. 연구 결과, 데이터를 공유하는 고유한 방법이 있었으나 일관된 저장 계획을 가지고 있지는 않았으며, 파일은 개인용 노트북, 플래시 드라이브 및 이메일에 분산되었고 대부분의 프로젝트에는 백업 전략이 부족하였고, 장기 보관 디지털 데이터는 프로젝트가 끝난 후에만 처리된다는 것을 지적하였다. 이러한 문제점을 해결하기 위해 전체 부서의 연구데이터 관리를 위한 중앙 집중식 공유 지침 세트를 상향식 방식으로 만드는 것을 목표로 TI-One 이니셔티브를 수립하였다.

박미영, 안인자, 남승주(2018)는 과학기술분야의 정부 출연연구기관 27개 기관의 데이터 업무 담당자를 대상으로 심층 인터뷰를 실시하였으며 이를 통해 9개의 우수기관에 대한 연구데이터 수집 및 관리현황 사례를 구체적으로 제시하였다. 해당 연구를 통해 도출된 시사점으로는 먼저 데이터 관리 규정 및 DMP와 같은 기관 차원의 연구데이터 수집 및 관리를 위한 제도 구축 그리고 연구데이터 관리 업무 수행을 위한 인력 확충 다음으로 연구자 참여를 독려하기 위한 보상 체계를 들 수 있다. 다음으로는 연구데이터가 개인 자산보다는 기관 및 국가 자산임을 인식시키는 연구자 인식 개선 그리고 데이터 공유 및 활용을 위한 국제표준에 부합하는 데이터 품질관리와 표준화에 대한 개선방안을 도출하였다.

Wiley와 Mischo(2016)는 연구 활동에서 교수진의 데이터 관리 업무와 요구 사항을 결정하기 위해 일리노이 어바나-삼페인 대학의 공학 및 대기 과학 교수진에 대한 21개의 심층 인터뷰를 분석하였다. 인터뷰에는 현재 연구 프로젝트 및 펀딩 출처, 데이터 유형 및 형식, 학문 및 기관 리포지터리의 이용, 데이터 공유, 대학도서관 데이터 관리 및 보존 서비스에 대한 인식, 펀딩 지원 기관의 검토 경험 등이 포함되었다. 해당 연구 결과 연구자들은 연방 기관 및 저널 출판사의 데이터 공유 의무에 대해 폭넓은 인식을 갖고 있으며 데이터 공유 가치에 대해 인식이 확산되고 있음을 나타냈다. 또한 데이터 관리 요구에 있어 학문적 차이는 중요하며 일관되고 성공적인 서비스를 위해 도서관의 역할이 중요함을 기술하였다.

이상의 연구데이터 관리 관련 10편의 국내외

연구를 분석한 결과, 철학을 제외하고 대부분이 과학기술에 한정되어 있음을 알 수 있었다. 또한 대부분 논문의 시자점은 연구데이터 관리 현황에 대한 문제점을 제시하고 개선방안을 도출하였으며 특히, 도서관에서 데이터 관리 교육 등을 지원하도록 하는 등 도서관 역할을 제시하는 제안도 확인할 수 있었다. 본 연구는 문헌정보학, 통계학, 생태학 및 한국음악학 등 인문사회와 자연과학 그리고 예술 분야까지 아우르는 학문분야의 연구데이터 관리 현황을 살펴보고 요구사항을 도출하고자 한다.

3. 학문분야별 특성 및 데이터 유형

본 장에서는 이번 연구의 대상인 ‘문헌정보학’, ‘통계학’, ‘생태학’ 및 ‘한국음악학’의 학문 특성 및 생산되는 데이터 유형에 대해서 살펴보고자 한다.

먼저, ‘사회과학’은 집단적 실체로서 인간의 활동에 초점을 맞추는 학문으로써 타 학문분야보다 출현시기가 늦어 학문의 최신 분야라고 할 수 있다. 환경과 현상을 연구하는 자연과학과는 다른 인간을 연구하는 학문으로서 인간의 특정 인지 능력을 통해 행동에 영향을 미치고 복잡한

상호작용 패턴을 생성하는 학문이다. 인간에 대해 연구하는 학문으로 복잡한 특징을 가지고 있으며 인간의 의지나 지향성이 개입된다. 학문이 가지는 복잡성과 의도성으로 인하여 인간 행동에 대한 일반화가 쉽지 않다(Collaborative Research Group(CRG) USA, 2016). 다음 <표 2>는 사회과학에서 생산되는 데이터 유형과 포맷을 나타낸 것이다(UCLA Library, n.d.)

다음으로 통계학은 도구적인 성격이 강한 학문으로 데이터의 생산 및 분석이 요구되는 모든 분야에 걸쳐서 활용되고 있다. 이는 국가통계, 자연과학뿐 아니라 계량적인 방법을 사용하는 인문학, 경영학 등도 포함한다. 또한 통계학은 시대의 요구에 따라 발전하는 학문으로 시의성이라는 특성을 지니고 있다. 과거에는 조사통계, 실험계획 등 연구자의 주도적인 설계 하에 자료를 수집하고 이를 가공 및 분석하는 방법에 대한 연구가 주류였고, 2000년대 전후로 인간 유전체 연구가 활성화되면서 생물통계(Biostatistics)가 크게 각광받기도 하였다. 최근에는 데이터마이닝, 빅데이터, 머신러닝 관련 분야가 활성화되면서 대용량 자료의 분석이나 예측모델링 등에 대한 연구에 많은 자원이 집중되고 있다. 다음 <표 3>은 통계학에서 생산되는 데이터 유형과 포맷을 정리한 것이다.

<표 2> 사회과학 분야 데이터 유형과 포맷

데이터 유형	데이터 포맷
<ul style="list-style-type: none"> • 설문조사 데이터(여론조사, 투표 기록) 및 비설문조사 데이터(문서, 지도, 사운드, 비디오, 멀티미디어) • 원시 측정값, 수치 표, 정부 통계 및 지수 • 텍스트(필드 노트, 스크립트, 블로그, 이메일) • 공간(영역, 이벤트, 공간 참조) 	<ul style="list-style-type: none"> • 텍스트 - ASCII, Word, PDF • 수치 - ASCII, Excel, Access, SPSS, STATA • 멀티미디어 - jpeg, tiff, dicom, mpeg, quicktime • 공간 - geotiff, KLM, KMZ • 모델 - 3D, statistical

1) 문헌정보학은 사회과학 분야 학문으로서 본 연구에서는 사회과학의 학문 특성 및 데이터 유형에 대해 살펴보았다.

〈표 3〉 통계학 분야 데이터 유형과 포맷

데이터 유형	데이터 포맷
<ul style="list-style-type: none"> • 이미지 데이터 • 함수 데이터 • 텍스트 데이터 • 지리정보 포함 데이터 등 	<ul style="list-style-type: none"> • 텍스트 - txt • 이미지 - PNG, JPEG 등 • 스프레드시트 - csv, xls, dat, txt 등

빅데이터 시대와 함께 데이터의 형태 및 유형 또한 다양하게 인식되고 있어 그에 따른 연구들이 진행되고 있다. 대표적으로는 이미지 데이터, 함수 데이터, 텍스트 데이터, 지리 정보포함 데이터 등을 들 수 있다. 위에서 기술된 다양한 유형의 데이터들 역시 실제 분석목적에 위해서는 표준적인 2차원 데이터테이블 혹은 그와 유사한 형태로 변환 및 처리하는 것이 보통이다. 가장 흔하게 채택하고 있는 파일 형식은 2차원 데이터 테이블의 저장 및 관리에 용이한 스프레드시트 형태의 데이터 파일이다. 이 파일들은 통계분석을 위한 소프트웨어 R/Python/SAS/SPSS/MATLAB을 이용하여 입출력이 용이한 형태로 볼 수 있다. 또는 dat나 txt 형태로 파일을 관리하기도 하는데 이 또한 스프레드시트형태의 연장선상에 있는 것으로 볼 수 있다. 최근에는 머신러닝 및 인공지능 관련 연구를 위해 이미지 데이터가 많이 활용되고 있다. 덧붙여, 텍스트 데이터 분석을 위해서는 txt 형태의 파일이 주로 활용된다.

생태학은 일정한 환경 내에서의 생물군 또는 생물군 집단의 풍부성, 분포, 상관성 등을 연구하는 생물학의 분야이다. 단일 유기체의 기관과 그 생명 과정보다는 유기체를 둘러싼 환경과 다른 유기체와의 상호 작용을 연구한다(노태호 외, 2000). 생태 분야의 데이터의 특성을 살펴보면 다음과 같이 정리될 수 있다(DBBM, n.d.).

- 복잡성: 다른 도메인보다 복잡하여 데이터의 관계뿐만 아니라 복잡한 하부 구조를 표현할 수 있어야 함
- 다양성: 다양한 분야의 종합적인 연구를 지원할 필요가 있어 범위가 넓으며 데이터양 또한 많음
- 무질서성: 같은 데이터에 대해 개별 연구자들이 각자 상이하게 표현하는 경우가 있음. 따라서 상이한 생물학 스키마를 호환하는 방법이 필요함
- 맥락정보 관리 필요성: 해당 도메인의 데이터가 복잡하게 구성되어 있어 최대한 해석하려면 가능한 한 맥락정보가 필요함

생태분야 데이터는 타 도메인보다 복잡하므로 복잡한 하부 구조를 표현하기 위하여 모델링 시 정보가 손실되지 않도록 주의해야 한다. 또한, 데이터가 범위가 넓으므로 다양한 유형과 값을 기술할 수 있도록 유연성을 지녀야 한다. 해당 데이터베이스 스키마가 빠르게 변화하므로 스키마 진화와 데이터 마이그레이션을 지원해야 한다. 또한 개별 연구자들의 상이한 표현을 위해 호환할 수 있는 스키마 개발이 필요하다. 특히 데이터 해석을 위해 적시에 맥락정보가 유지, 관리 및 전달되는 것이 중요하다.

마지막으로 한국음악으로 대표되는 한국전통 음악은 한국에서 전통적으로 전해 내려오는 민

요, 음악 등을 말하며 일반적으로 전통 악기를 바탕으로 연주되면 해당 범주에 포함시키고 있다. 특히 한국전통음악은 음악자료의 특성을 지니고 있는데 다음과 같이 정리할 수 있다(이혜원, 김태수, 2007).

“시간의 흐름에 따라 개체 자체에 대한 변화가 가장 많이 일어난다. 하나의 곡이 작곡되어서 특정 시간에 연주되고, 다른 자료 유형을 통해 생산되어지는 생명주기를 가지고 있다. 그러므로 음악 자원은 이벤트 즉 시간의 흐름으로 개체 자체의 변화가 뚜렷하며, 새로운 이벤트에 대한 생성이 계속 일어난다고 할 수 있다.”

최근에는 악보와 같은 인쇄 형태의 저작물을 PDF 등으로 변환하거나 개정, 개작, 해설 비평, 복제 등 파생되는 저작물이 급격히 늘어나고 있다. 이러한 음악은 일반적으로 들을 수 있는 녹음 자료와 연주와 공연을 촬영한 영상자료로 구분되고 음악 자체를 악보와 음악에 대한 이문서, 해설서 등의 문헌자료 및 공연을 안내하는 팸플릿 등의 인쇄자료로 구성될 수 있다(이성숙, 2010).

4. 학문분야별 연구자들의 연구데이터 관리 현황

본 연구는 학문 분야별 연구데이터 관리 현황을 파악하기 위하여 문헌정보학, 통계학, 생

태학 및 한국음악학 연구자를 대상으로 하였다. 해당 목적을 위해 선택된 학문분야의 연구자들을 대상으로 한 설문조사를 실시하였다. 다음은 설문조사의 개요를 나타낸 것이다.

• 설문조사 개요

- (1) 설문목적: 학문 분야별 연구자들의 연구데이터 관리 현황 조사
- (2) 설문대상: 문헌정보학, 통계학, 생태학 및 한국음악 전공 연구자들(115명)
 - 문헌정보학: 정부출연연구소 소속 연구자
 - 통계학: 통계청 및 국내 대학 소속 연구자
 - 생태학: 국립생태원 소속 연구자
- (3) 설문방법: DAF를 활용한 온라인 설문조사
- (4) 설문기간 및 응답인원
 - 생태학: 2020년 09월 10일 ~ 2020년 09월 21일 / 30명
 - 문헌정보학, 통계학, 한국음악: 2022년 07월 25일 ~ 2022년 08월 20일 / 85명

설문지는 2009년에 작성된 DAF(Data Asset Framework, 이하 DAF)²⁾ 실행 가이드의 내용을 준용하였으며, 국내 실정과 맞지 않는 부분은 수정하여 배포하였다. 응답자는 총 115명으로, 각 분야별로 박사학위 및 박사과정³⁾ 이상의 가진 연구자를 대상으로 하였다. 이 중 생태학은 2020년에 진행한 설문조사를 활용하였다. 설문대상자의 경우 문헌정보학, 통계학 및

2) DAF는 조직이 연구데이터 자산을 어떻게 관리하고 있는지 데이터의 식별, 위치, 설명 및 평가할 수 있는 수단을 제공하며, 이를 위해 연구데이터 지원 서비스의 범위를 넓히는 데 필요한 정보, 견해 및 경험을 수집하는 일련의 방법을 제공한다(JISC, University of Glasgow, & DCC, 2009).

3) 한국음악의 경우에는 박사학위 이상의 연구자가 많지 않아 박사과정 이상으로 설문 대상자를 확보하였다.

생태학은 대학 및 전문 연구기관의 연구자를 대상으로 하였으며 한국음악학은 시도 국악단에 소속된 실연자를 대상으로 하였다. 설문조사 기간의 연도 차이가 나지만 기존 조사를 활용한 이유는 동일한 설문조사 양식인 DAF를 사용하였기 때문에 타 학문 분야와 비교하는 연구가 가능하다고 판단하였기 때문이다.

설문 문항은 항목별로 단/복수 응답이 가능하도록 설계하였으며, 문항은 크게 연구데이터 생산, 연구데이터 공유 및 관리, 연구데이터 저장, 연구데이터 보존 및 연구데이터 재사용 이렇게 5개의 항목별로 관련된 문항끼리 배치하여 구성하였다. 설문조사 응답 결과를 활용하여 연구데이터 관리 현황에 대한 학문 분야에 따른 차이를 피셔의 정확 검정과 카이제곱검정을 이용하는 교차분석을 통해 확인할 수 있는데 교차분석을 통한 검정 과정에서의 가설은 다음과 같다.

- 가설: 학문분야별 연구데이터 관리 현황은 차이가 없다.

4.1 연구데이터 생산

학문분야 별 연구자들의 연구데이터 관리 현황 중 생산에 대한 질문에 대한 결과는 다음과 같다. 우선, 연구자들의 데이터 생성 방법을 질문한 결과는 <표 4>와 같이 얻을 수 있었다.

응답자 전체를 기준으로 통계분석을 통해 데이터를 생성한다고 응답한 수가 제일 많았으므로 조사, 관찰, 실험 순으로 많이 나타났다. 생태학 분야의 연구자들은 통계분석이나 조사로 데이터 생성을 한다고 답한 경우가 거의 없었고 대부분 관찰과 실험을 통한다고 응답하였다. 문헌정보학과 통계학 연구자들은 통계분석과 조사를 통해 데이터를 얻는 것으로 나타났다. 한국음악 분야의 연구자들은 공연을 통한 데이터 생성이 높은 비율을 보였다.

학문분야에 따라 차이 유무를 보기 위해 각 분야에 따라 데이터 생성 방법에는 차이가 없다는 귀무가설을 세운 후 교차분석을 실시하였다. 각 기댓값이 5미만인 셀의 비율이 높아 카이제곱분포를 활용한 검정이 아닌 피셔의 정확

<표 4> 학문분야별 데이터 생산 방법(복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
관찰	8(12.50%)	23(47.91%)	4(7.01%)	4(7.40%)	39(17.48%)
실험	3(4.68%)	15(31.25%)	3(5.26%)	0(0.00%)	21(9.41%)
조사	25(39.06%)	1(2.08%)	10(17.54%)	15(27.77%)	51(22.86%)
통계분석	24(37.50%)	5(10.41%)	21(36.84%)	6(11.11%)	56(25.11%)
시뮬레이션	4(6.25%)	2(4.16%)	19(33.33%)	2(3.70%)	27(12.10%)
공연	0(0.00%)	0(0.00%)	0(0.00%)	27(50.00%)	27(12.10%)
기타	0(0.00%)	2(4.16%)	0(0.00%)	0(0.00%)	2(0.89%)
전체	64(100%)	48(100%)	57(100%)	54(100%)	223(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수			P-value		
2000			0.0004998		

검정을 수행하였다. 전체 table의 수가 많기 때문에 Monte Carlo 시뮬레이션을 통해 p-value를 계산하였다. 그 결과 p-value가 0에 가까울 정도로 낮은 값이 나왔고 따라서 귀무가설을 기각하게 된다. 즉 학문분야에 따라 데이터 생성 방법에는 차이가 있다고 할 수 있다.

연구데이터 현황 및 편당기관에 대한 문항들을 살펴본 결과 분야에 따라 상이한 면이 있는 것으로 보인다. 이는 분야별로 다루는 학문 주제가 다름에 따라 연구데이터의 현황 또한 상당히 달라진다고 볼 수 있다. 따라서 융복합 연구일 경우 연구자들은 학문분야별로 생산되는 연구데이터가 많은 차이가 있는 것을 숙지하고 있어야 한다.

4.2 연구데이터 공유 및 관리

다음의 <표 5>는 학문분야별 공동 작업 유무에 대한 문항의 응답 결과를 나타낸 것이다.

한국음악과를 제외한 나머지 분야의 연구자들은 공동 작업을 하는 연구자들이 대부분이었다. 한국음악과의 경우 다른 분야에 비해서는 공동 작업을 하지 않는 연구자들의 비율이 높은 것을 확인할 수 있었다. 가설 검정을 활용하여 분야별로 공동 작업 유무에 차이가 있는지 확인한 결과 역시 위의 표에서 확인할 수 있다. 앞의

문항과 다르게 공동 작업 유무에 대한 문항의 응답 결과는 각 응답이 카이제곱 검정을 수행할 정도의 기대빈도를 가지므로 카이제곱 분포를 활용한 검정을 수행하였다. 그 결과 p-value가 0.08692로 유의수준 0.05 하에서 귀무가설을 기각할 수 없게 되므로 각 분야에 따라 공동 작업 유무는 차이가 없다는 결론을 얻을 수 있다.

다음의 <표 6>은 학문분야별 공유방법에 대한 문항의 응답 결과를 나타낸 것이다.

먼저, 전체 응답자들의 빈도를 봤을 때 이메일로 공유를 하는 경우가 가장 많은 것을 알 수 있다. 다음으로 외부스토리지 서비스와 CD, DVD, USB 등을 사용하는 순서로 나타났다. 모든 분야에서 이메일로 데이터를 공유하는 응답자의 수가 가장 많았지만 나머지 공유 방법들에 대한 선호도는 학문 분야에 따라 상이한 것으로 보였다. 학문분야별 연구데이터 공유 방법에 차이 유무를 보기 위해 각 분야에 따라 연구데이터 공유 방법에는 차이가 없다는 가설을 세우고 피셔의 정확 검정을 통한 검정 결과 p-value가 낮은 값이 나왔고 그 결과 귀무가설을 기각하게 된다. 따라서 각 분야별 연구데이터 공유 방법에는 차이가 있다고 할 수 있다.

마지막으로 공동 작업 시 발생하는 문제에 대한 결과는 다음의 <표 7>과 같다.

<표 5> 학문분야별 공동작업 유무

보기	문헌정보학	생태학	통계학	한국음악	전체
예	24(77.4%)	24(80%)	18(75%)	16(53.3%)	82(71.30%)
아니오	7(22.6%)	6(20%)	6(25%)	14(46.7%)	33(28.69%)
전체	31(100%)	30(100%)	24(100%)	30(100%)	115(100%)
카이제곱검정 결과					
검정통계량	자유도		P-value		
6.5705	3		0.08692		

〈표 6〉 학문분야별 데이터 공유 방법(복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
이메일	21(42.9%)	22(52.4%)	12(41.4%)	14(53.8%)	69(47.26%)
외부 스토리지 서비스	17(34.8%)	5(11.9%)	10(34.5%)	3(11.5%)	35(23.97%)
소속기관 스토리지 서비스	6(12.2%)	5(11.9%)	5(17.2%)	0(0%)	16(10.95%)
CD, DVD, USB 등 사용	4(8.1%)	8(19%)	2(6.9%)	7(27%)	21(14.38%)
기타	1(2%)	2(4.8%)	0(0%)	2(7.7%)	5(3.42%)
전체	49(100%)	42(100%)	29(100%)	26(100%)	146(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수			P-value		
2000			0.02549		

〈표 7〉 학문분야별 공동 작업 시 발생한 문제(복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
없음	4(8.3%)	7(16.7%)	6(20%)	8(36.4%)	25(17.60%)
소속기관 스토리지 서비스 기능 한계	2(4.2%)	9(21.4%)	2(6.7%)	0(0%)	13(9.15%)
외부 스토리지 서비스 기능 한계	8(16.7%)	4(9.5%)	6(20%)	0(0%)	18(12.67%)
데이터 이름 생성 규칙 부재	9(18.7%)	5(12%)	2(6.7%)	1(4.5%)	17(11.97%)
데이터 버전 관리의 어려움	11(23%)	8(19%)	6(20%)	4(18.3%)	29(20.42%)
데이터 전송 관련 법적 문제	2(4.1%)	4(9.5%)	3(10%)	3(13.6%)	12(8.45%)
데이터 소유권 문제	6(12.5%)	4(9.5%)	1(3.3%)	3(13.6%)	14(9.85%)
보안 문제	6(12.5%)	0(0%)	4(13.3%)	3(13.6%)	13(9.15%)
기타	0(0%)	1(2.4%)	0(0%)	0(0%)	1(0.70%)
전체	48(100%)	42(100%)	30(100%)	22(100%)	142(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수			P-value		
2000			0.01699		

전체 응답자들의 빈도를 보면 공동 작업 시 발생한 문제가 없다고 답한 경우가 가장 많았고 다음으로 데이터 버전 관리의 어려움, 데이터 이름 생성 규칙의 부재 순으로 많았다. 한국음악과의 경우 데이터 공유 방법에 있어서 스토리지 서비스를 거의 이용하지 않는다는 것을 앞선 문항에서 확인할 수 있었는데, 이 때문에 공동 작업 시 발생 문제의 원인으로 외부, 소속 기관 스토리지 서비스라고 응답한 경우가 단

한명도 없었다.

학문분야별 공동 작업 시 발생하는 문제 유형에는 차이가 없다는 귀무가설을 세우고 피셔의 정확 검정을 이용한 교차분석 결과 p-value가 0.01699로 유의수준 0.05하에서 귀무가설을 기각하게 된다. 따라서 공동 작업 시 발생하는 문제의 유형은 학문분야 마다 차이가 있다고 할 수 있다.

4.3 연구데이터 저장

연구데이터 저장에 관련된 설문 문항들 중 연구데이터의 크기로 인한 저장 문제에 대해 극복 방법을 묻는 문항의 응답 결과는 다음의 <표 8>과 같이 얻을 수 있었다.

먼저 전체 응답의 빈도를 살펴보면 추가 저장 장치 구입의 빈도가 가장 많은 것을 확인할 수 있었다. 다음으로 외부 스토리지 서비스 이용을 통해 극복한다는 응답이 많았다. 또, 모든 분야에서 소속기관의 스토리지를 이용하거나 소속기관에게 저장 공간을 요청하여 저장 문제를 극복한다는 응답은 없거나 미미하다는 것을 볼 수 있었다. 교차분석의 검정 가설은 각 분야

별로 저장 문제를 극복하는 방법에는 차이가 없다이며 그 검정 결과 p-value가 유의수준 0.05 하에서 귀무가설을 기각하기 어렵기 때문에 ‘연구데이터 저장 문제로 인한 극복 방법에는 분야마다 차이가 없다’라는 결론을 내릴 수 있다.

4.4 연구데이터 보존

학문 분야에 따른 연구자들의 연구데이터 보존 현황 항목의 문항들 중 먼저 연구데이터의 백업 여부에 대한 문항의 응답 결과는 다음의 <표 9>와 같다.

분야에 상관없이 응답자들 전체의 빈도만 봤을 때는 대부분 데이터의 일부만 백업을 한다

<표 8> 학문분야별 저장 문제 극복 방법 (복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
외부 스토리지 서비스	5(38.5%)	2(9.5%)	6(46.1%)	2(22.2%)	15(26.78%)
소속기관 스토리지 서비스	0(0%)	3(14.3%)	0(0%)	0(0%)	3(5.35%)
추가 저장 장치 구입	7(53.8%)	12(57.1%)	5(38.5%)	7(77.8%)	31(55.35%)
소속 기관에 저장 공간 요청	1(7.7%)	3(14.3%)	2(15.4%)	0(0%)	6(10.71%)
기타	0(0%)	1(4.8%)	0(0%)	0(0%)	1(1.78%)
전체	13(100%)	21(100%)	13(100%)	9(100%)	56(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수			P-value		
2000			0.2789		

<표 9> 학문분야별 연구데이터 백업 여부

보기	문헌정보학	생태학	통계학	한국음악	전체
전체 백업	8(25.8%)	8(26.7%)	11(45.8%)	8(26.7%)	35(30.43%)
일부 백업	20(64.5%)	21(70.0%)	9(37.5%)	18(60%)	68(59.13%)
백업 하지 않음	3(9.7%)	1(3.3%)	4(16.7%)	4(13.3%)	12(10.43%)
전체	31(100%)	30(100%)	24(100%)	30(100%)	115(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수			P-value		
2000			0.2549		

고 응답하였다. 통계학과 연구자들의 경우 데이터 전체를 백업한다는 응답이 일부만 백업한다는 응답보다 많이 나왔음을 확인할 수 있었다. 학문 분야에 따라서 연구자들이 연구데이터 백업을 하는지에 대한 정보가 차이가 없다는 가설을 세우고 피서의 정확 검정을 활용하여 교차분석을 실시하였다. p-value는 0.05보다 높은 값이 나와서 유의수준 0.05 하에서 귀무가설을 기각할 수 없게 된다. 즉, 학문 분야에 따라 연구자들의 데이터 백업여부에는 차이가 없다는 결과를 얻게 된다. 다음으로 데이터의 백업 위치에 대한 문항으로 응답 결과는 다음의 <표 10>과 같다.

다음으로는 데이터의 백업 위치에 대한 문항으로 응답 결과는 <표 4-12>에 나와 있다. 전체 응답의 빈도를 살펴보면 이동식 저장장치(USB/SD 카드 등)의 응답이 가장 많았고 출력물과 기타 응답을 제외한 나머지들은 비슷한 수준의 빈도를 보였다. '각 분야에 따라서 연구자들의 데이터 백업 위치가 차이가 없다'는 가설을 세운 후

교차분석을 통한 검정 결과 역시 <표 4-12>에 나와 있다. p-value가 0에 가까울 정도로 낮은 값을 보이므로 기존에 세웠던 가설을 기각하게 된다. 따라서 각 분야에 따라 연구자들이 데이터를 백업하는 위치는 차이가 있다고 할 수 있다. 다음의 <표 11>은 연구데이터의 소속기관 보존 의무에 대한 결과를 나타낸 것이다.

전체 응답은 소속기관이 보존해야 한다는 응답이 좀 더 많은 빈도를 보였으나, 학문 분야에 따라 상이한 빈도를 보였다. 문헌정보학, 생태학 연구자들은 전체 응답의 양상과 마찬가지로 소속기관이 연구데이터를 보존해야 한다는 의견이 지배적이었으나, 통계학과 한국음악 연구자들은 반대의 의견이 더 많았다. 위의 빈도분석 결과를 검정을 통해 상세히 알아보고자 연구데이터를 소속기관이 보존해야 한다고 생각하는지에 대해서는 학문 분야별로 의견 차이가 없다고 귀무가설을 설정한 후 교차분석을 실시하였다. 그 결과 p-value가 매우 작은 값이 나왔고 그에 따라 귀무가설을 기각하였다. 따라서 데이터를 소속기

<표 10> 학문분야별 연구데이터 백업 위치 (복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
출력물	3(4.5%)	1(2.1%)	0(0%)	4(7.7%)	8(3.92%)
개인 pc(집)	5(7.6%)	2(4.3%)	2(5.1%)	14(26.9%)	23(11.27%)
개인 pc(사무실/연구실)	18(27.3%)	4(8.6%)	10(25.6%)	4(7.7%)	36(17.64%)
외부 스토리지 서비스	9(13.6%)	5(10.6%)	10(25.6%)	3(5.8%)	27(13.23%)
소속 기관 스토리지 서비스	4(6.1%)	12(25.5%)	3(7.8%)	1(1.9%)	20(9.80%)
이동식 저장장치 CD/DVD/ 외장 하드 드라이브	17(25.8%)	11(23.4%)	9(23.1%)	17(32.7%)	54(26.47%)
기타	9(13.6%)	8(17%)	5(12.8%)	9(17.3%)	31(15.19%)
기타	1(1.5%)	4(8.5%)	0(0%)	0(0%)	5(2.45%)
전체	66(100%)	47(100%)	39(100%)	52(100%)	204(100%)

피서의 정확 검정 결과	
Monte Carlo 시뮬레이션 수	P-value
2000	0.0004998

〈표 11〉 학문분야별 소속기관 보존 의무

보기	문헌정보학	생태학	통계학	한국음악	전체
예	21(67.7%)	25(83.3%)	8(33.3%)	13(43.3%)	67(58.26%)
아니오	10(32.3%)	5(16.7%)	16(66.7%)	17(56.7%)	48(41.73%)
전체	31(100%)	30(100%)	24(100%)	30(100%)	115(100%)
카이제곱검정 결과					
검정통계량	자유도		P-value		
17.783	3		0.0004876		

관이 보존해야 하는지에 대한 생각은 분야별로 다르다는 결론을 얻을 수 있었다.

4.5 연구데이터 재사용

연구데이터 재사용에 관련된 설문 문항들 중 먼저 연구데이터 재사용이 가능한지의 여부를 묻는 문항의 응답을 정리한 결과는 다음의 〈표 12〉와 같다.

재사용이 불가능하다고 응답한 경우는 모두 재사용 가능하다는 응답과 일부 재사용 가능하다는 응답에 비해 훨씬 적은 빈도를 보였다. 또 모든 분야에 대해서 일부 재사용 가능하다고 응답한 빈도가 모두 재사용 가능하다고 응답한 빈도보다 높은 것을 확인할 수 있었다. '각 분야에 따라 연구데이터의 재사용 여부는 차이가

없다'라는 가설을 세우고 그에 대한 교차분석 결과 p-value가 유의수준 0.05하에서 귀무가설을 기각할 수 없었다. 따라서 연구데이터의 재사용 여부는 학문 분야에 따라 차이가 없다고 할 수 있다. 다시 정리하면 분야에 따른 차이가 없이 대부분의 연구자들이 가진 연구데이터가 일부 재사용 가능하거나 모두 재사용 가능하다. 다음의 〈표 13〉은 재사용할 수 없는 이유에 대해 나타난 것이다.

전체 응답들 중 기밀성 또는 데이터 보호 문제로 인해 재사용이 불가능하다는 응답이 가장 큰 빈도를 보였다. 다음으로는 타 연구자가 이해할 수 없는 데이터, 라이선스 계약으로 공유 금지 순서로 많은 응답을 보였다. 분야별로 살펴봤을 때 특이한 점으로는 문헌정보과 연구자들은 기밀성 또는 데이터 보호 문제보다는

〈표 12〉 학문분야별 연구데이터 재사용 가능 여부

보기	문헌정보학	생태학	통계학	한국음악	전체
모두 재사용 가능	5(16.1%)	7(23.3%)	10(41.6%)	14(46.7%)	36(31.30%)
일부만 재사용 가능	24(77.4%)	22(73.3%)	12(50%)	15(50%)	73(63.47%)
재사용 불가능	2(6.5%)	1(3.3%)	2(8.4%)	1(3.3%)	6(5.21%)
전체	31(100%)	30(100%)	24(100%)	30(100%)	115(100%)
피셔의 정확 검정 결과					
Monte Carlo 시뮬레이션 수					P-value
2000					0.09745

〈표 13〉 학문분야별 연구데이터 재사용할 수 없는 이유(복수 응답)

보기	문헌정보학	생태학	통계학	한국음악	전체
기밀성 또는 데이터 보호 문제	13(35.2%)	17(60.7%)	11(61.1%)	10(45.5%)	51(48.57%)
라이선스 계약으로 공유 금지	6(16.2%)	3(10.7%)	4(22.2%)	7(31.9%)	20(19.04%)
타 연구자가 이해할 수 없는 데이터	14(37.8%)	3(10.7%)	3(16.67%)	3(13.6%)	23(21.90%)
현재의 소프트웨어로 인식할 수 없음	0(0%)	1(3.6%)	0(0%)	1(4.5%)	2(1.90%)
기타	4(10.8%)	4(14.3%)	0(0%)	1(4.5%)	9(8.57%)
전체	37(100%)	28(100%)	18(100%)	22(100%)	105(100%)

피서의 정확 검정 결과

Monte Carlo 시뮬레이션 수	P-value
2000	0.08396

타 연구자가 이해할 수 없는 데이터라는 이유가 더 많은 응답을 보였다는 것이 있었다. 학문 분야에 따라서 연구데이터 재사용이 불가능한 이유의 차이 유무를 보기 위해 분야에 따라 연구데이터 재사용이 불가능한 이유가 차이가 없다는 가설을 세운 후 교차분석을 진행하였다. 그 결과 p-value가 0.08396로 유의수준 0.05하에서 귀무가설을 기각할 수 없다. 따라서 연구데이터 재사용이 불가능한 이유는 학문 분야에 따라 차이가 없다는 결론을 내릴 수 있다.

4.6 종합

본 절에서는 4장 1절부터 5절까지 분석한 내용을 가설 검정을 기준으로 정리하고 시사점을 도출하고자 한다. 우선, 연구데이터 생산에서는 통계 분석을 통해 데이터를 생성한다는 응답이 가장 많은 것을 확인할 수 있었다. 데이터 생성 방법의 경우, 학문분야별로 차이가 있었으며 특히, 한국음악의 경우 공연을 통한 데이터 생성이 높았다. 연구데이터 공유 및 관리의 경우, 이

메일로 공유한다는 의견이 가장 많았지만 공유 방법에 대해서는 학문 분야에 따라 다른 것으로 나타났다. 연구자간 공동 작업 시 발생하는 문제에 대해서는 데이터 버전 관리 및 데이터 이력 생성 규칙이 가장 많은 것으로 나타났다. 또한, 공동 작업 시 발생하는 문제도 학문 분야마다 차이가 있는 것으로 파악되었다. 연구데이터 저장에서는 데이터 크기로 인하여 발생한 문제에 대하여 추가 저장 장치를 구입하여 해결하는 경우가 가장 많은 것으로 확인되었다. 학문분야별로 연구데이터 저장 문제를 극복하는 방법에는 차이가 없는 것으로 나타났다. 다음으로 보존의 경우, 데이터 일부만 백업한다는 의견이 가장 많았고, 학문분야에 따라 백업여부가 차이가 있는 것으로 나타났다. 또한, 데이터 백업 위치의 경우 USB와 같은 이동식 저장장치가 가장 많았으며 학문 분야별로 차이가 있는 것으로 확인되었다. 소속기관이 연구데이터를 보존해야 하는지에 대해서는 소속기관이 보존해야 한다는 빈도가 조금 더 많았으나 학문분야별로 차이가 있는 것으로 나타났다. 다음으로 연구데이

터 재사용에서는 데이터 재사용이 불가능하다는 빈도가 가장 많았으며 재사용이 불가능한 이유는 기밀성 또는 데이터 보호 문제와 연구자가 이해할 수 없는 데이터가 가장 많았다. 해당 항목에 대해서도 학문분야별로 차이가 없는 것으로 확인되었다.

다음은 연구데이터 현황을 파악하면서 확인한 문제점과 개선방안에 대해서 기술하고자 한다. 우선, 데이터를 이메일로 공유한다는 의견에서 이메일 보다는 소속기관의 스토리지 또는 실시간으로 공유 가능한 웹 저장 매체를 이용하는 것이 데이터 저장 및 공유 차원에서 더 나을 수 있다. 또한, 공동 작업 시 데이터 버전 관리와 데이터 이름 생성 규칙이 힘들다는 의견에서는 연구자 대상으로 한 연구데이터 관리 교육이 필요하다고 판단된다. 연구데이터 크기로 인한 문제를 극복하는 방안에 대해 추가 저장 장치가 가장 많은 것으로 보아 연구자를 위해 소속기관에서 대용량 스토리지를 구입하여 제공하는 서비스를 제공하는 것도 하나의 방법이 될 수 있을 것이다. 다음으로 데이터 백업 위치도 USB와 같은 이동식 저장장치라고 답한 의견이 많아 데이터 멸실 위험이 큰 것으로 파악되었다. 이 문제를 해결하기 위해 소속 기관에서는 연구데이터 관리 규정을 통해 생산된 모든 연구데이터는 보존이 가능하게 시스템을 지원해야 할 것이다. 마지막으로, 연구자가 보유하고 있는 연구데이터를 재사용 불가능하다는 의견이 많았으며 또한 불가능한 이유 중 '타 연구자가 이해할 수 없다'인 것으로 확인되었는 바, 이 문제는 메타데이터 등과 같은 데이터

관리 체계를 정립하는 것이 대안이 될 수 있을 것으로 판단된다.

5. 결 론

본 연구의 목적은 학문분야별 연구데이터 관리 현황을 파악하기 위함이다. 관리 현황을 파악하기 위하여 선택한 학문분야는 사회과학인 문헌정보학, 자연과학인 통계학 및 생태학 그리고 예술분야인 한국음악학을 대상으로 하였다. 연구데이터 관리 현황을 파악하기 위하여 DAF 조사 도구를 이용하여 설문을 진행하였다. 연구데이터 관리 현황 분석은 데이터 생산, 데이터 공유 및 관리, 데이터 저장, 데이터 보존 그리고 데이터 재사용까지 5개 파트로 나누어 분석을 진행하였다. 분석 결과 데이터 저장을 제외하고 데이터 생산, 데이터 공유 및 관리, 데이터 보존 및 데이터 재사용의 경우 학문 분야별로 차이가 있는 것으로 파악되었다. 또한, 연구데이터 현황을 파악하며 나타났던 문제점에 대해서는 대용량 데이터를 위한 소속기관의 스토리지 확보, 연구데이터 관리 교육, 연구데이터 관리 규정 수립 그리고 메타데이터를 통한 데이터 관리 등이 대안으로 제시하였다. 특히, 연구데이터 관리 교육의 경우, 해외에서는 도서관을 중심으로 교육이 이루어지고 있음을 참고하여 우리도 도서관을 중심으로 한 연구데이터 관리 교육이 이루어져야 할 것이다. 본 연구의 후속으로 연구데이터 관리 교육을 주제로 진행된다면 현장의 연구자들에게 도움이 될 것으로 판단된다.

참 고 문 헌

- 국가연구개발정보처리기준. 과학기술정보통신부고시. 제2020-102호.
- 김성욱, 김선태 (2020). 응집물질물리분야 연구데이터 관리 방안 연구. *정보관리학회지*, 37(3), 77-106.
<http://dx.doi.org/10.3743/KOSIM.2020.37.3.077>
- 노태호, 홍선기, 강대석, 권오석 공역 (2000). *생태학: 인간과 자연*. 서울: 아카데미서적.
- 박미영, 안인자, 남승주 (2018). 과학기술분야 출연연구기관 연구데이터 관리 및 공유 사례 분석 연구. *한국비블리아학회지*, 29(4), 319-344. <http://dx.doi.org/10.14699/kbiblia.2018.29.4.319>
- 박윤미, 김지현 (2021). 기계공학분야 연구자들의 연구데이터 생산과 관리에 관한 연구. *한국기록관리학회지*, 21(4), 137-162. <https://doi.org/10.14404/JKSARM.2021.21.4.137>
- 이성숙 (2010). 한국전통음악의 서지적 관계 특성에 따른 FRBR 모형 적용방안. 대전: 한국연구재단.
- 이혜원 (2007). 음악 자원을 대상으로 한 이벤트 중심 ABC 온톨로지 확장 모형에 관한 연구. 박사학위논문, 연세대학교 문헌정보학과.
- Barabucci, G., Eschweiler, M., & Speer, A. (2018) TI-One: active research data management in a modern philosophy department. 2018 IEEE 14th International Conference on e-Science (e-Science), 2018, 314-315. <https://doi: 10.1109/eScience.2018.00070>.
- Collaborative Research Group (CRG) USA (2016). Top 10 Characteristics and Features of the Social Sciences. Available:
<https://crgsoft.com/10-characteristics-of-the-social-sciences/>
- DBBM [n.d.]. Characteristics of Biological Data. Available:
http://www.dbbm.fiocruz.br/class/Lecture/d17/db_overview/characteristics_of_biological_data.htm
- Donaldson, D. R. & Koepke, J. W. (2022). A focus groups study on data sharing and research data management. *Sci Data*, 9. <https://doi.org/10.1038/s41597-022-01428-w>
- JISC, University of Glasgow, & DCC (2009). Data Asset Framework Implementation Guide. Available: https://www.data-audit.eu/docs/DAF_Implementation_Guide.pdf
- Krahe M. A., Toohey J., Wolski M., Scuffham P. A., & Reilly, S. (2020). Research data management in practice: results from a cross-sectional survey of health and medical researchers from an academic institution in Australia. 19th Australasian Research Management Society Conference, 108-116. <https://doi.org/10.1177/1833358319831318>
- Milewska, A., Wiśniewska, N., Cimoszko, P., & Rusakow, J. (2022). A survey of medical researchers indicates poor awareness of research data management processes and a role

- for data librarians. *Health Info Libr J*, 39(2), 132-141. <https://doi.org/10.1111/hir.12403>
- Raszewski, R., Goben, A., Bergren, M., Jones, K., Ryan, C., Steffen, A., & Vonderheid, S. (2020). A survey of current practices in data management education in nursing doctoral programs. *Journal of Professional Nursing*, 37. <https://doi.org/10.1016/j.profnurs.2020.06.003>.
- UCLA Library [n.d.]. Data Management for the Social Sciences. Available: <https://guides.library.ucla.edu/data-management-social-sciences>
- University of Washington (2022). Learning Research Data Management with the UW Libraries. Available: <https://sites.uw.edu/libstrat/2022/03/23/learning-research-data-management-with-the-uw-libraries>
- Wiley, C. & Burnette, M. (2019). Assessing data management support needs of bioengineering and biomedical research faculty. *Journal of eScience Librarianship*, 8. e1132. <https://doi.org/10.7191/jeslib.2019.1132>
- Wiley, C. & Mischo, W. (2016). Data Management Practices and Perspectives of Atmospheric Scientists and Engineering Faculty. *Issues in Science and Technology Librarianship*. <https://doi.org/10.29173/istl1688>

• 국문 참고자료의 영어 표기

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

- Kim, Sungwook & Kim, Suntae (2020). A study on the research data management methods for the condensed matter physics. *Journal of the Korean Society for Information Management*, 37(3), 77-106. <https://doi.org/10.3743/KOSIM.2020.37.3.077>
- Lee, Hyewon (2007). A Study on the Event-Driven ABC Ontology Extension Model for Music Resources. Doctoral dissertation, Yonsei University.
- Lee, Seong-sook (2010). A Method for Applying the FRBR Model According to the Bibliographical Characteristics of Korean Traditional Music. Daejeon: National Research Foundation of Korea.
- National R&D information processing standard. Ministry of Science and ICT Notice. No. 2020-102.
- Park, Miyoung, Ahn, Inja, & Nam, Seungjoo (2018). A study on the analysis of research data management and sharing of science & technology government-funded research institutes. *Journal of the Korean Biblia Society for Library and Information Science*, 29(4), 319-344. <https://doi.org/10.14699/KBIBLIA.2018.29.4.319>

- Park, Yunmi & Kim, Jihyun (2021). A study on research data creation and management by researchers in mechanical engineering. *Journal of Korean Society of Archives and Records Management*, 21(4), 137-162. <https://doi.org/10.14404/JKSARM.2021.21.4.137>
- Roh, Tae-ho, Hong, Seon-gi, Kang, Dae-seok, & Kwon, Oh-seok [Co-translated] (2000). *Ecology: Man and Nature*. Seoul: Academy Books.