2022-06

2021년 기초원천연구기획사업

과학기술 연구데이터 구축·활용 강화 방안 연구

(Measures to Enhance the Development and Utilization of Research Data in Science and Technology Fields)

연 구 기 관 : 과학기술정책연구원
연구책임자 : 신은정

2022. 6. 16

과학기술정보통신부

안 내 문

본 연구보고서에 기재된 내용들은 연구책임자의 개인적 견해이며 과학기술정보통신부의 공식견해가 아님을 알려드립니다.

과학기술정보통신부 장관 이 종 호

제 출 문

과학기술정보통신부 장관 귀하

본 보고서를 “과학기술 연구데이터 구축·활용 강화 방안 연구”과제의 보고서로 제출합니다.

2022. 6. 16.

연구기관명 : 과학기술정책연구원

연구책임자 : 신은정 연구위원

연구원 : 정일영 연구위원
이다은 선임연구원

자문위원 :
이종숙 한국과학기술정보연구원 HPC 융합플랫폼연구단 단장
한종규 한국지질자원연구원 지질자원데이터센터 센터장
이천무 한국생명공학연구원 KOBIC 바이오데이터스테이션 부장
권재민 한국핵융합에너지연구원 통합시뮬레이션연구부 부장
고정환 경희대 물리학과 CERN 연구팀 교수
이세민 울산과학기술원 바이오메디컬공학과 교수
김성준 남서울대학교 빅데이터융합과 교수
김경철 웰케어클리닉 대표원장
안재평 한국과학기술연구원 책임연구원

요약문

양식A201

연구과제명	국 문 : 과학기술 연구데이터 구축·활용 강화 방안 연구 영 문 : Measures to Enhance the Development and Utilization of Research Data in Science and Technology Fields			
연구책임자	소 속	과학기술정책연구원	성 명	신은정
연구기간	2021.12.17～2022.06.16 (6개월)		연구비	40,000(천원)

○ (연구배경) 연구데이터는 새로운 과학기술혁신의 원천으로서 전 세계적으로 연구데이터의 구축·활용 노력이 가속화되고 있으며, 우리나라도 최근 들어 연구데이터의 구축 및 활용을 위한 정책과 사업을 확대하고 있음

– 하지만, 연구데이터의 구축·관리에 투입하는 개별적 노력의 효과를 제고하고 국가차원의 구축·활용 성과를 창출하기 위한 종합 대책의 필요성이 여전히 존재하며 이를 위한 기초자료와 정책협의가 필요한 시기임

○ (연구목표) 본 연구는 과학기술분야 연구데이터사업이 증가하는 현 시점에서 연구데이터사업 고도화방안, 개별 데이터센터간의 연계·운영방안 등을 중점 검토하여 국가차원의 연구데이터 구축·활용 효과를 높이기 위한 정책과제를 제시하는데 목적이 있음

○ (연구내용) 연구의 효과적 수행을 위하여 국내외 연구데이터 정책 동향을 분석하는 한편, 국내 연구데이터 구축·활용 현황분석과 관련 이해관계자와의 정례적 협의를 통해 정책이슈를 발굴·심화함

– 구체적으로 미국, 유럽연합 등지의 연구데이터 정책 동향과 함께 국내에서 진행된 연구데이터 정책을 검토함. 또한 연구데이터의 구축·활용 현황 분석을 위하여 데이터 프로세스 관점에서 연구 전 과정에서 이루어지는 분야별 데이터 구축·활용 과정과 이슈를 점검함. 마지막으로 국가차원에서 정책적으로 고려해야 할 정책 이슈와 과제, 정책의 방향성을 검토 ·제시함

○ (기대효과) 본 연구는 개별 사업 기획이나 분석을 넘어서는 연구데이터 정책 진단이자 분야별 데이터 프로세스를 관통하는 정책 이슈와 과제를 발굴한 연구로서 국가차원의 연구데이터 구축·활용 효용을 제고하는데 기여할 것으로 기대됨

〈 SUMMARY 〉

Purpose & Contents	o Research data is a source of new scientific and technological innovation. Efforts to develop and utilize research data have been accelerated around the world. However, there is still a need for a comprehensive measure to enhance the effectiveness of research data management at the national level that is beyond individual efforts dedicated in research data management. o This study aims to devise policy recommendations to enhance the effectiveness of research data management at the national level by identifying policy gaps to be fulfilled and making a linkage among individual data center programs.		
Results	o This study analyzes domestic and international trends in research data policies and services, and surveys the current status and policy issues related to the development and utilization of research data in science and technology fields, based on a series of interviews and taskforce team meetings with relevant domestic stakeholders. o In detail, this study reviews research data policies and practices in Korea along with the ones in other countries and regions, such as the United States and the European Union. In addition, it examines the data life cycle embedded in research process which is mostly field-specific, based on case studies. Given the study results and consultation with domestic stakeholders, this study presents policy issues on the development and utilization of research data, which is followed by policy recommendations to be considered at the national level in the near future		
Expected Contribution	o This study uncovers policy issues which are not limited to individual projects and programs and it also provides policy recommendations on research data that are applicable to the national level, which is expected to contribute to enhancing the overall effectiveness of national research data policies and programs.		
Keywords	Research Data	Scientific Data	Research Data Platform

| 목 차 |

보고서 요약문 ····· i

제1장. 연구개발과제의 개요 ····· 1

제1절 연구개발 필요성 및 목적 ····· 1

제2절 연구의 범위와 주요내용 ····· 4

제3절 연구의 방법 및 추진체계 ····· 6

제2장. 국내외 동향 분석 ····· 9

제1절 글로벌 동향 ····· 9

제2절 국내 동향 ····· 31

제3장. 연구데이터 구축·활용 현황 및 정책이슈 ····· 43

제1절 연구데이터 구축·활용 현황 ····· 43

제2절 연구데이터 구축·활용 관련 정책이슈 ····· 71

제4장. 연구데이터 구축·활용 촉진을 위한 정책(안) ····· 86

참고 문헌 ····· 96

〈 CONTENTS 〉

Summary ····· i

Chapter 1. Introduction ····· 1

1. Importance and Purpose of the Study ····· 1

2. Scope and Contents of the Study ····· 4

3. Research Methods and processes ····· 6

Chapter 2. Domestic and International Trends Analysis ····· 9

1. International Trends ····· 9

2. Domestic Trends ····· 31

Chapter 3. Current Status and Policy Issues regarding the Development and Utilization of Research Data ····· 43

1. Current Status on the Development and Utilization of Research Data ····· 43

2. Policy Issues on in the Development and Utilization of Research Data ····· 71

Chapter 4. Policy Recommendations to Enhance the Development and Utilization of Research Data ····· 86

References ····· 96

| 표 목 차 |

<표 1-1> 연구데이터 정책협의 회의/행사 일정 ····· 7
<표 1-2> 연구추진일정 ····· 8
<표 2-1> Horizon Europe 프로그램의 연구데이터 관련 지침 ····· 19
<표 2-2> 유럽 주요국의 연구데이터관리(RDM) 가이드 개발 및 이행 현황 ····· 20
<표 2-3> EC의 신뢰성있는 리포지터리(trusted repository)의 기준 ····· 22
<표 2-4> 미국 NSTC의 바람직한 리포지터리 특성 ····· 24
<표 2-5> 일본 연구데이터 관련 주요 지표 ····· 29
<표 2-6> 국가 차원의 데이터 정책 및 이니셔티브 ····· 31
<표 2-7> 국가연구개발정보처리기준의 연구데이터 관련 조문 ····· 35
<표 2-8> 연구과제 수행절차에 따른 DMP 수행내용 ····· 36
<표 3-1> 연구에 활용되는 데이터의 스펙트럼 ····· 44
<표 3-2> 연구데이터와 공공·산업데이터의 비교 ····· 46
<표 3-3> 데이터 사용목적에 따른 CERN의 연구데이터 공개 수준 구분 ····· 50
<표 3-4> 주요국 연구데이터 정책 동향 개괄 ····· 71
<표 3-5> 데이터-SW 접근성에 따른 실질적 접근성의 차별화 ····· 74
<표 3-6> 데이터 특성에 따른 데이터 관리 특성 차별화 ····· 78
<표 3-7> 국내 연구자가 활용 중인 해외 리포지터리 데이터 ····· 82
<표 4-1> 데이터중심연구 활성화를 위한 선도연구사업(안) ····· 88
<표 4-2> 데이터중심연구 활성화를 위한 연구데이터 커먼즈사업(안) ····· 89
<표 4-3> 정책과제 제안 종합(안) ····· 93
<표 4-4> 정책과제(안)의 추진 ····· 94

| 그 림 목 차 |

[그림 1-1] 디지털 시대 연구데이터연맹(RDA)의 성장 ····· 1
[그림 1-2] 연구데이터의 범위 ····· 4
[그림 1-3] 연구추진체계 ····· 7
[그림 2-1] 디지털 기술의 발전과 데이터·인프라의 성장 추세 ····· 9
[그림 2-2] 데이터과학 및 관련 분야의 성장 추세 ····· 10
[그림 2-3] re3data에 등록된 연구데이터 리포지터리의 증가 추세 ····· 11
[그림 2-4] 글로벌 연구데이터 리포지터리 현황(re3data 자료) ····· 12
[그림 2-5] 글로벌 연구데이터 리포지터리 현황(Fairsharing 자료) ····· 13
[그림 2-6] 글로벌 OA 정책 증가 추세(ROARMAP 자료) ····· 14
[그림 2-7] 분야별 데이터정책 보유 저널의 비중: Taylor & Francis 사례 ····· 16
[그림 2-8] 유럽 연구도서관의 연구데이터서비스 현황(LIBER 자료) ····· 17
[그림 2-9] 유럽데이터스페이스 구축 계획 ····· 18
[그림 2-10] 미국 연방부기관의 공공액세스/데이터정책 도입 ····· 23
[그림 2-11] 미국 NITRD 프로그램의 연도별·기관별 지원 예산 ····· 25
[그림 2-12] 미국 연방데이터 전략: 2021 실행계획 ····· 26
[그림 2-13] 일본의 포괄적 데이터 전략 ····· 27
[그림 2-14] 일본 연구데이터정책 이행체계 ····· 28
[그림 2-15] 중국 국가자연과학기금위원회(NSFC)의 데이터 관련 지원 ····· 30
[그림 2-16] 국가 차원의 데이터 플랫폼 관련 주요 계획의 내용 ····· 32
[그림 2-17] 연구데이터 공유·활용 전략의 주요 내용 ····· 34
[그림 2-18] 데이터관리계획 적용 사업의 확대 ····· 36
[그림 2-19] 소재 연구데이터 플랫폼 개념도 ····· 38
[그림 2-20] 바이오 연구 데이터 확보관리 체계도 ····· 40
[그림 2-21] 국가 바이오 데이터 스테이션 운영 프로세스 ····· 41
[그림 3-1] 연구데이터의 범위 ····· 43
[그림 3-2] 연구-공공-산업데이터 관계도 ····· 45
[그림 3-3] 연구데이터 프로세스 ····· 47
[그림 3-4] 입자물리학 CMS 연구데이터 프로세스 사례 ····· 50
[그림 3-5] CERN Open data portal과 데이터셋/분석 도구 탑재 ····· 52
[그림 3-6] CERN ALICE 인터페이스 및 software ····· 52
[그림 3-7] 지질자원데이터 리포지터리 (GDR) ····· 58
[그림 3-8] KIST 소재데이터플랫폼 구축을 통한 실험데이터 수집 ····· 61
[그림 3-9] 연구데이터와 관련 매체 ····· 73
[그림 3-10] 연구데이터플랫폼의 다층적 구조와 기능 ····· 79
[그림 4-1] 정책이슈에 대응하는 정책개선과제(안) ····· 86

제1장. 연구개발과제의 개요

제1절 연구개발 필요성 및 목적

1. 연구의 필요성

(1) 연구의 배경 및 필요성

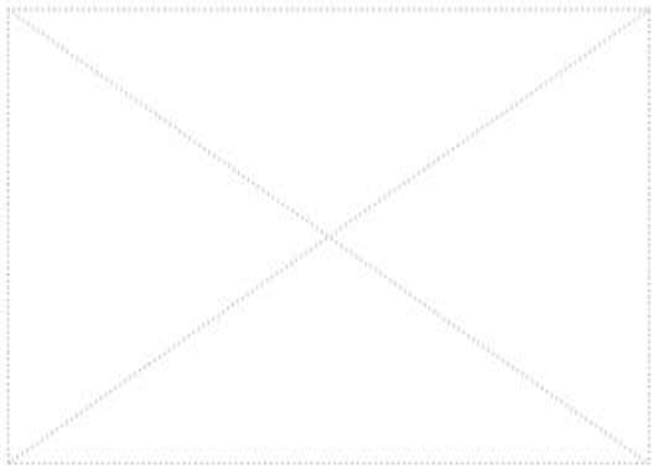
○ 컴퓨팅기술의 급속한 발전으로 분절적 정보에 불과했던 데이터가 지식 창출에 직접 기여하는 시대로 전환(신은정 외, 2021)

- 연구데이터는 연구결과의 검증에 필수적인 사실 기록뿐만 아니라 연구과정에서 생산·활용되는 다양한 형태의 텍스트, 음성, 이미지 자료를 통칭할 수 있음

○ 국내외적으로 연구데이터의 가치와 수요가 증가함에 따라 연구데이터의 생산·관리 기반을 확대하고 공유·활용을 촉진하는 정책과 서비스 증가

- 전 세계적으로 연구데이터의 효율적 관리 및 활용을 지원하는 리포지터리와 서비스기관 증가
- 연구데이터 생산, 관리, 공유, 활용을 촉진하기 위한 글로벌 연합체로서 2013년 발족한 연구데이터연맹(research data alliance, RDA)도 지속적으로 성장

[그림 1-1] 디지털 시대 연구데이터연맹(RDA)의 성장



자료: Research Data Alliance 홈페이지, rd-alliance.org (접속일: 2021.5.6.), 신은정(2021), p. 2 재인용

- 개별 국가 차원에서 공공자금 지원을 받은 연구데이터의 관리·활용도를 높이거나 연구·혁신수요가 있는 데이터의 구축 및 활용을 장려하는 정책 확대
- OECD는 2021년 초 「공공자금으로 산출된 연구데이터의 접근성에 관한 권고」를 재개정·공표

○ 우리나라의 경우, 2018년 「혁신성장 촉진을 위한 연구데이터 공유·활용 전략」(2018.1.19.), 국가과학기술심의회 운영위원회)을 채택·이행 중
- 연구데이터 관리체계를 정비하고 바이오, 소재, 인공지능, 대형연구장비 등의 데이터집약형 연구분야부터 연구데이터센터 구축·운영 계획을 현실화*

* 소재분야 연구데이터플랫폼사업(2020-2028), 바이오데이터스테이션사업(2021-2025)

- 연구데이터의 체계적 구축·관리·활용을 촉진하기 위한 데이터관리계획(Data Management Plan, DMP)을 부분적으로 도입·적용 중

* 2019년 3월「국가연구개발사업의 관리 등에 관한 규정」에 연구데이터의 관리 및 활용을 촉진하기 위한 데이터관리계획의 이행근거 포함되었으나, 2020년「국가연구개발혁신법」의 제정으로 인하여 관련 조항은「국가연구개발정보처리기준」에 포함

○ 분야별 연구데이터사업이 본격화되고 있으나 국가 차원의 연구데이터 구축·연계·활용 전략이 충분히 구체화되지 못하여 사업실행단계 공백이 존재함
- `연구데이터에 대한 이해도가 분야별·부문별로 상이하여 국가 차원에서 보편적으로 합의 가능한 연구데이터의 생산·관리·활용 기준이 마련·정착되지 않음
- `연구데이터의 수집뿐만 아니라 중장기적 운영과 관리, 활용이 이루어져야 하지만, 분야별 데이터 수집·관리 기반을 마련하는 개별 사업 단위에서 국가 차원의 중장기적 운영, 관리, 연계, 활용 전략을 모색하는데 한계가 있음

(2) 선행연구 검토 및 본 연구의 차별성

○ 연구데이터의 가치가 부상하면서 관리·활용의 필요성이 증가한다는 주장을 펼친 선행연구는 이미 다수 존재
- Hey, et al. (2009)는 일찍이 제4차 패러다임을 주창하며 데이터중심과학의 미래를 예견했고, 이후에도 Clauset, et al. (2017) 등의 후속 연구에서 데이터중심성이 강화되는 최근 연구의 경향성과 연구데이터의 가치가 재확인됨

○ 연구데이터의 구축 및 활용을 장려하는 기술적 방편으로서 FAIR 원칙이 제안된 바 있으며, 이외에도 데이터 관리/연계/표준화를 위한 기술적 제안과 데이터센터 구축·운영을 위한 설계 노력 증가

- Wilkinson et al. (2016)은 연구데이터의 공개·활용 촉진을 위하여 FAIR원칙, 즉 연구데이터를 검색가능하고(findable), 접근가능하고(accessible), 상호호환가능하고(interoperable), 재사용가능하도록(reusable) 제공하자는 지침 제시
- 2013년 연구데이터연맹(RDA)의 창설 이래 워킹그룹별 데이터 관리/연계/표준화에 대한 기술적 대안 논의 급성장
- 이외에도 연구데이터사업의 기획 혹은 데이터 리포지터리의 설계를 위해 수행한 다수의 국내 선행연구* 존재

* '데이터 기반 과학기술 R&D를 위한 클라우드 서비스 기획연구(2017)', '국가연구데이터플랫폼 구축 및 운영(2019) 등

○ 하지만 연구데이터에 대한 포괄적 이해 아래 국가 차원의 연구데이터 구축·활용 방안을 제시한 정책연구는 상대적으로 미흡
- 신은정 외(2019a)에서 국가연구개발사업 수행주체를 대상으로 연구데이터의 관리 실태를 조사하고 데이터관리계획의 이행 근거를 마련하기 위한 정책연구를 수행한 바 있으며, KISTI(2019)는 과기분야 출연연 대상 연구데이터의 활용실태조사와 국가과학기술연구회 차원의 데이터관리계획 이행방안을 제시한 바 있음
- 황석원 외(2020)은 데이터 기반 연구개발 관리체계의 혁신을 위한 정책연구를 수행
- 다만, 국가 차원의 연구데이터 구축·활용 전반의 정책 방향성을 검토·점검하고 데이터중심연구 활성화를 위한 정책과제를 제시한 연구는 부족

2. 연구의 목적

○ 본 연구는 과학기술분야 연구데이터의 가치와 수요가 증가하는 시점에서 국내적으로 구축되고 있는 연구데이터의 중장기적 관리·활용 효과를 제고하기 위한 정책방향과 실천과제를 도출하는데 목적이 있음
- 연구데이터의 가치와 범위, 특수성을 살펴 국가 차원의 연구데이터 구축·활용정책의 방향성 제시
- 연구데이터의 중장기적 구축·관리, 공유·활용 수준을 높이기 위해 분야별·사업별 지침을 관통하는 국가 차원의 연구데이터정책 실천과제 모색·제안

제2절 연구의 범위와 주요내용

1. 연구의 범위

○ 본 연구는 **과학기술분야 국가연구개발사업에서 산출되는 연구데이터**를 대상으로 함
- 국가연구개발사업이란 중앙행정기관이 법령에 근거하여 연구개발을 위하여 예산 또는 기금으로 지원하는 사업을 일컬음(국가연구개발혁신법 제2조)
- **과학기술분야 국가연구개발사업**이란 과학기술정보통신부 소관 법령에 근거하여 연구개발과제를 특정하여 그 연구개발비의 전부 또는 일부를 출연하거나 공공기금으로 지원하는 과학기술분야의 연구개발사업을 지칭함(과학기술정보통신부 소관 과학기술분야 연구개발사업 처리규정 제2조)

○ 본 연구에서 다루는 연구데이터는 現 「국가연구개발정보처리기준」 연구데이터 정의보다 포괄적으로, **연구개발과제 수행과정에서 수집, 생산, 가공, 처리, 분석, 해석되는 데이터**를 일컬음
- 연구결과의 검증에 필수적인 데이터뿐만 아니라 결과 검증에 직접적으로 활용되지 않지만 연구개발과정에서 생성되는 중간산출 데이터를 포함

* 국가연구개발정보처리기준」 제2조에서는 연구데이터를 “연구개발과제 수행과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실자료로서 연구결과의 검증에 필수적인 데이터“로 정의하고 있음

[그림 1-2] 연구데이터의 범위

자료: 신은정 외(2019a), p. 36 발췌 및 수정

2. 연구의 주요내용

○ 연구데이터 구축·활용기반에 관한 개념화 및 연구방향 설정
- 연구데이터의 구축 및 활용에 관한 선행연구 검토
- 연구의 필요성 구체화, 연구범위 및 방향 설정

○ 연구데이터의 구축 및 활용을 촉진하는 국내외 동향 분석
- 연구데이터의 구축, 관리, 공유, 활용에 관한 글로벌 동향 분석
- 국내 연구데이터의 구축, 관리, 활용에 관한 국내 정책 동향

○ 국내 연구데이터 구축·활용 현황 및 정책이슈
- 연구데이터 프로세스 개괄
  (연구데이터의 특성 분석, 연구데이터와 유사 개념과의 관계 설정 포함)
- 바이오·소재 등 분야별 연구데이터 프로세스 분석
- 연구데이터 구축 및 활용 관련 주요 정책 이슈

○ 연구데이터의 구축·활용 촉진을 위한 정책과제 제시
- 국가차원의 중장기 연구데이터 구축·활용 촉진을 위한 추진 방향
- 연구데이터 구축·활용 효과 제고를 위한 실천적 정책과제

제3절 연구의 방법 및 추진체계

1. 연구의 방법

○ (동향분석) 웹정보, 정책·사업자료 및 선행연구 등을 검토하며 국내외 동향 분석

○ (실태분석) 한국연구재단 주관 연구데이터 관련 사업의 현황 개괄, 바이오, 소재 등 분야별 주요 연구데이터사업에 대한 추가 조사 및 관계자 면담, 연구 분야별 데이터 프로세스에 대한 연구자 인터뷰 등을 통해 연구데이터의 생산·관리·연계·활용 이슈 분석

○ (정책협의) 정부부처 담당자와 주요 연구데이터사업 담당자, 분야별 데이터중심연구 선도 연구자 등 주요 이해관계자로 구성된 정책자문 실무팀(T/F)를 구성하여 의견 수렴 및 정책 협의

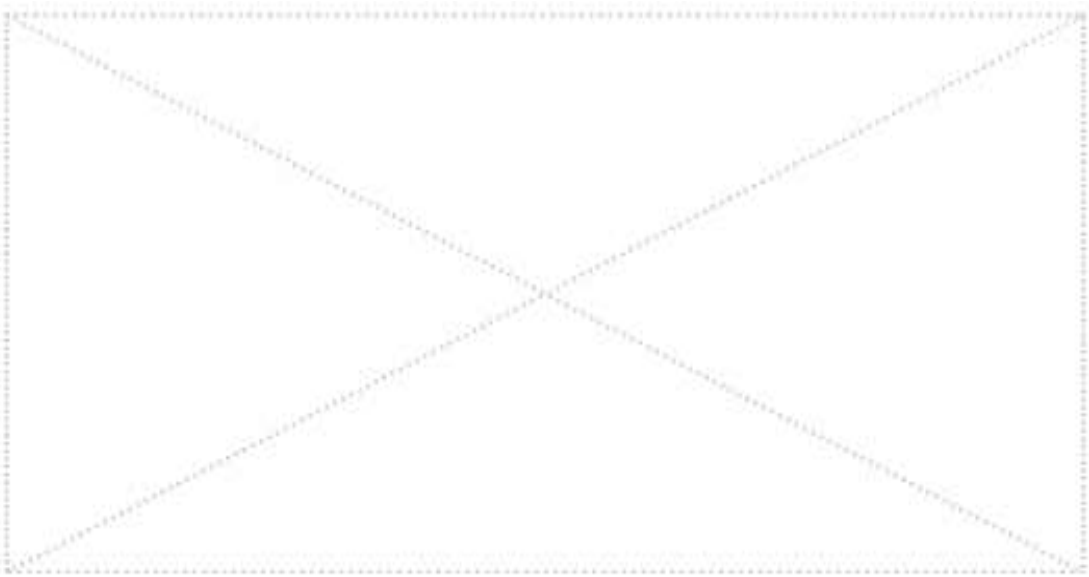
2. 연구의 추진체계

○ (내부 연구팀) 연구전체일정과 내용을 공유하면서 동시에 두 개의 실무 모듈을 운영하여 연구추진 효율성 제고

○ (외부 T/F) 연구진 이외 국내 주요 이해관계자 중심 정책자문 T/F를 구성하여 도출된 주요 연구결과의 검토·협의 진행

- 주무부처인 과학기술정보통신부와 주요 연구데이터센터 운영 관계자, 분야별 데이터중심 연구 활동이 활발한 연구자 후보 추천을 받아 정책자문 T/F 구성, 정례회의 진행

[그림 1-3] 연구추진체계



○ 일차 조사를 완료한 연구 중반부터 담당부처 협의 및 T/F 회의 진행

<표 1-1> 연구데이터 정책협의 회의/행사 일정

구 분	일시	참석자
1차 정책 부처 협의	22.2.22	담당 과장, 사무관, 연구진
2차 정책 부처 협의	22.3.13	담당 국장, 과장, 사무관, TF 위원 일부, 연구진
1차 정책 자문 T/F	22.3.22	담당 국장, 과장, 사무관, TF 위원 전체, 연구진
2차 정책 자문 T/F	22.4.30/22.5.3.	TF 위원 전체(두 그룹으로 회의 진행), 연구진
3차 정책 자문 T/F	22.5.25	TF 위원 전체, 연구진
기술혁신학회 특별세션	22.5.27	TF 위원 일부, 연구진, 학회 특별세션 공동발표자 및 참여자
3차 정책 부처 협의	22.5.31.	담당 과장, 사무관, 연구진

3. 연구의 추진일정

○ 연구 전반부 국내외 동향분석 및 실태조사를 실시하고, 연구 후반부 정책이슈 발굴 및 대안 도출

<표 1-2> 연구추진일정

구 분	연구기간					
	1개월 (‘21.12)	2개월 (‘22.1)	3개월 (‘22.2)	4개월 (‘22.3)	5개월 (‘22.4)	6개월 (‘22.5)
주요 개념 및 연구방향 설정						
국내외 동향분석						
실태분석 및 정책이슈 발굴						
정책과제 협의 및 도출						

제2장. 국내외 동향 분석

제1절 글로벌 동향

1. 디지털 전환기 연구데이터의 성장

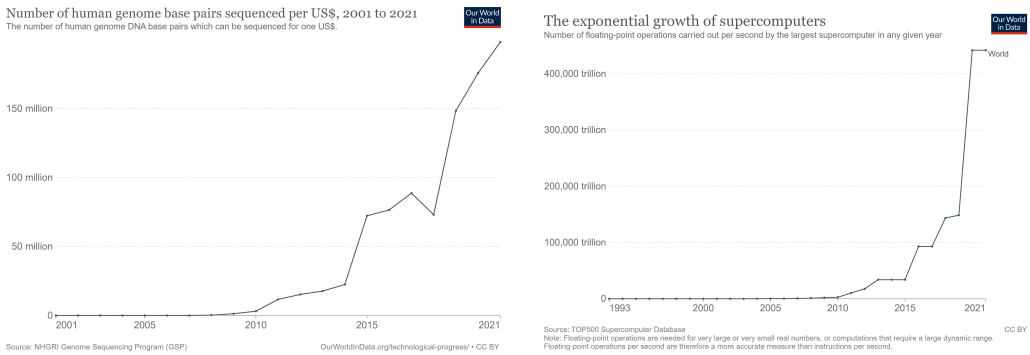
(1) 디지털 기술의 발전과 데이터중심연구의 성장

○ 디지털 기술 및 도구의 급속한 성장과 함께 데이터를 생산, 수집하고 가공, 분석, 활용하는 범위와 규모가 확대

– 데이터 분석 및 마이닝 방법, 빅데이터 및 기계학습 기술의 발전과 함께 연구개발 과정에서도 데이터의 생산, 수집, 활용의 규모와 속도 증가

[그림 2-1] 디지털 기술의 발전과 데이터·인프라의 성장 추세

< 1달러당 분석된 인간게놈데이터의 규모 > < 수퍼컴퓨터의 기하급수적 성장 >

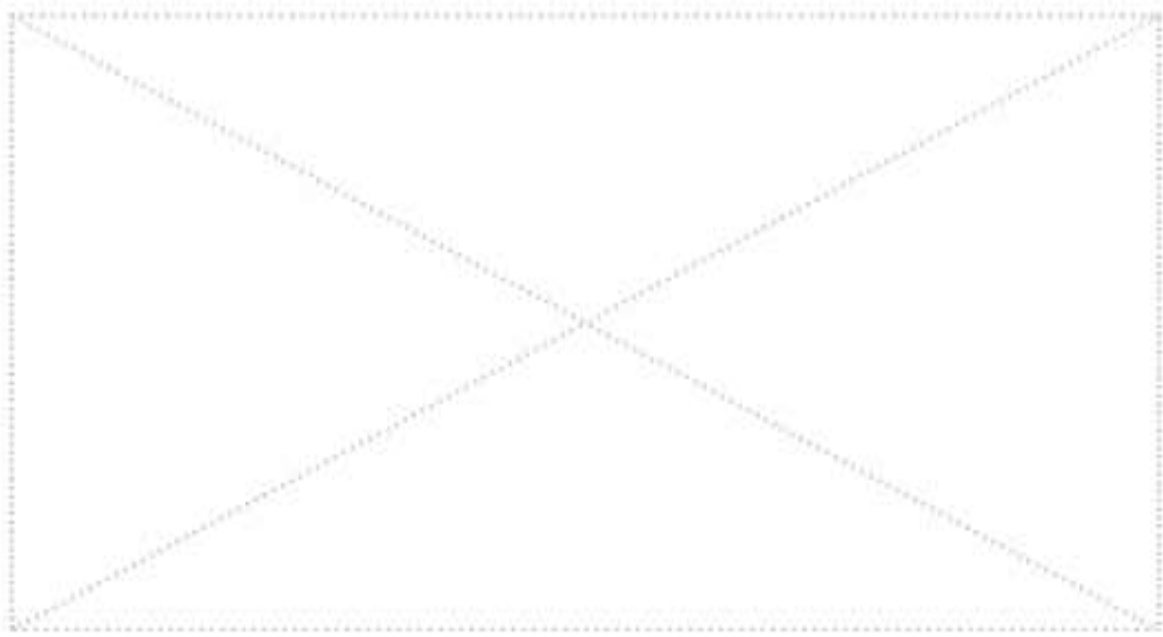


자료: Our World in Data 웹 자료 발췌, https://ourworldindata.org/technological-progress (접속일: 2022.3.28.)

○ 디지털 기술이 연구에 적용되면서 연구개발과정에서 생산되는 데이터의 규모와 종류가 증가하고 데이터 분석을 위한 연구 활성화

– 4차 산업혁명에 비견되는 디지털 전환 추세에 따라 증가하는 데이터의 관리·활용도를 고도화는 데이터과학(data science) 및 데이터공학(data engineering)도 함께 성장

[그림 2-2] 데이터과학 및 관련 분야의 성장 추세



자료: Sarker(2021) p. 377 발췌

○ 디지털시대 연구데이터는 기존 실증·검증의 영역을 확장할 뿐만 아니라 새로운 연구의 방향을 개척

- 기존 연구데이터가 이론적 가설이나 설계를 지엽적으로 검증하는 수단이었다면, 디지털 시대 연구데이터의 확대는 기존 실증·검증의 범위와 속도를 확대시킴
- 연구데이터를 통해 기존 이론을 검증하는 수준을 넘어서서 새로운 발견이나 이론의 발전을 이끄는 데이터중심의 4세대 연구모델 부상(Hey et al. 2009; Kitchin 2014)
- 디지털 데이터중심연구의 성장세로 인하여 디지털 지질학(digital geography), 디지털 인문학(digital humanity) 등 디지털과학(digital science)의 영역도 확장

(2) 연구데이터 리포지터리 및 서비스의 성장

○ 연구데이터에 대한 수요와 생산이 모두 증가하면서 전 세계적으로 연구데이터 서비스를 제공하는 조직과 기능 증가

- 차세대 연구·혁신을 위한 데이터의 중요성에 주목하여 연구데이터연맹(RDA)이 2013년 결성되었으며, 당시 유럽연합, 호주, 미국의 국립과학재단(NSF)과 국가표준원(NIST) 등 주요국 펀딩기관과 연구기관, 데이터서비스기관 참여[1)]
- 연구현장에서 데이터 저장, 관리, 제공을 지원하는 연구데이터 리포지터리 또한 증가

1) Research Data Alliance 홈페이지, https://www.rd-alliance.org/about-rda (접속일: 2022.3.28.)

- 전 세계 연구데이터 리포지터리 목록을 제공하는 리포지터리 레지스트리 서비스, re3data, 독일연구재단(German Research Foundation, DFG) 지원으로 2012년 시작
- re3data에 등록된 연구데이터 리포지터리의 수는 해를 거듭하면서 증가하여 2022년 2월 현재 2,792개에 이름

[그림 2-3] re3data에 등록된 연구데이터 리포지터리의 증가 추세

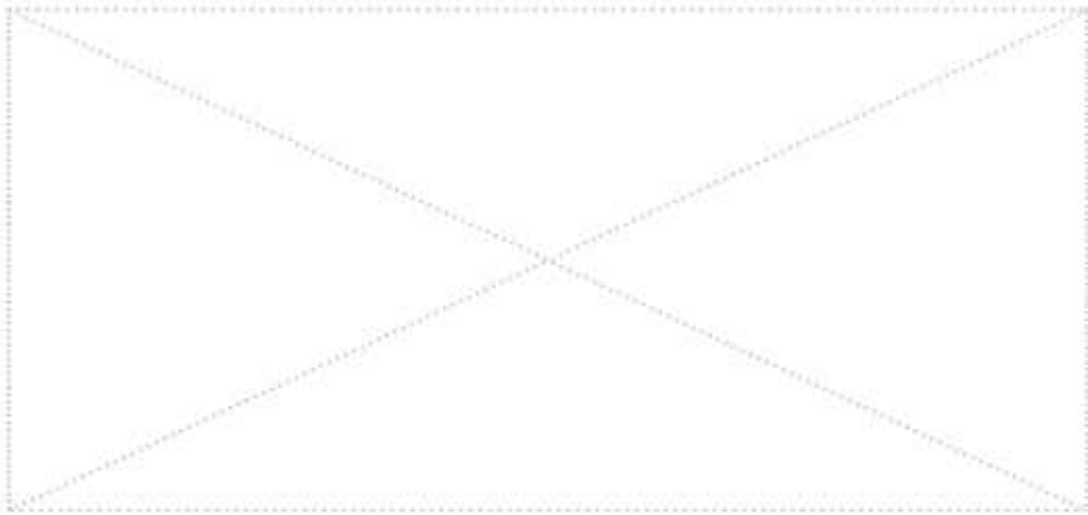
자료: Datacite 블로그(20222016.4.13. 작성)[2], 그림 발췌 후 re2data coref 블로그(2022.3.15. 작성)[3] 데이터로 갱신하여 도식화

○ 전 세계적인 연구데이터 리포지터리의 구축·활용 증가하고 있으며, 특히 리포지터리 운영을 위한 미주 및 유럽 지역 활동이 두드러짐
- re3data에 등록된 연구데이터 리포지터리 2,792개 중 490개가 국제기구 및 유럽 연합 기관과 연계 운영 중
- 전체 등록 리포지터리의 12.4%가 두 개 이상의 국가간 협력에 의해 운영되고 있으며, 미국과 영국, 미국과 캐나다, 미국과 독일 및 유럽국가간 협력이 활발한 편임[4]

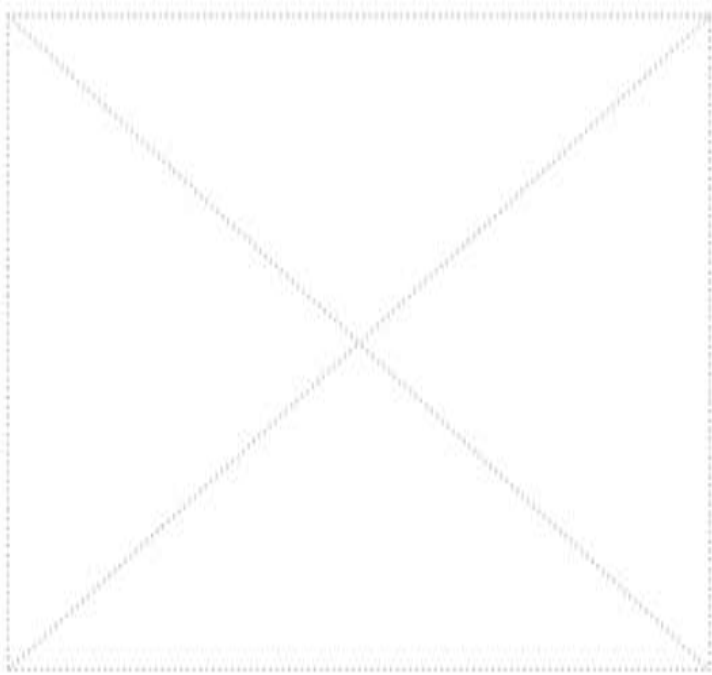
2) Datacite 블로그 (2016.4.13. 작성), https://blog.datacite.org/re3data-reaches-a-milestone-and-begins-offering-badges/#:~:text=re3data.org%20has%20reached%20a,disciplines%20from%20around%20the%20world. (접속일: 2022.3.31.)
3) re2data coref 블로그 (2022.3.15. 작성), https://coref.project.re3data.org/blog/mapping-the-global-repository-landscape (접속일: 2022.3.31.)
4) re2data coref 블로그 (2022.3.15. 작성), https://coref.project.re3data.org/blog/mapping-the-global-repository-landscape (접속일: 2022.3.31.)

[그림 2-4] 글로벌 연구데이터 리포지터리 현황(re3data 자료)

< re3data 등록 연구데이터 리포지터리의 국가 분포 >



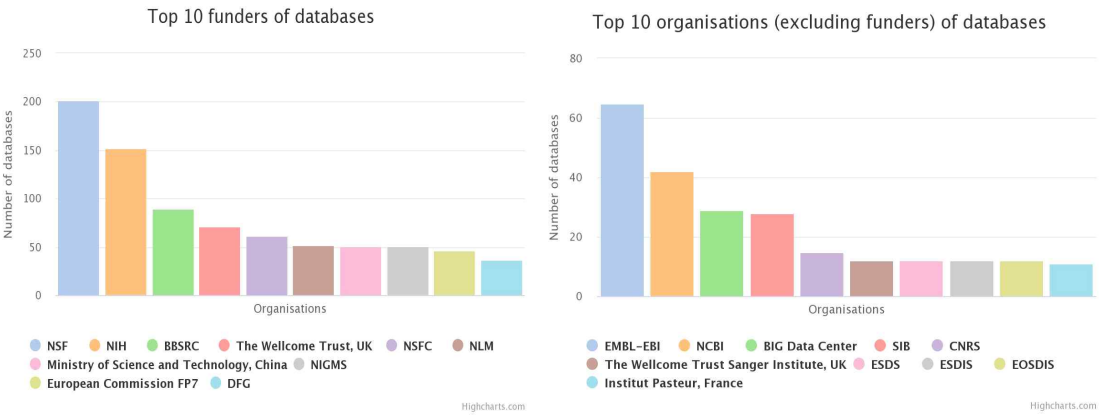
< re3data 등록 연구데이터 리포지터리의 국가간 협력 현황 >



자료: re3data coref 블로그(2022.3.15. 작성) 그림 발췌,
https://coref.project.re3data.org/blog/mapping-the-global-repository-landscape (접속일: 2022.3.31.)

– Fairsharing 통계에서도 미국, 영국, 유럽연합 국가 등지에서 연구데이터 리포지터리 활동이 활발하며 상대적으로 많은 연구데이터를 생산·제공하는 것으로 확인됨
– 미국, 캐나다 등의 미주 지역과 유럽 지역 이외에도 중국의 연구데이터 활동 부상

[그림 2-5] 글로벌 연구데이터 리포지터리 현황(Fairsharing 자료)



자료: Fairsharing 통계[5] 그림 발췌

5) Fairsharing 통계, https://fairsharing.org/summary-statistics (접속일: 2022.3.31.)

2. 글로벌 차원의 연구데이터정책 확산

(1) 연구데이터 정책의 증가

○ 연구데이터의 생산·공유·활용이 증가하면서 정책적으로 데이터의 체계적인 생산과 관리, 공유와 활용을 장려하는 추세 확대

– 연구기관이나 연구커뮤니티, 연구펀딩기관이나 정부에서 연구데이터의 체계적 생산과 관리, 활용을 장려하기 위한 연구데이터정책(research data policy)과 연구데이터관리(RDM, research data management) 지침을 제공하는 사례 증가

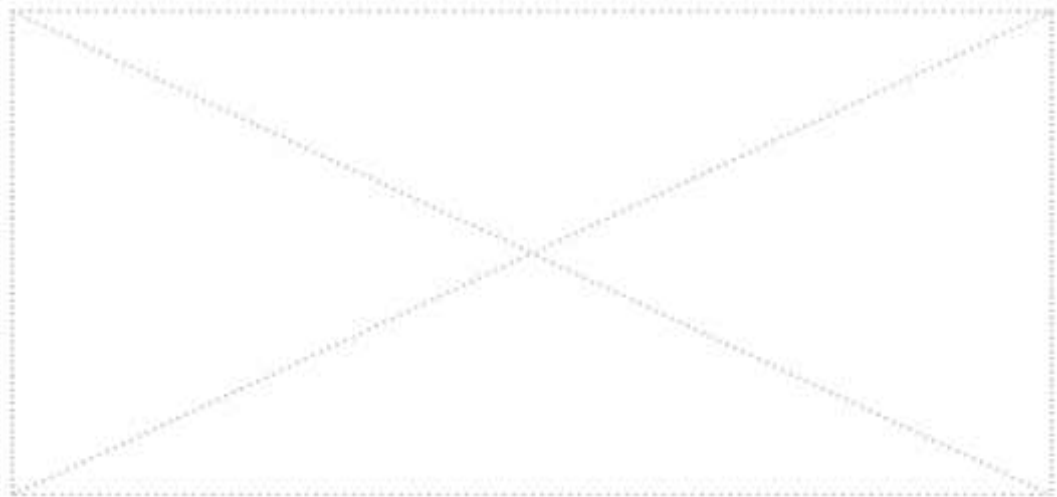
○ ROARMAP(Registry of Open Access Repository Mandates and Policies)에 따르면, 2021년 4월 기준 OA 정책을 도입한 기관은 1,100개가 넘음

– 연구과정에서 생산된 논문 포함 출판물에 관한 OA 정책이 주를 이루지만, 연구데이터의 관리 및 등록, 공개에 관한 사항을 명시한 OA 정책도 부상 중임

– ROARMAP 분석에 따르면, 등록된 OA 정책 중 유럽연합의 [Horizon 2020] OA 정책을 일부 혹은 전면 반영한 경우는 779건이며, 연구데이터의 OA를 제시한 정책도 35건 존재[6)]

– 한편 Sherpa Juliet 통계에서 집계된 전체 178건의 OA 정책 중에서 데이터 아카이빙에 관한 사항을 명시한 경우는 48%(30%는 의무적, 18%는 자발적 데이터아카이빙 정책)[7)]

[그림 2-6] 글로벌 OA 정책 증가 추세(ROARMAP 자료)



자료: ROARMAP 홈페이지(http://roarmap.eprints.org/, 접속일: 2020.4.6.)

6) ROARMAP 홈페이지, http://roarmap.eprints.org/dataviz.html, 접속일: 2022.4.6.
7) JISC 홈페이지, https://v2.sherpa.ac.uk/view/funder_visualisations/1.html, 접속일: 2022.4.6.

(2) 국제기구의 연구데이터 정책

○ OECD는 2006년 공공지원을 받은 연구데이터의 접근성을 제고하기 위한 권고*를 마련, 제시한 바 있으며, 그간의 환경적 변화를 고려하여 2021년 해당 권고를 개정·공표함

* Recommendation of the Council concerning Access to Research Data from Public Funding

- 2021년 개정안은 공공지원을 받은 연구데이터 및 관련 디지털 매체의 접근성을 제고하기 위하여 다음과 같은 7가지 실천과제를 제안함(신은정, 2021)
- **(거버넌스)** 책임성있는 신뢰 기반 데이터 거버넌스의 구축 및 운용
- **(기술표준)** 연구데이터 등이 FAIR*할 수 있는 기술표준과 실천 확대

* Findable, Accessible, Interoperable, and Reusable

- **(책임·권리)** 연구데이터 등의 접근 및 이용에 관한 책임과 소유권, 스튜워드쉽의 구체화
- **(인센티브·보상)** 기존 보상체계에 준하는 연구데이터 접근성 제고 보상체계 개발·적용
- **(인프라)** 지속가능한 연구데이터 등의 접근 및 이용을 위한 인프라·서비스의 지속
- **(인적자본)** 연구데이터 등에 관한 전문지식과 숙련 인력의 양성
- **(국제협력)** 연구데이터 등(민감데이터 포함)의 접근을 위한 국제공조 및 협력 촉진

○ 이외에도 여러 국제기구 및 회의를 통해 연구데이터 접근성 제고에 관한 규범 확산

- 세계은행(World Bank)은 2012년부터 자체적인 Open Data Portal을 구축·운영하고 있으며, UNESCO는 2021년 말 연구데이터를 포함한 연구 전 과정의 포용과 개방을 강조하는 오픈사이언스 권고*를 마련·공표함

* UNESCO Recommendation on Open Science

- G8 정상회의(2013년)를 시작으로 G20 연구장관회의(2021년)에 이르기까지 연구데이터 접근성 제고를 위한 지침과 실천 강조
- 최근 코로나19 팬데믹의 출현은 전 세계적 해결책 공동 모색을 촉구하는 계기가 되었고 WHO, CORD-19, GISAID, ICODA* 등을 통해 연구데이터 공유 활동을 확대시킴

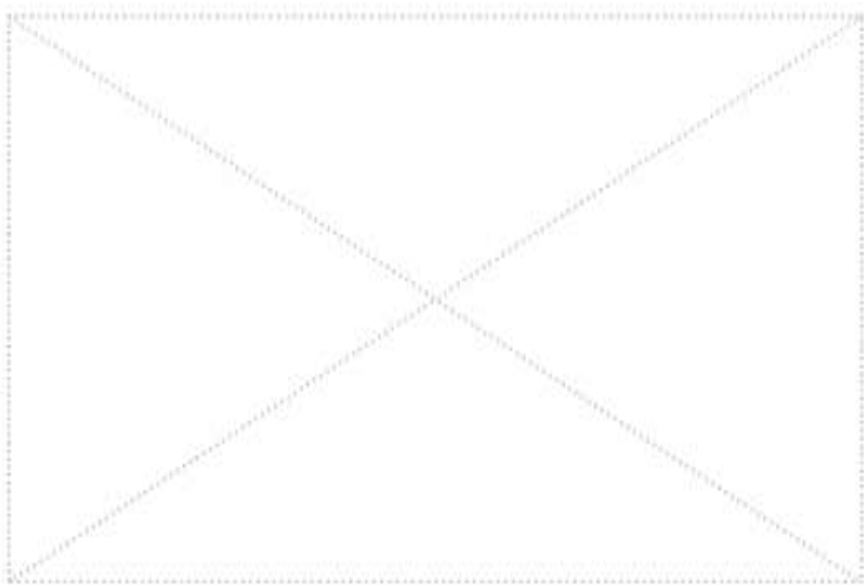
* CORD-19(COVID-19 Open Research Dataset), GISAID(ICODA(International GOVID-19 Data Alliance)

(3) 주요 출판사의 연구데이터 정책

○ 연구논문 검증을 위하여 필요한 데이터를 등록·검토·공개하는 출판사·저널 증가

- <Nature>는 2007년부터 논문 출판을 위한 데이터공유를 공론화했고, PLOS는 2014년부터 출판 전 데이터 공개가 안 된 논문은 거절하는 강경책을 도입함(Briney, Goben, and Zilinski, 2015)
- Springer Nature의 경우 2016년부터 표준적인 연구데이터 정책을 마련·적용하고 있으며, Talyor & Francis는 자사 저널에 대해 2018년부터 데이터 공유 정책을 권고(Jones, Grant and Hrynaszkiewicz, 2019)

[그림 2-7] 분야별 데이터정책 보유 저널의 비중: Taylor & Francis 사례



자료: Jones, Grant and Hrynaszkiewicz (2019) 그림 1 발췌

- 전 세계에서 발행하는 학술지의 23%를 담당하는 Elsevier 社는 연구데이터에 관한 원칙을 제시하고 이를 위한 구체적 실천지침을 권고하고 있으며, 데이터 공유를 위한 리포지터리 <Mendeley Data>와 데이터를 직접 출판할 수 있는 데이터 저널, <Data in Brief>를 발행 중임(Briney, Goben, and Zilinski, 2015)

○ 주요 출판사나 저널에서 논문출판을 위한 데이터정책을 마련·제시하는 추세가 확산되고 있지만, 연구데이터 설명·등록·공개의 수준과 세부 요건은 다양함

- Springer Nature社의 경우, 데이터 가용성(data availability)에 대한 설명, 데이터 리포지터리를 통한 데이터 공유, 출판 전 동료심사를 위한 데이터 제공 등을 단순 권장하는 저널부터 의무화한 저널까지 여러 수준의 데이터 정책 공존

- <Science> 등의 학술저널을 발행하는 AAAS는 보다 엄격한 데이터정책을 추진 중인데, 출판 전 데이터 리포지터리에 데이터를 등록하여 접속번호(accession number)를 제공하도록 하고 있으며 MIBBI 지침*을 준용하도록 권고[8)]

* Minimum Information for Biological and Biomedical Investigation guidelines

(4) 주요 연구기관의 연구데이터 정책

○ 미국 대학에서 기관 차원의 연구데이터정책과 연구데이터서비스(RDS) 확대

- 미국 연구중심대학의 경우, 2012년 초반만 해도 시작 단계였던 기관의 데이터정책이 확산되어 절반가량의 조사기관에서 독자적인 데이터정책을 마련하거나 지식재산정책에 연구데이터에 대한 고려를 포함(Briney, Goben and Zilinski, 2015)
- 미국 다수의 대학 도서관에서 학내 구성원의 데이터관리계획(DMP) 작성을 지원하고 데이터 설명, 보존, 관리 등에 관한 교육 서비스를 제공하는 등 연구데이터관리(RDM) 서비스를 제공하고 있음(김지현, 2014; Yoon and Schultz, 2017; Murray, et al., 2019)

○ 유럽연합 및 유럽 국가 단위 연구데이터 정책이 확대되면서 유럽 대학 및 연구기관의 데이터정책과 서비스 점차 일반화

- 유럽연구도서관연합(LIBER)의 최근 조사에 따르면, 미국과 같이 유럽에서도 도서관 서비스의 일환으로 연구데이터 서비스가 중요하게 여겨지고 있으며 이를 위한 기술적·조직적 지원과 컨설팅, 교육·훈련 등이 확대 추세(Tenopir, et al., 2017)

[그림 2-8] 유럽 연구도서관의 연구데이터서비스 현황(LIBER 자료)

주: 유럽 LIBER 회원 중 22개국 119 기관의 응답을 바탕으로 서비스 운영/계획 비중 계상
자료: Tenopir, et al.(2017), pp. 10-11 그림 발췌

8) Berkeley Library, (https://update.lib.berkeley.edu/2017/05/04/elsevier-springer-nature-aaas-publisher-research-data-policies/, 접속일: 2022.4.5.)

3. 주요국의 연구데이터정책

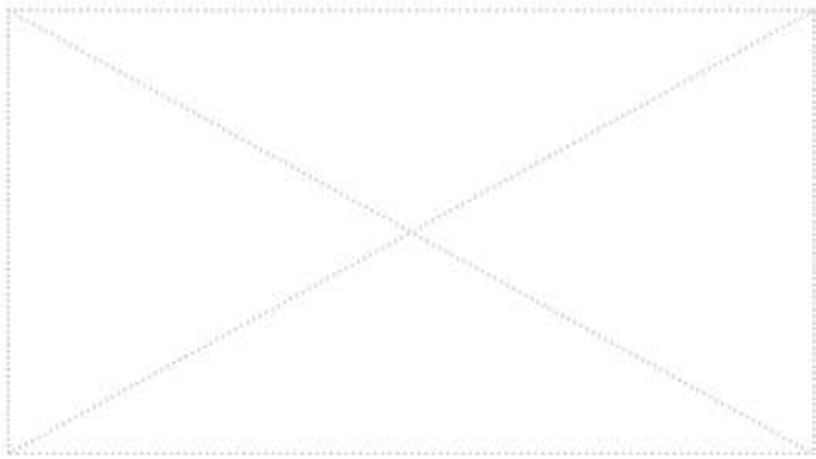
(1) 유럽연합의 연구데이터·데이터 정책

□ 유럽데이터전략

○ 유럽연합(EU)은 디지털 경제·사회로의 전환을 전망하며 유럽지역을 디지털 단일시장으로 육성하고 유럽전역의 데이터 생산, 관리, 활용을 촉진하는 이니셔티브 추진[9)]

– 유럽연합 차원의 디지털 단일시장(Digital Single Market) 구상 아래 유럽데이터스페이스(Common European Data Spaces) 구축 방안 제시

[그림 2-9] 유럽데이터스페이스 구축 계획



자료: Real-time Linked Dataspaces – Common European Data Spaces, http://dataspaces.info/common-european-data-spaces/#page-content (접속일: 2022.4.13.)

– 유럽 차원의 데이터 인프라 강화를 위하여 디지털유럽프로그램(DIGITAL, 2021-2027)을 통한 재정 지원, 호라이즌유럽프로그램(Horizon Europe, 2021-20270)을 통한 연구·혁신활동 지원, 유럽연합의 회복·복원시설(RRF, Recovery and Resilience Facility) 연계 등을 추진하고 국가간공동투자, 부문별개발활동 촉진 중(EC, 2022)

– 한편, 법제도적으로 공공정보(PSI)의 공개·활용을 촉진하는 한편 책임성있는 데이터 보호와 활용을 위하여 일반데이터보호규정(GDPR, 2018), 데이터거버넌스법

9) 유럽위원회(EC)는 2014년 데이터경제에 관한 협의(Communication on the data-driven economy)를 바탕으로 2017년 유럽데이터경제 구축안(Communication on building a European data economy), 2018년 유럽데이터스페이스 구축안(Communication on building a European data economy)을 연이어 마련하며 유럽연합 회원국의 의견수렴과정을 거침. 이에 따라 2018년 유럽데이터경제전략(European data economy strategy), 2020년 유럽데이터전략(European Strategy for data) 제시(EC, 2018.4.15.; EC, 2022)

(Data Governance Act, 2021), 데이터법(Data Act, 2022) 등 정비

□ 유럽연합의 연구데이터정책·사업

○ 연구·혁신활동과 직접적으로 관련이 있는 유럽연합의 데이터정책 또한 지속 추진되고 있으며, 현재 주요 정책은 다음과 같음

- 첫째, Horizon Europe 프로그램을 통해 데이터공유 및 분석 기술, 컴퓨팅 기술과 데이터 역량을 강화하는 연구개발활동을 촉진하고 산업적으로 클라우드에서 사물인터넷 엣지(IoT Edge)로 전환하는 데이터생태계를 만들어가는 활동을 지원 중임(EC, 2022)
- 둘째, Horizon Europe 프로그램을 통해 연구데이터의 체계적 생산과 관리를 촉진하고 연구데이터의 폭넓은 확산과 공유·활용을 위하여 FAIR 원칙 아래 데이터관리계획(DMP)을 포함한 연구데이터관리(RDM)*, 유럽오픈사이언스클라우드(EOSC)의 구축·활용 장려

※ 이전 Horizon 2020 프로그램에서부터 오픈연구데이터시범사업(ORDP)을 통해 연구데이터를 FAIR 원칙에 따라 검색가능하고(Findable) 접근가능하고(Accessible) 상호호환가능하며(Interoperable) 재사용가능한(Reusable) 형태로 관리, 공유하도록 권고해 왔으며, Horizon Europe 프로그램에서 적용 대상과 수위 확대[10)]

<표 2-1> Horizon Europe 프로그램의 연구데이터 관련 지침

구 분	요 약
기본 방향	• 정책의 주안점이 “오픈연구데이터(ORD)”에서 “연구데이터관리(RDM)”로 전환 • Horizon Europe 프로그램의 모든 과제 책임자는 디지털 연구데이터를 FAIR 원칙에 따라 책임성있게 관리해야 하며 다음과 같은 연구데이터관리(RDM) 조치를 취해야 함
주요 규정	• (과제 착수 6개월 이내) 데이터관리계획(DMP) 작성 (정기적 갱신) • DMP에 명시한 시점 내에 ‘신뢰성있는’ 리포지터리에 연구데이터 등록 – 신뢰성있는 리포지터리는 EOSC 요건에 맞추어 EOSC와 연계된 곳 • 연구데이터 등록시 ‘as open as possible, as closed as necessary’ 원칙 아래 데이터 접근 조치를 취함 (단, 필요시 접근 제한 상황을 DMP에 명기하면 예외 허용)
세부 규정	• 데이터 등록·공개시 데이터 검증 및 재사용을 위해 필요한 메타정보(관련 연구성과, 연구수단 등에 관한 정보) 제공, 메타데이터도 FAIR 원칙 아래 공개, 공개 데이터에 대해 오픈라이선스(CC BY 혹은 CC 0) 적용 • (공개) 지원프로그램에 명시되어 있거나 펀딩기관 요청이 있는 공공위기상황(Public emergencies)이나 과학적 출판결과의 검증을 위해 필요한 경우 연구데이터의 제공 의무 • (제한) 연구주체의 상업적 개발 수요나 유럽연합 경쟁력 관련 이슈 등 기타 사유 발생시

자료: EC(2021.12), EC(2021.11), European Research Council(2022), EC(2020) 자료 [11)] 요약

10) OS4OS(2022) “Where is Open Science in Horizon Europe”, https://openscience.eu/Open-Science-in-Horizon-Europe (접속일: 2022.6.1.)

□ 유럽 주요국의 연구데이터정책·지침

○ 유럽지역의 경우, 유럽연합뿐만 아니라 개별 국가, 연구펀딩기관, 연구기관 차원에서 연구데이터관리(RDM) 정책과 가이드라인을 마련하고 데이터관리계획(DMP) 작성이나 데이터 리포지터리 선정 및 활용에 관한 지침을 제공하는 사례가 일반화되고 있음

– 핀란드학술원(AKA)과 네델란드연구회(NWO)는 일찍이 연구데이터 및 오픈사이언스에 관한 국가 정책을 마련했고, 스위스국가과학재단(SNSF)도 2017년 오픈 연구데이터(ORD)정책을 마련

– 유럽연합 연구개발프로그램의 Model Grant Agreement에 반영된 연구데이터관리(RDM) 가이드에 따라, 핀란드와 네델란드는 기존의 DMP 지침을 갱신하고 프랑스국가연구청(ANR), 폴란드국가과학센터(NCN), 스웨덴연구회(VR)은 관련 DMP 요건을 신설함

<표 2-2> 유럽 주요국의 연구데이터관리(RDM) 가이드 개발 및 이행 현황

국가/기관	RDM 가이드 개발·협의 절차		RDM 가이드의 주요 내용	
	기관 내부 프로세스	기관 외부 프로세스	DMP 요청 시기	기타 사항
오스트리아 과학펀드(FWF)	• 특정 (전략) 부서 주관 • 타 부서 및 상위 조직 지원 • 집행이사회 승인	• 연구계 의견수렴 진행, e-infrastructure Austria 프로젝트의 WG에서 기타 이해관계자 의견 수렴	• 그랜트 승인 이후 DMP 요청 •	• DMP 갱신 • 과제 종료시 최종 DMP
프랑스 국가연구청(ANR)	• 특정 (디지털과학) 부서 주관 • 타 부서 지원 • 집행이사회 승인	• 프랑스과학기술정보원(INIST) 및 펀딩기관교류네트워크 통해 DMP 요건 협의 • 프랑스오픈사이언스위원회를 통해 기타 이해관계자 의견 수렴	• 프로젝트 착수 후 6개월 이내 DMP 요청	• DMP 갱신(30개월 이상 과제는 중반부 갱신 요청) • 과제 종료시 최종 DMP

11) EC(2020) “Exploitation & Open Science in Horizon Europe”, https://ec.europa.eu/research/participants/data/ref/h2020/other/events/2020-10-09/3_exploitation-ipr-open_science_en.pdf?msclkid=4862130abad211ecb6c1f7c64bf68939(2022.3.1.접속)

국가/기관	RDM 가이드 개발·협의 절차		RDM 가이드의 주요 내용	
	기관 내부 프로세스	기관 외부 프로세스	DMP 요청 시기	기타 사항
핀란드 학술원(AKA)	• N/A	• (협력구조) DMPTuuli Network Group 구성, DMP 요건 협의 • 동 Group 멤버/연구기관의 의견 수렴	• 그랜트 신청시 DMP 요청	• DMP 갱신
네델란드 연구회(NWO)	• 특정 (오픈사이언스) 부서 주관 • 타 부서 지원 • 집행이사회 승인	• (협력구조) UKB 연구데이터 WG 통해 DMP 요건 협의 • 연구기관 Data stewards 통해 연구계 의견 수렴 • DANS 통해 기타 이해관계자 의견 수렴	• 그랜트 승인 후 4개월 이내 DMP 요청	• 해당사항 없음
폴란드 국가과학센터(NCN)	• 특정 (감사) 부서 주관 • 타부서 및 과학위원회 대표 지원 • 과학위원회 최종 승인	• 해당사항 없음	• 그랜트 신청시 DMP 요청	• 과제 종료시 최종 DMP
스웨덴 연구회(VR)	• 특정 (연구인프라) 부서 주관 • 타 부서 및 과학위원회 대표 지원	• RDM WG 통해 DMP 요건 협의, 연구기관 및 기타 이해관계자 의견 수렴	• 그랜트 승인 후 프로젝트 시작 전 DMP 요청	• DMP 갱신
스위스 국가과학재단 (SNSF)	• 공동 개발 • 확대권리위원회 지원 • 집행이사회 승인	• 서로 다른 스위스연구기관들의 그랜트 사무소를 통해 연구계 및 기타 이해관계자 의견 수렴	• 그랜트 신청 시 DMP 요청	• DMP 갱신 • 과제 종료시 최종 DMP

주: 원 자료에서 특정 학문분야에 국한된 아일랜드 사례 제외
자료: Science Europe (2020), p.12 & p.18 표 발췌 및 요약

□ 유럽연합의 연구데이터 리포지터리 기준

○ 유럽차원의 연구데이터관리(RDM) 확대를 위하여 “신뢰성있는” 리포지터리(trusted repository)의 기준에 관한 협의가 발전하고 있고 해당 논의 일부는 유럽연합 Horizon Europe 그랜트 표준협약(MGA) 지침에 반영·활용 중임(Science Europe 2020)

<표 2-3> EC의 신뢰성있는 리포지터리(trusted repository)의 기준

구 분	상 세 내 용
리포지터리	• 리포지터리는 온라인 아카이브로서 연구자가 디지털 연구 성과(출판물, 데이터, 소프트웨어, 기타 성과 등)를 등록하고 해당 성과에 대한 (오픈) 액세스를 제공하며 이를 장기 보전할 수 있는 곳임 • 리포지터리는 대학이나 연구기관 등 기관 차원에서 구축된 곳과 생물학, 물리학 등 특정 분야 중심적으로 구축·운영되는 곳, 이외에도 전 학문분야를 지원하는 범기관 성격의 범용 리포지터리(예: CERN에서 개발한 Zenodo)가 있음 – 단, 개인 웹사이트나 DB, 출판사 웹사이트, 클라우드서비스(Dropbox, Google drive 등)을 리포지터리라고 할 수는 없으며, 요구조건에 따른 오픈액세스를 제공하지 못하는 유사 플랫폼(예: Academia.edu, ResearchGate 등)도 리포지터리로 간주하지 않음
신뢰성있는 리포지터리	• 신뢰성있는 리포지터리는 공인된 인증(예: CoreTrustSeal, nestor Seal DIN31644, ISO16363 등)을 받은 리포지터리나 학계에서 분야/도메인 리포지터리로 인정받은 리포지터리로서 글로벌 차원의 인지도가 있는 곳임 □ 신뢰성있는 리포지터리의 요건 • 콘텐츠/데이터의 진실성과 사실성을 보장하기 위한 조직적, 기술적, 절차적 특성 정보를 구체적으로 표기하고 콘텐츠의 장·단기적 이용 및 재이용을 촉진시킴(특정 규정을 갖추고 자신의 서비스와 정책에 관한 명확한 정보* 제시) * (예) 콘텐츠의 확보, 접근 및 보안, 서비스 재원을 포함한 지속가능성에 관한 정보 등 • 가능하면 이용 시점에 무료로 콘텐츠를 오픈 액세스할 수 있도록 포괄적이고 평등한 서비스 제공(단, 법적·윤리적 제한 고려) – 콘텐츠/데이터와 주요 주체에 대해 영구적이고 고유한 식별자(PIDs)를 부여하여 일반적으로 인식, 참조, 인용할 수 있도록 함 – 콘텐츠/데이터의 검색, 재이용, 인용이 가능하도록 충분히 상세한 메타데이터를 함께 제공하고 메타데이터도 기계실행(machine-actionable)가능하며 표준화 – 콘텐츠/데이터 액세스 및 이용에 대한 라이선스 구체화 및 관련 정보 제공(오픈시 CC BY 나 이에 준하는 라이선스 권장, 경우에 따라 상업적 활용이나 이차 활용 제한 규정 도입 • 등록 자료의 중장기적 보존 장려, 데이터/메타데이터/관리절차의 정확성과 진실성을 유지하고 문제 발생시 대응하는 절차와 조직을 보유하며, 대체로 국제적/국내적으로 인정된 보안기준과 절차, 여건을 충족하여 불법적 액세스나 민감·개인정보 침해 이슈를 방지

자료: EC(2021.11), p. 155-156 발췌 및 요약

(2) 미국의 연구데이터·데이터 정책

□ 미국의 공공액세스정책과 연구데이터규정

○ 미국 연방정부는 2013년부터 공공액세스정책(Public Access Policy)을 마련·공표하고 이에 따라 연방재정지원을 받는 주요 연구(펀딩)기관은 연구출판물뿐만 아니라 연구데이터에 대한 접근성 제고 조치를 마련·이행(신은정 외, 2021)

- 2013년 2월 미 백악관 과학기술정책국(OSTP)에서 공공액세스 메모렌덤(Memorandum on Increasing Access to the Results of Federally Funded Scientific Research)을 발표하고, 이에 따라 연간 1억 불 이상의 연방재정지원을 받는 연구(펀딩)기관은 지원과제의 결과(논문과 데이터)에 대한 접근성 제고 계획 및 조치를 마련·이행 중*

* 연방정부기관의 FOIA 이행 요건 일환으로 2 CFR 200.315 (e) 법령에 근거, 재정관리측면에서 관련 조항 OMB Circular A-110에도 명시

○ 연방정부 차원의 공공액세스정책 및 오픈사이언스정책 추진을 모니터링하고 재조정하기 위하여 대통령 산하 과학기술정책국(OSTP)에서 매년 공공액세스정책 추진 경과와 성과를 모니터링·보고하고 있으며, 국가과학기술위원회(NSTC) 산하에 오픈사이언스소위(Subcommittee on Open Science, SOS)를 구성·운영 중임

○ 연방정책에 따라 산하 연구(펀딩)기관 차원의 연구데이터 규정과 서비스 확대

- 구체적으로, 연간 1억 불 이상의 연방재정지원을 받는 20여개 이상의 모든 연구(펀딩)기관에서 공공액세스정책에 따른 계획과 정책을 마련하였고, 연간 1억 불 이하의 연방재정지원을 받는 다수의 기관에서도 관련 계획과 정책 마련[12)]

[그림 2-10] 미국 연방부기관의 공공액세스/데이터정책 도입

자료: 신은정 외(2019a) p. 56 발췌 (원 자료: CENDI(2017) 웹페이지, 신은정·정원교(2017) 및 SPARC(2016)를 바탕으로 도식화)

12) OSTP(2021) 2021 Public Access Congressional Report, https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf

- 연방정부 지원을 받은 학술논문(출판물)의 접근성을 제고하기 위하여 학술논문의 OA출판이나 OA리포지터리 등록을 의무화하고 있으며, 연구데이터의 경우 데이터 관리계획의 작성 및 체계적 연구데이터관리를 통해 데이터의 관리·활용가치를 제고하도록 하고 있음
- DMP의 작성 및 계획을 통해 과제 단위 데이터 관리를 체계화하고 연구기관 등의 관리 책임을 구체화하도록 유도하며, 연구데이터에 대한 고유식별자 부여, 메타데이터의 제공, 과제기간을 초월한 중장기적 데이터의 보존과 관리, 공유·활용을 장려
- 국립보건원(NIH), 지질조사국(USGS), 에너지부(DOE), 국가기술표준원(NIST) 등은 기관 차원의 데이터(서비스)센터를 구축·운영하고 있고, 국립과학재단(NSF) 등은 분야·기관 단위 OA 리포지터리의 요건과 권장 리스트를 제시하고 자발적 선택을 권장하고 있음
- 에너지부(DOE) 지원 과제에서 산출된 소프트웨어의 공개 검색 및 활용을 위하여 디지털 플랫폼, DOE CODE가 발족되었고[13] 국가기술표준원(NIST)는 오픈소스 코드 플랫폼[14] 구축·운영 중이며 지질조사국(USGS)은 Model Catalogue를[15] 통해 과학모델 공유(OSTP, 2021)

○ 연방지원을 통해 생산된 연구데이터의 접근과 관리, 공유 및 활용을 촉진하기 위하여 NSTC 산하 오픈사이언스소위(SOS)는 2022년 바람직한 데이터 리포지터리의 요건과 특성에 관한 권고안 발표

<표 2-4> 미국 NSTC의 바람직한 리포지터리 특성

구 분	상 세 내 용
조직적 인프라	**(Free and Easy Access)** 법·정책적 요건을 고려하여 적시에 비용없이 데이터와 메타데이터에 대한 최대한의 오픈액세스 제공
	(Clear Use Guidance) 데이터의 접근 및 활용, 재이용 등에 관한 분명한 정책 문서 보유
	(Risk Management) 보안, 위기관리, 민감데이터 관련 모니터링 등을 위한 기술적, 행정적, 물리적 세이프가드의 구축 상황 명시
	(Retention Policy) 데이터 보유에 관한 정책 문서 제공
	(Long-term Organizational Sustainability) 예측불가능 사건에 대비한 컨틴전시 플랜을 포함한 장기적 데이터관리 계획 마련
디지털 매체 관리	**(Unique Persistent Identifiers)** 데이터의 검색, 보고, 평가 등을 가능하게 하기 위하여 데이터 등에 DOI와 같은 고유한 영구식별자 부여
	(Meta data) 데이터의 검색과 재이용, 인용을 위하여 널리 활용 가능한 메타데이

13) https://www.osti.gov/doecode/ (접속일:2022.6.2.)
14) https://code.nist.gov/ (접속일:2022.6.2.)
15) https://data.usgs.gov/modelcatalog/ (접속일:2022.6.2.)

	터의 적절한 제공
	(Curation and Quality Assurance) 데이터와 메타데이터의 정확성과 신뢰성 제고를 위한 품질관리 및 전문가 큐레이션 제공·장려
	(Broad and Measured Reuse) 데이터 재이용에 관한 사항을 다루는 메타데이터의 제공 및 데이터 추출, 가공·정제, 재이용을 가늠할 수 있는 역량 지원
	(Common Format) 데이터와 메다테이서가 분야에서 널리 활용되며 표준에 부합하는 일관성있는 포맷으로 접근, 다운로드, 추출될 수 있도록 허용
	(Provenance) 제출된 데이터와 메타데이터의 기원, 관리 경과, 버전 관리, 기타 수정 이력 등을 기록하는 메커니즘 보유
기술	**(Authentication)** 데이터와 데이터 제공자의 고유식별자 연계 등을 통해 데이터 제공자의 진위 검증 지원
	(Long-term Technical Sustainability) 안정적 기술 인프라와 펀딩 계획을 바탕으로 장기적 데이터관리 계획 마련
	(Security and Integrity) 데이터 민감성에 맞는 차등적 보안 절차를 마련하고 승인받지않은 데이터의 접근, 가공, 유출을 방지하는 잘 정비된 사이버 보안 기준과 조치를 문서화

자료: NSTC(2022), p. 4-5 발췌 및 요약

□ 미국의 데이터중심연구지원사업

○ 데이터중심연구, 디지털인프라 고도화를 위한 연방 R&D 프로그램 지속

- 90년대 고성능컴퓨팅 기반 연구개발에서부터 인공지능, 빅데이터, 유무선 네트워크 등 관련 기술개발이 고도화되면서 미국 연방정부 차원의 범부처 사업으로 네트워킹·정보기술 R&D (NITRD) 프로그램을 91년 이래 지속 운영[16)]

* 빅데이터 관리·분석(LSDMA), 고성능컴퓨팅 인프라·응용(HCIA), 인공지능(AI) 등 11개 분야에 대해 연간 약 78억 불 지원

[그림 2-11] 미국 NITRD 프로그램의 연도별·기관별 지원 예산

[최근 20년간 NITRD 예산 규모 추이]

[2022 회계연도 기관별 NITRD 예산요구비중]

16) https://www.nitrd.gov/about/ (접속일:2022.6.2.)

자료: (왼쪽) NITRD 웹사이트 그림 발췌, https://www.nitrd.gov/about/ (접속일: 2022.4.1.)
(오른쪽) NITRD (2021), p.13 발췌

□ 미국의 연방데이터전략

○ 한편, 미국 연방정부 차원에서 데이터기반 행정, 데이터기반 산업을 활성화하기 위한 정책 또한 추진

- (연구) 일찍이 국가과학기술위원회(NSTC)의 NITRD 소위에서 “Federal Big Data R&D Strategic Plan (2016)”을 제시한 바 있으며, 관련 R&D 프로그램 지속 증가[17]
- (행정) 2019년 연방데이터전략(Federal Data Strategy)으로 전환하면서 모든 연방기관에 최고데이터책임자(CDO) 직제를 마련하는 등 데이터중심 행정체계 강화 (US Federal Data Strategy Team, 2019)
- (상업) 미 상무부의 “Commerce Data Strategy(2021-2024)”를 통해 데이터 관리를 전략적 자산화하고 적절한 데이터의 접근 및 이용을 장려하기 위한 조치와 역량 강화, CDO/CIO/CFO/EO/SO 등의 핵심 업무 연계 등을 제시(US Department of Commerce, 2021)
- (기타) 2021년 미 농무부도 ‘데이터 전략’을 마련하고 ‘국방부 데이터 명령’도 발표됨[18]

[그림 2-12] 미국 연방데이터 전략: 2021 실행계획

」

자료: U.S. Federal Data Strategy, https://strategy.data.gov/2021/action-plan/ (접속일: 2022.5.1.)

17) 이러한 연장선상에서 NSF는 2022년 사회과학분야 연구데이터 생태계 구축을 위한 3,800만불 프로그램 추진 계획을 발표(NSF 홈페이지, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1946932&HistoricalAwards=false (접속일: 2022.4.1.))

18) 이외에도 바이든 행정부는 2022년 3월 “디지털 자산의 책임성있는 개발”에 관한 행정명령을 통해 미국의 금융 안전성과 리더십을 유지하기 위한 디지털 자산관리방침 구체화

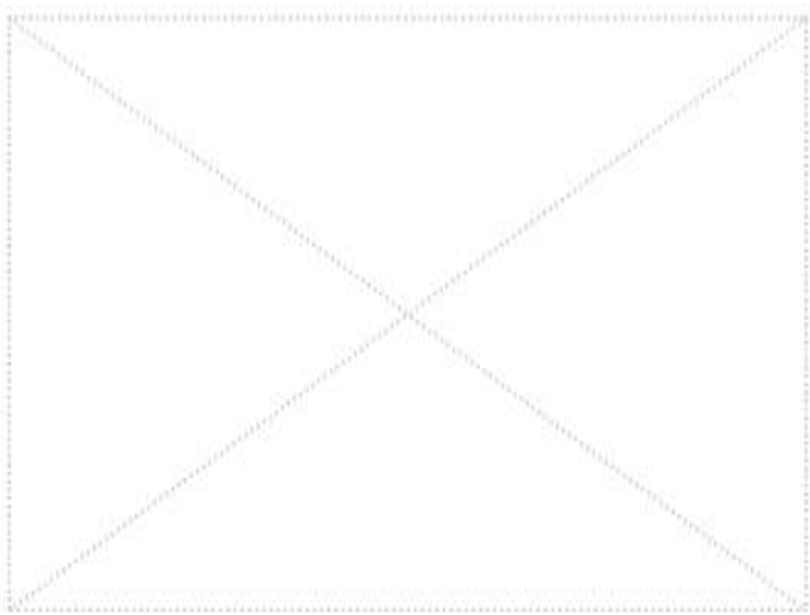
(3) 일본의 연구데이터·데이터 정책

□ 일본의 포괄적 데이터전략

○ 전 산업·사회적 차원에서 데이터 관리·활용 수요가 증가하고 있고 미국, 유럽 등지에서도 전방위적 데이터 정책·전략이 마련되는 동향 고려, 일본 정부는 21년 6월 「포괄적 데이터 전략」 마련(KIAT 산업기술정책센터, 2022.5.9.)

- 20년 디지털 정부 각료회의에 '데이터 전략 T/F' 설치, 데이터 전략(안) 마련 후 실무작업반 등을 통해 검토·수정을 거쳐 21년 최종안 제시
- 기존 디지털 전략을 총괄하던 내각의 IT 종합전략본부 대신 21년 '디지털사회추진회의'와 '디지털청' 신설, 「디지털 사회 실현을 위한 중점 계획」 추진
- 21년 제시된 「포괄적 데이터 전략」은 7단계 과제와 2개 횡단 요소로 구체화되어, 범부처 차원에서 개별 요소·단계별 대응방안과 체계를 마련

[그림 2-13] 일본의 포괄적 데이터 전략



자료 : KIAT 산업기술정책센터(2022.5.9.), p.12 발췌 (원 자료: 内閣, 包括的データ戦略)

□ 일본의 연구데이터정책·사업

○ 일본은 과학기술혁신을 촉진하는 차원에서 연구데이터의 개방·폐쇄 균형 전략을 강조
  – 일찍이 제5기 과학기술혁신기본계획에서 연구데이터의 개방·폐쇄 전략을 제시한 바 있으며, 국가 차원의 계획·전략을 개별 부처, 공공펀딩기관, 국립연구기관, 대학, 연구자협의회 등 다층적 차원에서 적용·이행하며, NII를 비롯한 디지털 연구인프라·서비스 확대 및 연계를 통해 연구데이터관리 장려(신은정 외, 2019b)

[그림 2-14] 일본 연구데이터정책 이행체계

자료: 신은정 외(2021) p. 171 그림 발췌 (원 자료: Hayashi (2020.9.15) UNESCO 아태지역회의 발표자료 수정)

  – 21년 수립된 제6기 과학기술혁신기본계획에서도 연구데이터의 개방과 폐쇄의 균형 전략을 유지하며[19] 세부과제로서 연구계 디지털 전환에 대응하며 신뢰성있는 데이터의 생산·관리 및 적절한 데이터의 공유·활용을 통해 연구과정의 효율화, 선진화, 연구성과의 수월성/영향력 제고, 다양한 주체의 연구참여 확대 제시
  – 전반적으로 일본은 연구의 디지털 전환(DX)을 지원하는 차원에서 필요한 데이터 생산·공유·활용 기반을 확충하고 네트워크·컴퓨팅·데이터 역량을 강화하며 새로운 DX 기반 연구활동을 장려하는 투자 지향

19) 閣議決定(2021) 科学技術・イノベーション基本計画(令和３年３月２６日)

<표 2-5> 일본 연구데이터 관련 주요 지표

(현재)

- 연구데이터 정책을 수립한 국립연구개발법인의 수('20년 9월 기준) : 11개소
- 데이터관리계획(DMP)를 도입한 연구지원 부처/기관의 수('20년 9월 기준) : 8개소
- 일본 역내 기관 리포지터리의 수('19년 기준) : 811개소

(목표)

- 데이터 정책 수립 기관 비중 : 2025년까지 기관 리포지터리 보유 대학/국립연구개발법인 등의 60-100% 수준으로 향상
- 데이터관리계획(DMP) 도입 및 메타데이터 부여 비중 : 2023년까지 공모형 연구지원 신규 과제 전체에 대해 적용

자료: 제6기 과학기술혁신기본계획 전문, p. 58 - 60 일부 발췌

(4) 중국의 연구데이터 정책

□ 중국의 디지털 전환 정책과 데이터연구지원

○ 중국은 디지털 첨단기술분야에서 국가 경쟁력을 확대하기 위한 정부 주도 투자와 지원을 지속해오고 있으며, 2021년 '14.5 규획'에서 디지털 전환(DX) 정책 전면화 (KIAT, 2022)

- 성장 둔화, 글로벌 산업 전환 국면을 돌파하기 위하여 14.5 규획 및 2035 비전 목표 강요에서 디지털 전환(DX) 과제 제시
- 디지털 발전의 가속화와 디지털 중국의 건설을 위하여 디지털 경제, 디지털 사회, 디지털 정부, 디지털 생태계 관련 정책 과제 목표 제시

○ 국가 차원의 과학기술정책이 2010년 이후 확대되고 있으며, 2015년 13차 과학기술혁신 5개년 계획에 3개 유형의 데이터·정보 기반 연구사업/정책과 인력양성사업, 기업지원사업 등 총 5개 유형의 데이터지원사업 추진(Zhang, et al. 2021)

- 지난 30년간 데이터중심 연구에 대한 국가자연과학기금위원회(NSFC)[20] 지원 증가 추세

20) 중국 국가자연과학기금위원회(NSFC)는 2018년 과학기술부(MOST)로 통합됨

[그림 2-15] 중국 국가자연과학기금위원회(NSFC)의 데이터 관련 지원

자료: Zhang, et al. (2021) p.9 그림 4 발췌

□ 중국의 과학데이터관리정책

○ 국가 차원의 과학데이터의 체계적 생산과 관리, 공유를 위한 제도·인프라 정비 중
- 2018년 국가 차원의 과학데이터 관리 제도 '科学数据管理办法' 공표,[21] 과학데이터의 전 과정 관리 방침 제시, 이에 따라 중국농업과학원(CAAS), 중국과학원(CAS)과 같은 주요연구기관 및 지방정부 차원의 과학데이터관리 규정 확산 중(Zhang, et al. 2021)
- 국가 차원의 과학데이터 수집, 관리, 공유를 위하여 2019년 20개의 국가과학데이터센터 구축, 주요 연구인프라시설의 공유 장려 중(Ma, 2020.9.16.)

○ 연구기관/대학 단위 데이터센터 운영 및 연구데이터관리 정책 이행 현황
- 중국과학원(CAS)는 대표적 국립연구원으로서 2010년 이후 데이터 그리드, 데이터 클라우드 등을 통해 데이터의 공유 및 데이터중심연구를 장려해왔고, '빅데이터 엔지니어링' 프로그램 추진, 2018년 '과학데이터관리제도(科学数据管理办法) 공표 이후 관련 데이터 책임과 의무를 더욱 구체화(Zhang, et al. 2021)
- 대학의 경우, 2014년 중국학술도서관 연구데이터관리 이행그룹이 결성된 바 있으나 아직까지 개별 기관의 연구데이터관리 정책과 인프라간 편차는 큰 편임(Huang, Cox, and Sbaffi, 2021)

21) 중국 国务院办公厅关于印发 科学数据管理办法的通知 – 国办发 [2018] 17号, http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm (접속일: 2020.4.1.)

제2절 국내 동향

1. 국가차원의 데이터정책

(1) 데이터정책 및 법제도의 마련

○ 부문과 지역을 넘나드는 전방위적 디지털 전환과 4차 산업혁명을 예견하면서 국내 정책과 제도도 지속 보강·정비되고 있음

– 2017년 대통령직속 4차산업혁명위원회가 마련되고 21년부터는 4차산업혁명위원회 산하 데이터특별위원회가 신설·운영됨

– 데이터산업 및 인프라 육성을 위하여 일련의 국가정책과 전략, 국책사업 추진

<표 2-6> 국가 차원의 데이터 정책 및 이니셔티브

연도	주요 정책 및 제도 개선	주관(실무) 기관
2017.11.	4차 산업혁명 대응방향	4차산업혁명위원회
2018.6.	데이터산업 활성화 전략	4차산업혁명위원회
2019.1.	데이터·AI 경제 활성화 계획	관계부처합동(혁신성장전략회의)
2020.1.	데이터3법(개인정보보호법, 정보통신망법, 신용정보법) 개정	국회(과학기술정보통신위원회)
2020.4.	한국판 뉴딜(디지털 뉴딜 포함) 종합계획	관계부처합동(비상경제회의)
2020.6.	한국판 디지털 뉴딜(데이터댐 포함) 사업계획	(과기정통부)
2021.2.	국가데이터정책 추진방향	4차산업혁명위원회
2021.2.	디지털 지식재산 혁신 전략	국가지식재산위원회
2021.6.	데이터플랫폼 활성화 방안	4차산업혁명위원회

○ 데이터기반 행정, 데이터산업, 데이터기반 지식체계 등을 위한 제반 법제와 조직, 사업 등도 확대 추세임

– 데이터기반 행정을 위하여 기존 공공정보법과 공공데이터법 이외 2020년 [데이터기반 행정 활성화에 관한 법률]이 신설되어 관련 조직과 이행체계를 보강

– 2021년에는 [국가지식정보의 연계 및 활용 촉진에 관한 법률]이 신설되고 국가지식정보연계를 위한 조직과 이행체계를 정비하고 관련 계획을 수립

– 2019년 말 이루어졌던 데이터3법(개인정보보호법, 정보통신망법, 신용정보법) 개정에 이어, 2022년 초에는 [산업디지털전환촉진법]이 새롭게 제정됨에 따라 산업데이터의 생산, 관리, 활용규범을 구체화하고 데이터거래를 활성화하는 법제 기반이 확대됨

※ 2021년 10월 신설된 [데이터산업·이용촉진 기본법]은 2022년 4월 시행

○ 국가 차원의 데이터정책과 법제도가 확대되면서 관련 활동을 지원하는 데이터 플랫폼 구축 및 연계·활용 계획 증가

[그림 2-16] 국가 차원의 데이터 플랫폼 관련 주요 계획의 내용

<디지털 집현전 통합 플랫폼 계획>

<데이터기반 행정 활성화 계획>

< 디지털 산업혁신 빅데이터 플랫폼>

산업부 산하 한국산업기술시험원(KTL) 주관으로 가이온, 에프앤가이드, 한국M&A 거래소, 대덕넷, 알리콘, 에스티에이치에이에스, 이엔씨지엘에스 등 총 7개 데이터센터와 함께 디지털 산업혁신 빅데이터 플랫폼 구축 계획('22)

(2) 신정부 국정과제

○ 2022년 5월 발표한 윤석열정부 110대 국정과제에서 정부, 산업, 사회부문에 걸친 전방위적 플랫폼의 활용을 제시

- (11) 디지털플랫폼정부, 범정부 데이터·서비스의 개방, 연계, 활용 인프라 구축

  * 디지털플랫폼정부 추진을 위한 민관합동위원회, 국가 데이터정책 컨트롤타워, 범부처·민관합동 디지털 국가전략 수립 및 디지털 혁신위원회 신설 검토

- (22) 수요자 지향 산업기술 R&D 혁신 위한 산업기술 빅데이터 플랫폼 구축 및 (23) 주력산업 고도화를 위한 산업데이터 플랫폼 구축(디지털 혁신 허브 지정·확대 및 디지털 혁신 기업 지원), (77) 디지털플랫폼 등 신산업 육성
- (2, 25) 보건의료, (28) 국토교통, (65) 재난안전, (71) 농림축산 빅데이터 공유·활용 촉진

○ 한편, (74) 국가혁신을 위한 과학기술시스템 재설계 차원에서 창의·혁신적 연구성과 창출을 위한 국가 연구데이터 플랫폼 구축, 대학·연구기관의 디지털 연구환경 조성 등을 세부과제로서 제시

2. 과학기술 연구데이터정책

(1) 연구데이터사업 및 전략의 도입

□ 과학기술분야 개별 연구데이터사업의 전개

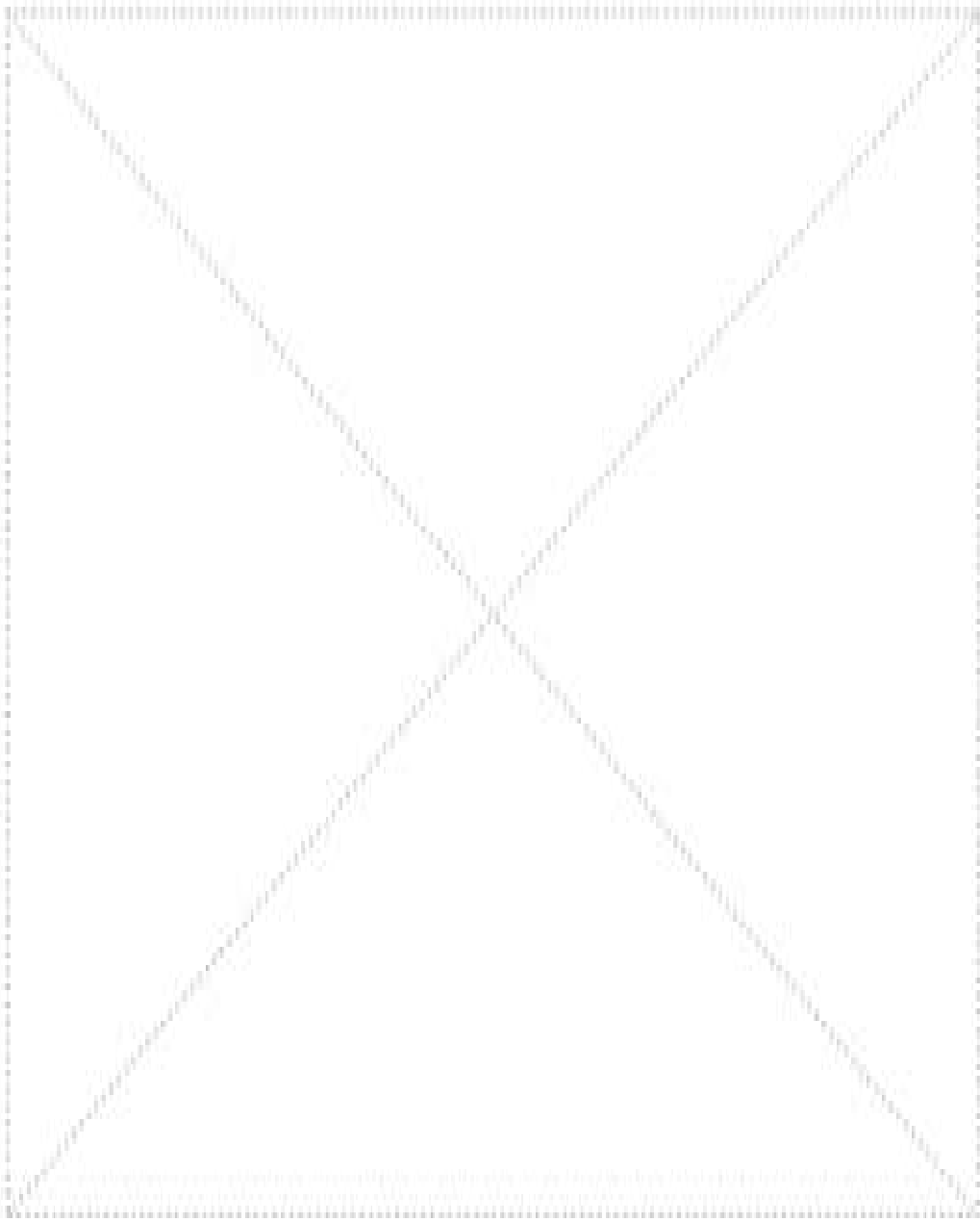
○ 과학기술분야에서는 데이터중심연구를 가속화하기 위한 일환으로 2014년 전후로 빅데이터사업을 기획·추진해왔고, 2018년 민관합동 빅데이터 T/F 활동의 일환으로 과학기술데이터의 공유·활용을 촉진하는 방안을 마련한 바 있음

○ 한편, 데이터중심연구를 지원하는 핵심인프라로서 슈퍼컴퓨팅자원에 대한 투자와 분야별 소요에 따라 데이터중심연구를 위한 대형국책연구사업이 기획·추진됨
- 물리학 분야에서 유럽입자물리연구소(CERN) 데이터 기반 국제공동연구는 국내 대표적인 빅사이언스 연구사례로서 2006년 한-CERN 협력사업 양해각서가 체결된 이후 지속 확대(신은정 외, 2018)
- 바이오연구분야에서 데이터중심연구를 촉진하기 위하여 시작된 <포스트게놈 다부처 유전체 사업>은 2014년 시작되어 2021년까지 8년간 추진
- 이외에도 2000년대 이후 대형연구시설장비 및 공공연구기관 차원에서 대량의 연구데이터를 지속 수집, 관리, 공유, 활용하는 활동이 분야별로 등장

□ 연구데이터 공유·활용전략의 도입

○ 국가 R&D를 통해 생산된 연구데이터의 공유·활용을 촉진하기 위한 정부 전략 제시
- 2017년 하반기부터 과학기술정보통신부 중심으로 논의가 시작되어 연구데이터의 빅데이터화를 위한 “연구데이터 T/F(모아서 새롭게 T/F)”를 구성, 12월까지 총 26회 회의
- 관련 논의를 집약하여 2018년 1월 「연구데이터 공유· 활용 전략」이 국가과학기술심의회 운영위원회 안건(2018. 1. 19.)으로 상정·의결됨
- 동 전략에서는 ’연구데이터의 지식자산화, 빅데이터화‘를 위하여 (1) 데이터 공유·활용의 기본 원칙 및 커뮤니티 형성, (2) 연구데이터플랫폼 등 인프라 구축·운영, (3) 데이터관리계획 이행근거를 포함한 관련 법제도 정비, (4) 데이터 집약형 연구 모범사례 발굴 및 성과 창출형 시범사업 추진 등을 제시

[그림 2-17] 연구데이터 공유·활용 전략의 주요 내용



자료: 국가과학기술심의회 운영위원회(2018), p. 5 발췌

○ 「연구데이터 공유· 활용 전략」에 따라 후속 조치 시행

- 2018년부터 연구데이터 공유 커뮤니티 형성을 촉진하고 국가연구데이터플랫폼 구축·운영을 촉진하기 위한 시범사업 추진*

  * 바이오, 미래소재, 인공지능, 대형연구장비 등의 데이터집약형 연구 시범사업 추진(신은정 외 2019a)

- 2019년 연구데이터 관련 법제도 개선 차원에서 「국가연구개발사업의 관리 등에 관한 규정」 개정을 통해 국가R&D사업 중 중앙행정기관의 장이 필요하다고 인정하는 사업에 대하여 데이터관리계획(data management plan)을 수립·이행할 수 있는 법적 근거를 마련함

  * 다만, 해당 규정은 「국가연구개발혁신법」의 제정·시행과 함께 「국가연구개발사업의 관리 등에 관한 규정」이 대체됨에 따라 데이터관리계획에 관한 조문은 「국가연구개발정보처리기준」으로 이관

<표 2-7> 국가연구개발정보처리기준의 연구데이터 관련 조문

구분	주요 내용
제2조 (정의)	(연구데이터) 연구개발과제 수행 과정에서 실시하는 각종 실험, 관찰, 조사 및 분석 등을 통하여 산출된 사실자료로서 연구결과의 검증에 필수적인 데이터 (데이터관리계획) 연구데이터의 생산, 보존, 관리 및 공동활용 등에 관한 계획
제23조 (연구데이터의 관리)	① 중앙행정기관의 장은 필요하다고 인정하는 국가연구개발과제에 한하여 과제를 수행하고 있거나 참여하려는 자에게 데이터관리계획을 작성하여 과제협약 시 제출하게 할 수 있다. ② 중앙행정기관의 장은 국가연구개발과제를 선정할 때 데이터관리계획에 따른 연구데이터 생산·보존·관리의 충실성 및 공동활용 가능성 등을 검토하여야 한다. ③ 연구개발기관은 데이터관리계획에 따라 소관 국가연구개발과제의 연구데이터를 관리하고, 그 결과를 최종보고서에 포함하여 제출해야 한다. ④ 중앙행정기관의 장은 연구데이터의 생산·보존·관리 및 공동활용 등에 관한 시책을 수립·추진할 수 있다.

(2) 연구데이터 전략의 이행 및 연구데이터플랫폼의 구축

□ DMP 적용 대상 국가 R&D 과제의 확대

○ 2019년 「국가연구개발사업의 관리 등에 관한 규정」 개정에 따라 과학기술정보통신부에서 지원하는 국가 R&D 사업(한국연구재단 및 IITP 지원 사업)의 일부에 대해 데이터관리계획(DMP) 작성 요건 적용

○ 2019년 DMP 작성 요건이 부과된 사업은 바이오, 나노소재, ICT 분야 총 94개 과제였으며 2020년에는 적용 대상이 확대되어 바이오, 나노소재, ICT, 반도체 분야와 다부처사업에 해당하는 215개에 대해 DMP 작성을 요청(신은정 외, 2021)

[그림 2-18] 데이터관리계획 적용 사업의 확대

자료: 신은정 외(2021), p. 49 & p. 55 발췌 후 일부 수정

○ 한국연구재단은 지원 사업에서 산출된 연구데이터의 수집·공유·활용을 위하여 '연구과제 신청, 평가와의 연계 및 사후관리를 위한 절차'에 DMP를 반영하여 운영

<표 2-8> 연구과제 수행절차에 따른 DMP 수행내용

① 과제 공고 및 접수	⇨	② 선정 평가	⇨	③ 과제 진행	⇨	④ 연차·최종 평가
• 공고 시 DMP 적용원칙 안내 • DMP 작성·제출 공지		• 선정 평가 시 평가지표 반영		• 저장·공유장소로 KISTI (Tera0) 활용		• 연구성과 평가항목에 반영 • 데이터관리계획의 이행 및 갱신

자료: 한국연구재단(2020) 내용 바탕으로 수정, 신은정 외(2021) p. 49 재인용

□ KISTI 주관 국가연구데이터플랫폼(DataOn)

○ 한국과학기술정보연구원(KISTI)의 기관주요사업으로서 국가연구데이터플랫폼 (DataOn) 구축·운영[22)]

- 20년 1월 오픈한 DataOn은 분야를 넘어서는 연구데이터 통합 검색, 관리, 활용을 위하여 국내·외 연구데이터 센터, 학술정보 및 R&D 정보와의 연계 지속
- 21년 2월 기준 KISTI의 GSDC, 한국정보화진흥원의 AI hub, 한국지질자원연구원, 한국표준과학연구원과 연계를 완료하고, 유럽 최대 연구데이터 플랫폼인 OpenAIRE, 호주 국가연구데이터 플랫폼인 ARDC, 일본 IRDB 등과도 MOU를 체결·연계(KISTI, 2021)

* 2021년 12월 현재 국내 약 3.3만 데이터셋 등록, 해외 데이터 약 113만 메타데이터 연계

□ 소재연구데이터플랫폼 구축사업

○ 2020년 4월 과학기술정보통신부 소재부품장비 기술특별위원회를 통해 「소재 연구데이터 플랫폼 구축방안」 수립 및 확정

- 연구데이터 관리를 위해 DMP를 적용하던 기존 대상(미래소재 디스커버리사업 등 36개 과제 과제, 350억원 규모)을 2020년 전체 나노·소재 R&D 사업 (84개 과제 약 895억원 규모)로 확대하고, 2021년에는 집단연구, 타부처 소재 R&D 사업 등으로 적용범위를 지속 확대
- 연간 1,000억원 규모 나노·소재분야 R&D 과제에서 생산되는 연구데이터의 체계적 수집과 관리, 공유·활용을 지원하는 소재연구데이터플랫폼(HW, SW)의 구축 및 운영 계획 제시
- 2020년부터 2027년까지 총 7년간 소재연구데이터플랫폼(HW, SW)에 640억원 투자 계획 제시
- 소재연구데이터의 체계적 수집·분석·활용을 위하여 국가 소재연구데이터 플랫폼 전용 컴퓨팅(HW) 자원 구축·운영, 분야별 특화된 기계학습 모델을 통해 신소재 탐색·후보물질 발굴 등을 지원하는 SW플랫폼 개발·서비스, 효율적 활용·관리·운영을 위한 연구데이터 수집 서식 및 메타데이터 표준화, 관련 제도와 인력 양성 등 계획 제시

22) 「연구데이터 공유·활용전략(18.1.)」 을 통하여 제안된 국가연구데이터플랫폼의 구축·운영을 위하여 2019년 예비타당성조사가 이루어졌으나 해당 사업계획이 최종 승인되지 못함

○ 국가소재연구데이터플랫폼 구축 사업('20-'28)의 본격화

- 과학기술연구회 산하 4개의 정부출연연구소가 소재 혁신 연구 생태계 육성을 위해 국가 차원의 계산 기반 소재 연구데이터 플랫폼을 구축·운영하는 국책사업(전체사업기간 및 예산 : '20년 9월 ~ '28년 3월, 총 490억 원) 착수
- 2020년 9월 한국표준연구원을 중심으로 4개 연구기관(KISTI, KIST, 재료연 포함) 역량을 결집한 '국가소재연구데이터센터(사업단)' 출범
- 플랫폼(HW, SW) 구축 및 운영센터(KISTI), 소재연구데이터 수집·활용 관련해서는 에너지 환경소재 데이터센터(KIST), 스마트소재 데이터센터(KRISS,) 안전(구조)소재 데이터 센터(재료연)에서 추진

* (에너지·환경소재) 다공성 소재, 고체화합물 소재, 이차원 소재 등, (스마트·IT소재): 차세대 반도체 소재, 센서소재, 헬스케어용 소재 등, (안전구조소재): 극한환경 소재, 고기능성 금속소재, 수소인프라 소재 등

[그림 2-19] 소재 연구데이터 플랫폼 개념도

자료: 과기정통부 보도참고자료(2020.4.19.), "소재 연구데이터 플랫폼 구축방안 수립(소부장 특위)", 붙임3.

○ 소재연구데이터플랫폼 구축·운영 현황 및 향후 계획

– (DMP) 소재분야에서 2019년 이후 DMP 작성을 요청한 R&D 과제 연구책임자에게 소재연구데이터 플랫폼에 데이터의 공개 여부와 시기를 DMP에 명시하도록 하고, DMP 작성 예시를 제공하여 데이터 접근성을 증대할 계획임

* 연구데이터 관리를 위해 DMP를 적용하던 기존 대상(미래소재 디스커버리사업 등 36개 과제 과제, 350억원 규모)을 2020년 전체 나노·소재 R&D 사업 (84개 과제 약 895억원 규모)로 확대하고, 2021년에는 집단연구, 타부처 소재 R&D 사업 등으로 적용범위를 지속 확대

* 연간 1,000억원 규모 나노·소재분야 R&D 과제에서 생산되는 연구데이터의 체계적 수집과 관리, 공유·활용을 지원하는 소재연구데이터플랫폼(HW, SW)의 구축 및 운영 계획

– (표준화) DMP를 작성해도 데이터 포맷과 표준이 다르면 데이터 접근성이 저하되므로 표준화 연구를 선행하여 데이터 접근 및 활용성을 극대화하고 이종데이터 간 결합 가능성도 고려한 표준화 작업 진행 중

– 단순히 연구데이터만 모아 놓은 저장소가 아니라, 플랫폼 상에서 머신러닝, 딥러닝 등 다양한 분석기법을 탑재하고, 연구 관련 논문까지 연구자 입장에서 필요한 모든 자원(HW자원 포함)을 모아 제공함으로써 데이터 생산자 - 소비자 - 관리자 세 주체가 모두 참여하는 플랫폼 구축을 목표로 함

· 분야별로 특화된 기계학습 모델을 적용함으로써 소재탐색 설계 및 실험 검증을 지원하고, 유망 소재에 대해서는 향후 상용화 지원 후속연구도 지원할 예정

□ 바이오데이터스테이션(K-BDS) 구축사업

○ 2020년 5월 「제3차 국가생명연구자원 관리·활용 기본계획('20～'25)」을 통해 '데이터 기반의 바이오 연구환경 구축' 전략을 마련

○ 2020년 7월 비상경제중앙대책본부회의에서 관계부처 합동으로 마련한 바이오산업 혁신 대책으로서 「생명연구자원 빅데이터 구축 전략」 의결, 세부추진전략 제시

– 부처·사업·연구자별 흩어져 있는 바이오 연구 데이터를 통합 수집·제공하는 '국가 바이오 데이터 스테이션'(K-BDS, Korea Bio Data Station) 조성

○ 바이오데이터스테이션(K-BDS) 구축사업은 다부처 국가생명연구자원 선진화 사업의 일환으로 2022년부터 2026년까지 총 1,313억원(적정성 재검토 통과) 예산 규모로 추진 중임(주관부처 : 과기정통부, 참여부처 : 농진청, 질병청)

※ 산업부, 식약처 등 독립된 데이터 센터가 없는 부처는 K-BDS에서 통합하여 공유형 데이터 센터 형태로 운영할 계획

- 바이오연구분야 국가 R&D 사업에서 생산된 연구데이터의 체계적 수집과 관리, 활용을 위하여 한국생명공학연구원 국가생명연구자원정보센터(KOBIC)와 KISTI 슈퍼컴퓨팅본부가 상호 협력하여 바이오데이터스테이션(K-BDS) 구축 추진(현재 시범운영 중)

- 바이오 데이터 스테이션(K-BDS)과 부처 데이터 센터*의 협력을 통해 바이오 연구데이터를 표준 등록양식에 따라 수집하고, 품질 관리를 위한 SOP를 적용하며, 부처 데이터 센터에 등록된 연구데이터의 메타데이터(meta data)를 K-BDS로 연계하여 통합·활용을 위한 바이오 데이터 플랫폼을 구축할 계획임

  * 과기부(Kobic), 해수부(MAGIC), 질병청(CODA), 농진청(NABIC), 환경부(국립생물자원관) 등

[그림 2-20] 바이오 연구 데이터 확보관리 체계도

자료: KOBIC 내부 자료 인용

[그림 2-21] 국가 바이오 데이터 스테이션 운영 프로세스

자료: KOBIC(2022) K-BDS 홍보자료 인용

○ (수집대상) 국가연구개발사업을 통해 생산된 바이오 데이터(실험+임상 데이터)
- 연구를 위한 참조자료를 제외하고 실험, 관찰, 조사 및 분석을 통해 얻어진 사실 자료로서 연구결과 검증에 필수적인 데이터를 수집하고 있으며, 최종 연구성과물의 경우 원데이터와 메타데이터 모두 수집 대상

* 생화학분석, 이미지(영상), 임상 및 전임상, 유전체, 분자구조, 표현형 정보, 화학반응 등

○ (DMP) 바이오데이터스테이션에 국가연구개발사업을 통해 생산된 바이오데이터를 등록·활용하도록 하기 위하여 '바이오의료기술개발사업' 등을 중심으로 데이터관리계획(DMP)의 작성 및 갱신, 검토 절차 체계화
- 19년부터 '바이오의료기술개발사업'의 데이터관리계획(DMP) 작성을 고지하였으며 20년부터 동 사업에서 데이터 생산을 목적으로 하는 과제에 대해 연차계획서 제출시 DMP를 제출하도록 하고 '21년부터는 신규과제계획서 제출 시 DMP 제출을 요청
- DMP 작성 과제에 대해 선정-최종평가 등 평가단계별로 DMP를 점검하고, 연구수행 중 또는 연구 종료 후 데이터를 지정된 저장소(K-BDS 또는 부처 데이터센터)에 등록하고 데이터 공개·공유를 하도록 하고 있음
- (DMP 양식 마련) 바이오 분야에 적합한 데이터관리계획(DMP) 양식(안) 마련 및 보급

○ (표준화) 연구현장의 다양한 표준화 수요를 반영하여 기 확정된 '20년도 표준 등록 양식의 고도화 및 신규 바이오 연구데이터의 표준 등록양식 마련

- 데이터를 구성하는 4가지 요소인 ① 원 데이터(raw data), ② 실험(조건)정보, ③ 소재(샘플)정보, ④ 프로젝트(과제) 정보를 모듈화하고, 모듈별 등록 양식의 통일성 확보
- 15대 데이터 표준 등록 실무작업반 중 연구현장의 표준화 수요*에 기반하여 필요한 실무작업반을 선택적 운영하여 진행

* 바이오 연구데이터 표준양식 현장 수요조사(국내 바이오 학회/협회 대상)

○ (품질관리) 데이터 분야별* 품질관리를 통한 양질의 데이터 제공을 전문가그룹으로 구성된 '바이오 데이터 품질선도센터'를 운영하여 ① K-BDS 등록 데이터의 품질관리 및 큐레이션, ② 데이터의 품질, 분석 및 해석 관련 연구현장의 문제를 해결하기 위한 상담 지원, ③ 품질선도에 역량을 갖춘 연구인력 양성을 수행(22년 하반기~)

* (분야 1) 단백체, (분야 2) 대사체, (분야 3) 화합물, (분야 4) 바이오이미징 등 4개소 운영(바이오 데이터 활용 수요조사를 통해 지원 분야 선정, 유전체는 KOBIC이 담당)

- 등록된 데이터의 품질관리를 위한 품질관리 프로토콜 구현하여, K-BDS 플랫폼에 탑재하여 철저한 품질관리 수행(유전체 데이터 파이프라인 구축완료, 타 분야 지속 확대)

제3장. 연구데이터 구축·활용 현황 및 정책이슈

제1절 연구데이터 구축·활용 현황

1. 연구프로세스와 연구데이터

(1) 연구데이터의 정의와 속성

□ 연구데이터의 정의

○ 연구데이터는 **연구개발과제 수행과정에서 수집, 생산, 가공, 처리, 분석, 해석되는 데이터**로 정의할 수 있음

- 現 「국가연구개발정보처리기준」에서 연구데이터를 “연구개발과제 수행 중 산출된 사실자료 중 연구결과 검증에 필수적인 자료”로 정의하고 있지만, 이는 정부에서 주관하는 관리를 위한 규정상 정의이며 실제 연구현장에서 인식하고 생산·활용되는 연구데이터의 범위는 훨씬 광범위함
- 연구의 전 과정에서 생산되는 연구데이터는 연구결과뿐만 아니라 연구과정정보를 담고 있으며, 결과데이터뿐만 아니라 중간산출물 혹은 원산출물 성격의 데이터 포함

[그림 3-1] 연구데이터의 범위

자료: 신은정 외(2019a), p. 36 발췌 및 수정

□ 연구데이터와 공공·산업데이터의 관계

○ 연구를 위하여 연구자가 실험·관찰·측정·시뮬레이션 등을 통해 직접 생산하는 데이터 이외에도 이미 생성된 데이터를 재이용하거나 가공·연계·융합하는 가운데 새로운 연구결과를 도출하기도 하므로 연구데이터의 범위는 폭넓음

<표 3-1> 연구에 활용되는 데이터의 스펙트럼

구분	연구에 활용하는 주요 데이터 소스
연구팀 실험·조사자료	[바이오] 한국인1만명 WGS 데이터, 국립병원 및 연구원 제공 바이오의료데이터, 신약후보물질 물성 DB, 생리학/생화학 실험데이터 및 이미지 등 [소재]] 소재특성 관련 실험데이터, 시료분석 데이터, SNUMat 물성데이터 등 [환경/기후/천문] 수계/공기질 측정 데이터, 기후변화 시뮬레이션 데이터, 물환경정보시스템 측정망 데이터, 생태 탐사·관측데이터, KMTNet 관측데이터 등 [에너지/물리] 전력/에너지 데이터, KSTAR 데이터 등
공공데이터	[바이오] 건강보험공단(건강검진 DB), 심사평가원, 질병관리본부 오픈데이터 등 [환경] 환경부 물관리통합정보, Kwater 수질 모니터링 데이터, 국가상하수도통계 등 [안전/재난관리] 국가재난통계, 기상청 기후데이터/레이터자료, 산림청 산악기상관측망 자료, 공공시설물 안전점검/진단데이터 등 [에너지] 한전 전력망 DB
민간보유 데이터/산업데이터/기타	[바이오] 삼성서울병원 등 병원 자체 임상 데이터, EMR 데이터, 진단 이미지/영상, 게놈데이터, 바이오파운드리 생산 데이터, 상용 화합물 DB 등 [공학/안전] 산업체 공정 측정 데이터, 무인기 촬영 이미지 데이터, 사회기반시설물 안전점검/진단데이터 [융합] 학술자료 DB, Shutterstock 데이터, ImageNet 데이터, GPS 데이터, Kaggle, GitHub 데이터/소스코드 등

자료: 한국연구재단 기초연구본부 주관 데이터기반연구 현황 조사자료(2022.1.19.) 요약

○ 연구자가 직접 생산한 데이터 이외에도 국내외 학계에서 공개·활용되는 데이터, 공공부문 업무와 서비스를 위하여 제공되는 데이터, 산업·상업적으로 유통되는 데이터 등도 연구를 위하여 활용

○ 연구과정에서 생산된 연구데이터라 할지라도 산업적·상업적으로 활용되면서 산업데이터로서 취급·거래되거나, 일반 국민에게 공개·확산되면서 공공데이터로서 인식·활용되기도 함

[그림 3-2] 연구-공공-산업데이터 관계도

□ 연구데이터의 고유 속성

○ 연구과정에서 생산·활용되는 데이터와 다른 속성을 가진 데이터 간 연계·융합이 가속화되고 있지만, 연구데이터는 기본적으로 연구과정과 불가분의 관계를 가진다는 측면에서 고유한 특성을 가지고 있음

- 연구과정과 불가분의 관계에 있는 연구데이터의 속성이 관리·제도적 측면에서 연구과정을 수반하지 않는 공공데이터나 산업데이터와의 차별성을 만들어 냄
- 예를 들어, 연구데이터는 경제적 가치나 정치사회적 가치로 환원·평가되기 이전에 연구적 가치, 학술적 신뢰성을 중심으로 발전하고 검증되며 공유되는 특징이 있음
- 공개를 우선시하는 공공데이터나 엠바고 설정 후 사업화를 우선시하는 산업·상업데이터와 달리, 연구데이터는 연구개발과정의 불확실성에 따라 공개-비공개 여부, 공개 데이터의 범위와 방식에 대한 결정을 사전적으로 확정짓기 어려운 경우도 존재함

– 연구과정에 배태된 특성에 따라 연구데이터는 데이터 권리와 책임, 품질을 평가·검증하고 공동 활용하는 과정에서도 연구계의 공개검증체계, 성과인정체계, 출판체계 등의 영향을 강하게 받음

<표 3-2> 연구데이터와 공공·산업데이터의 비교

구분	연구데이터	공공데이터	산업데이터
일반적 정의	연구과정에서 생성·취득·관리·활용되는 자료이자 연구결과 검증에 필수적인 사실자료	공공기관이 법령 등에서 정하는 목적 혹은 직무상 생성·취득·관리하는 전자(기록)자료	제품 또는 서비스 개발·생산·유통·소비등 산업활동 전과정에서 생성 또는 활용되는 전자(전자적 처리가능) 자료·정보
공개와 보호 비중	데이터 공개와 보호의 균형 필요 공개를 통한 연구 활성화와 보호를 통한 사업화 트랙 공존	데이터 공개를 우선하되 필요시 보호 활용 제고를 위한 선제적 데이터 공개 장려	데이터 보호를 우선하되 필요시 공개 데이터 보호와 거래, 산업적 활용 동시 강조
IP 및 라이선스	저작권과 특허 등 지식재산체계 우선, 예외적으로 오픈라이선스(CC) 적용	공공저작물로서 정부라이선스나 오픈라이선스 부여, 자유로운 배포/복제/가공 허용	저작권과 특허 등 지식재산체계 우선, 이해관계자간 협상과 계약 우선
인센티브	학계의 인용, 출판 체계에 기반한 인센티브시스템 존재	연구자 친화적 인센티브보다 국가 차원에서 공공기관 평가 등에 가산점 부여	산업데이터 가치 기반 거래와 사업화를 통해 자율적 시장 인센티브 개발
소유와 권리 특성	연구를 위해 데이터를 생성 혹은 활용한 자에게 기본적인 권리를 부여하되 해당 연구 지원-수행주체의 특성과 연구과제협약내용에 따라 권리관계 변경가능	공공기관 직무성과이자 자료로서 공공기관 중심 권리 이전 및 자료 공개 일반화	산업데이터를 생성한 자에게 기본적인 권리를 부여하되 해당 데이터를 제공할 경우 발생한 약정이나 계약, 협약 등에 따라 권리관계 재설정
글로벌 데이터 연계	학계 인지도와 신뢰도, 활용도 제고를 위하여 글로벌 연계 필수, 단 경우에 따라 국외이전 제한 가능	국가정보로서 국민 대상 서비스가 일차적 목표, 글로벌 정보 연계는 부수적 서비스	국가의 산업경쟁력 제고를 위하여 글로벌 호환성과 연계성을 높이되 필요시 데이터의 국외이전 제한 가능

자료: 「공공데이터법」 및 「산업디지털전환법」 참고하여 저자 작성

(2) 연구과정으로서의 연구데이터관리

□ 연구데이터 프로세스

○ 연구데이터가 연구 전 과정에 걸쳐 생산·갱신될 수 있음을 고려할 때 연구의 라이프사이클과 같이 연구데이터의 라이프사이클도 연동·발전

- 연구탐색 및 계획 단계에서부터 연구를 수행하고 관련 결과를 산출·검증하며 추가적인 수정·보완을 해나가는 일련의 과정이 데이터계획-데이터수집-데이터분석·검증-데이터출판 혹은 재수정 등의 사이클과 연동됨

○ 연구과정에 따라 연구데이터의 규모와 형태, 처리방식 등이 변화하는 가운데 연구의 필요에 따라 특정 데이터를 등록, 보존, 출판, 공유하기도 하는 바 연구 전 과정에 대한 이해가 바탕이 되는 가운데 연구데이터의 접근 및 활용을 고도화할 수 있을 것임

- 연구가 진행되는 과정에서 수집·생성된 원 데이터가 지속 갱신되고 가공·분석·결합될 수 있으며 이 중에서 유의미하게 해석·재해석되는 데이터는 일부이며 최종 연구결과(출판물)에 반영된 데이터 또한 일부에 국한될 수 있음
- 특정 시점에 생성된 특정 형태의 연구데이터를 정확히 검색, 접근, 활용하기 위하여 연구과정에서 생성된 데이터 관련 메타데이터가 중요하며, 연구 전 과정 워크플로우에 대한 이해가 높을수록 연구데이터의 활용 효과가 높아질 수 있음

[그림 3-3] 연구데이터 프로세스

자료: (상단) 저자 작성, (하단) 추후 보완

□ 연구데이터 관리

○ 연구데이터의 생산, 관리, 공유, 활용을 촉진하고 데이터중심연구를 활성화하기 위하여 연구과정의 일부로서 연구데이터관리(RDM)가 이행될 필요가 있음
- 연구데이터관리가 연구 전 과정에 걸쳐 이루어지는 바 데이터관리계획(DMP)을 통해 연구기획단계에서부터 데이터관리 방침과 절차를 구상하고 준비할 수 있다는 장점이 있으나 DMP만으로 RDM을 갈음할 수는 없음
- 연구데이터관리는 연구과정의 일부로서 연구데이터를 취급·관리하는 것으로 데이터관리계획, 데이터의 가공 및 보존, 체계적 저장과 설명, 고유식별자의 부여 및 처리, 메타데이터나 알고리즘, 코드, 소프트웨어, 워크플로우 등의 제공 등을 포함함(OECD 2021, ERC Grant 지침 참고)
- 연구데이터관리는 기술적 품질관리나 데이터처리뿐만 아니라 민감정보 보호, 보안 등에 관한 관리 책임을 다하거나 라이선싱을 포함한 공유 절차와 규칙 관련 조치를 이행하는 활동을 포함함(ERC Grant 지침 참고)

2. 분야별 연구데이터 프로세스[23)]

(1) 입자물리학 사례(유럽핵입자물리연구소, CERN)

□ CERN의 대용량 실험데이터와 공동연구

○ 빅사이언스의 대표적인 사례인 입자물리실험 혹은 고에너지물리연구의 경우 가속기를 이용한 높은 에너지를 통해 특정 조건의 입자 충돌, 반응 등을 관찰, 분석
- CERN의 LHC(Large Hadron Collider) 가속기를 이용한 주요 실험[24)]으로서 CMS, ATLAS, LHCb, ALICE 등의 검출기 기반 실험이 있으며, 한국은 주로 CMS, ALICE 실험에 참여하고 있음

○ LHC 가속기를 활용한 한 번의 충돌 실험에서 발생한 데이터는 연간 약 수 페타바이트 규모의 원 데이터 생성[25)]

23) 분야별 전문가, 연구자 대상 면담조사(2022.2.17. - 2022.2.28.)를 바탕으로 작성
24) 에너지 필요수준이 높아져서 둘레가 27km짜리 가속기를 만들어서 스위스 제네바 도시 지역부터 프랑스 국경을 넘어서까지 가속기를 만들 어 놓고 그 링을 따라 지하 100m에서 가속을 시켜 특정 포인트에서 입자를 충돌시키면 튀어나오는 것을 관찰하는 실험
25) 연간 데이터 생산량 160PB (매일 0.2PB), 참고로 LHC upgrade를 통해 매일 4PB수준까지 요구될 수 있음 (자료:

– 데이터 규모의 방대함으로 모든 원 데이터를 저장하는 것이 아니라 의미있어 보이는 데이터를 선별 저장하는 과정을 하드웨어 및 소프트웨어적으로 구현, 선별된 원 데이터를 CERN 컴퓨팅 센터에 저장하고 이를 전 세계 Tier Grid 체계를[26] 통해 저장·분산 처리

○ 대규모 실험데이터를 공동 활용하는 연구커뮤니티기반 집단연구 제도화

– 전 세계 Tier Grid에 분산·저장된 데이터를 실험연구에 참여하는 글로벌 연구커뮤니티에서 공동 활용, 연구주제별 연구 그룹이 상시 운영되고 있으며[27] 논문 기획 아이디어부터 실험데이터의 분석 알고리즘, 소프트웨어 등을 커뮤니티 내에서 공개 검증하는 체계가 갖추어져 있음

– 실험데이터를 기반으로 한 연구커뮤니티의 집단연구가 일상화되어 있어 검출기 관리부터 데이터 정제·분석, 논문 작성 및 영문 교정까지 모두 공동 작업하여 저자 표기시 주 저자를 명시하지 않고 2-3천명에 달하는 저자를 알파벳 순으로 기재[28]

– 데이터 분석과 논문 작성 이외에도 모든 연구자가 가속기 운영과 검출기 관리 등의 연구기본환경을 유지·보수·업그레이드하는 서비스 활동에 의무적으로 참여

– 실험연구커뮤니티에 신규 가입하기 위해 연구커뮤니티 기여도에 대한 긍정적 평가를 받아야 하며 동료 추천과 심사절차를 거쳐 구성원의 투표에 따라 가입여부 결정됨

□ 실험데이터의 유형과 계층

○ 연구과정에 따라 연구데이터의 규모와 형태, 접근성이 차등화되어 있음

– 초기 생성되는 원 데이터가 분석·가공·정제되면서 다층적 데이터로 전환되고 연구결과를 보고·출판하는 데이터 또한 별도로 관리·공유됨

---

https://doi.org/10.1051/epjconf/202125104023) CHEP2021 conference, https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSOfflineComputingResults)

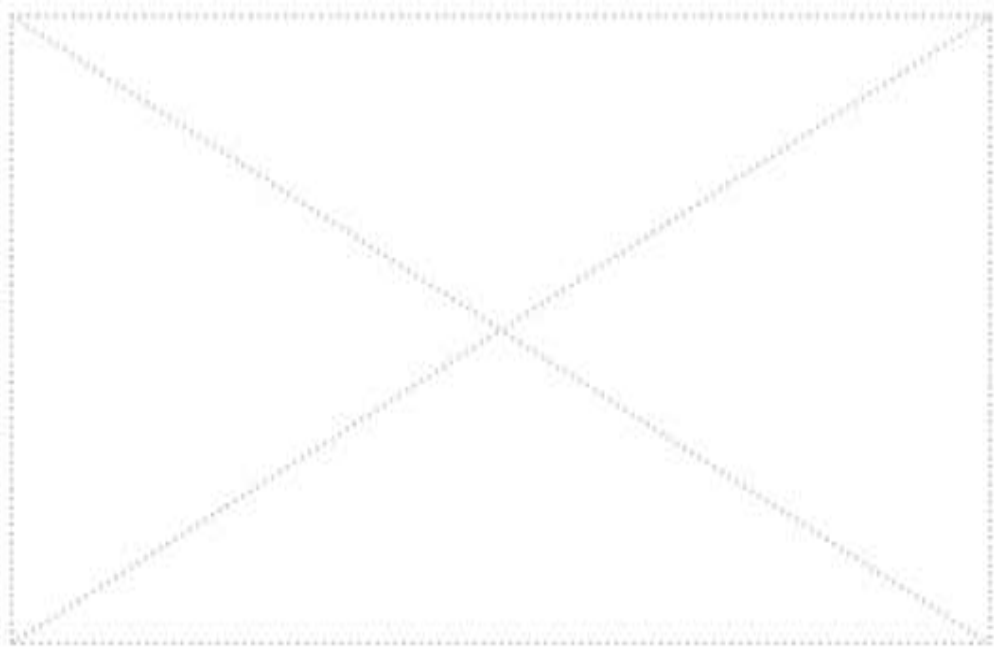
26) LHC에서 받은 데이터가 세계적으로 뻗어나가는 grid시스템. Tier 0~3까지 연결되어 있고 이용자 인증이 엄격하게 작동함. 0에서 3 Tier로 내려 갈수록 요구사항이 적어짐. LHC 데이터를 직접 받는 CERN 연구소가 Tier 0이고 각 대륙별 1개의 Tier 1 센터가 있어 Tier 0 데이터를 전송받음. 2022년 현재 KISTI의 GSDC는 ALICE 실험의 Tier1(아시아 지역) 기능을 담당하고 있으며, 국내 대형 연구소와 대학에서 Tier 2 센터 역할을 담당하고 있음. CMS 실험의 경우, KISTI가 Tier2 센터 역할을 맡고 있고 Tier3는 경북대, 서울시립대, KISTI 자체 센터 등이 맡고 있음

27) CMS실험의 경우 Physics Analysis Group은 Standard model, TOP, B physics, Higgs, SUSY, Exotica, Heavy ions, Beyond Two Generations 그룹이 운영 중이며, 대항 그룹 내 여러 연구주제별 모임이 존재함. 예를 들어 블랙홀 그룹은 Exotica 연구그룹의 연구 주제 중 하나임. Alice나 LHCb등의 실험에서도 다수의 연구그룹이 존재함. 다만, 연구그룹의 구성 및 운영 방식은 실험에 따라 다를 수 있음.

28) CMS, ATLAS실험 기준. Belle실험은 약 1000명 수준이고, 실험에 따라 100명 수준이나 더 적은 수인 경우도 있으며 실험에 따라서는 ABC순 표기를 하지 않는 실험도 있음. CMS와 ATLAS실험이 함께 공동으로 작성한 논문은 저자 수가 5천명 수준에 이름.

– 원 데이터는 CERN에서 직접 통제·관리하며 원 실험정보를 관련 정보와 결합·분석한 재구성 데이터는 연구커뮤니티 중심으로 공동 활용

[그림 3-4] 입자물리학 CMS 연구데이터 프로세스 사례



– 재구성 데이터의 일부는 일정기간의 엠바고 기간 이후 데이터셋 형태로 외부 연구자에게 공개하며, 대용량 실험데이터를 활용하기 용이하게 단순화한 데이터(skim data)를 교육·연구용으로 오픈데이터포털(CERN Open Data Portal)에서 제공

<표 3-3> 데이터 사용목적에 따른 CERN의 연구데이터 공개 수준 구분

	데이터 종류	데이터 접근지점	핵심 사용자
레벨 1	○ 출판/공개 데이터 (Published Results) – 그림, 표 등 출판된 논문 내 시각화 정보	○ 고에너지물리학 연구 데이터베이스 Inspire (http://inspirehep.net/) ○ HEP Data (https://hepdata.net/)	일반대중 (인터넷 공개 접근 가능)
레벨 2	○ 단순화한 데이터 (Simplified data formats) – 교육, 봉사활동 등 공적 목적으로 데이터를 활용할 수 있도록 간소화한 데이터를 제공	○ CERN Open Data Portal (http://opendata.cern.ch/)	봉사 및 교육 관련자 (예: 고등학교 교사 등)

	데이터 종류	데이터 접근지점	핵심 사용자
레벨 3	○ 재구성 데이터, 시뮬레이션데이터 등 (Reconstructed Data) – 연구에 활용 가능하도록 관련 분석 SW도구까지 함께 제공	○ CERN Open Data Portal (3~10년 엠바고 이후 일부 datasets만 공개) ○ The WLHC Grid (CERN 실험참여 연구자에 한해, 엠바고 없이 공개)	고에너지 물리학자 (high energy physicists)
레벨 4	○ 원 데이터, 분석가능한 모든 실험데이터 (Raw data) – 원 데이터 전체를 공개할 수는 없으나, 원 데이터의 주요 subset은 공개	○ CERN data centre (Tier 0)	강입자가속기 4개 실험에 참여하는 제한된 CERN 사용자 (데이터 처리 및 관리 전문가)

자료: Lipton(2020), p.12

□ 실험데이터와 소프트웨어

○ LHC 가속기 실험뿐만 아니라 입자물리연구커뮤니티 전반적으로 공통의 소프트웨어 라이브러리를 활용, CERN에서 만든 root 프로그램에 따라 데이터 형식은 .root으로 통일

– CMS 실험연구에서는 기본적으로 분석코드를 github 등 통해 공개

– 다만 실험데이터는 해당 실험에 참여하는 연구커뮤니티만 접근 가능하여 논문 게재 전까지는 데이터 접근이 제한됨

○ CERN Open Data Portal에서는 연구데이터뿐만 아니라 분석도구 등을 탑재·제공[29)]

– LHC에서 생성된 과거의 원데이터 뿐만 아니라 분석툴까지 탑재하여 연구데이터 활용성을 극대화

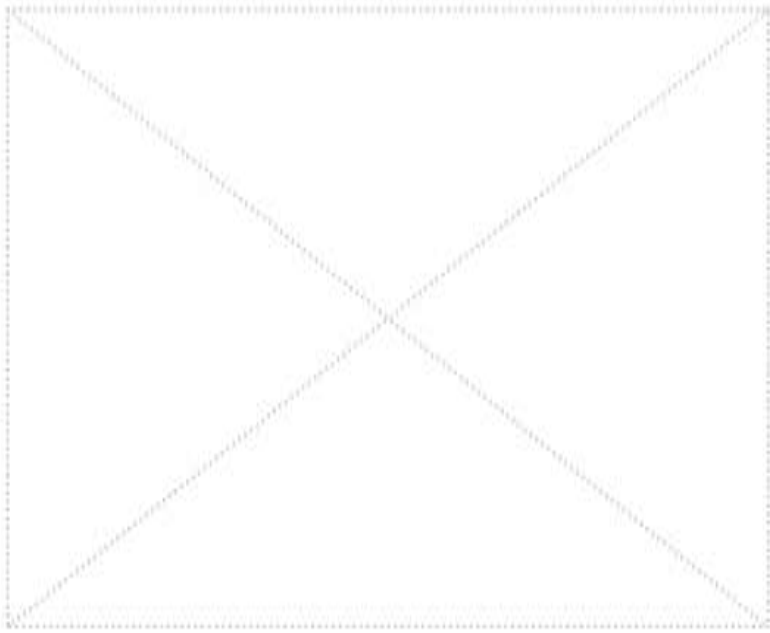
– 연구데이터 종류(충돌데이터, 파생데이터, 시뮬레이션데이터), 소프트웨어 (분석, 프레임워크, 도구, 검증, 워크플로우), 실험종류(ALICE, ATLAS, CMS, LHCb) 등에 따라 맞춤 연구데이터 및 분석환경을 제공

* 예를 들어, “ALICE Open Data”의 경우 소프트웨어를 설치하면 수동으로 별도 데이터를 다운로드 받을 필요 없이 함께 다운로드 됨[30)]

29) CERN 홈페이지, “CERN announces new open data policy in support of open science,” https://home.cern/news/press-release/knowledge-sharing/cern-announces-new-open-data-policy-support-open-science(2022.4.21.접속)

30) Getting Started with ALICE Open Data, https://opendata.cern.ch/docs/alice-getting-started#start(2022.4.21.접속)

[그림 3-5] CERN Open data portal과 데이터셋/분석 도구 탑재



자료: CERN Open data portal http://opendata.cern.ch/search?page=1&size=20&type=Software&type=Dataset(2022.4.26.접속)

[그림 3-6] CERN ALICE 인터페이스 및 software

자료: https://opendata.cern.ch/docs/alice-getting-started#start(2022.4.26.접속)

□ 출판데이터와 새로운 접근법

○ 실험데이터 기반 논문을 작성·출판한 경우, 출판논문에 활용된 데이터는 별도로 Inspire나 HEP Data를 통해 공개하여 논문 검증 및 후속 연구를 지원하고 있음
- 고에너지물리 연구에서 출판데이터의 공개는 최근 확대되어 표준으로 자리잡고 있음. HEP Data는 영국 Durham 대학에서 운영 중이며 출판 데이터는 HEP 데이터 형식으로 업로드되고, 각종 형식으로 자동 변환되어 다운로드 받을 수 있음

○ 논문으로 출판된 그림이나 그래프, 수치 등을 디지털 자료로 공개하여, 출판논문의 콘텐츠 자체를 데이터화하고 분석할 수 있도록 함
- 이론결과와 실험결과를 상호 비교할 수 있도록 분석 알고리즘을 소스코드 수준으로 구현하여 공유함
- 관련 라이브러리 코드들과 함께 실제 알고리즘을 반영한 소스코드를 같이 공개하여 별도로 알고리즘을 추정하거나 구현하는 과정없이 생성한 시뮬레이션 데이터에 대해 적용해 볼 수 있음

○ (향후 전망) 고에너지물리연구에서 핵심 시설인 입자가속기의 고도화·대형화 계속될 것으로 전망되는 가운데, CERN은 19년 미래원형충돌기(Future Circular Collider) 개념설계를 공개했고, 중국과학원 산하 고에너지물리연구소는 2030년 세계 최대 규모의 차세대 입자가속기, CEPC 구축을 목표로 준비 중이며, 일본도 ILC(International Linear Collider)를 구축 중인 바[31] 차세대 가속기 실험연구를 위한 국가 차원의 준비 필요

(2) 핵융합연구 사례

□ KSTAR 진단데이터·해석데이터

○ KSTAR라는 토카막 형태의 핵융합 연구설비를 이용하여 핵융합에너지 발전방식을 연구
- 핵융합반응은 보통 1억 도에서 가장 잘 이루어지는 바, 이 정도의 고온을 유지하면서 발전소로서 기능할 수 있도록 운전조건을 탐색하는 것이 연구의 주요 목적임

* 핵융합 장치는 핵융합 반응을 일으키는 초고온 플라즈마를 만들어서 장치에 가두는데, 이 플라즈마를 어떻게 만들지 - 공간적 밀도나 온도의 분포, 플라즈마 형상, 플라즈마 가열방식 등 -

31) https://www.dongascience.com/news.php?idx=21341

무수히 많은 파라미터 인자를 조합해서 최적의 성능을 낼 수 있는 핵융합 운전 조건을 찾음으로써 미래 핵융합 발전소를 위한 운전 시나리오를 고도화함

- 핵융합실험데이터들은 해석을 요하는'진단데이터'라 불리는데, 여러 SW 지식과 분석코드, 컴퓨터 시뮬레이션 등을 동원하여 진단데이터를 분석해야 그 의미가 드러남
- 핵융합설비에서 실험데이터, 장치운전 데이터와 같은 **진단데이터**가 생산되고, 이를 해석하기 위해 컴퓨터 모사를 통한 **시뮬레이션 데이터**를 생산
- (국내) 원자력연구원, 기계연구원, 재료연구원, 국내핵융합거점대학 등을 중심으로 KSTAR 연구데이터 활용 국내공동연구를 진행 중

○ 최적의 성능을 내는 KSTAR 운전조건을 찾기 위해 진단데이터를 분석하는 과정에서 분석 노하우와 소프트웨어를 가진 해외 국가·기관들과의 공조가 이루어짐

- 분석이 필요한 진단데이터는 KSTAR에서 생산하지만, 이를 분석하는 SW 기술력, 분석 및 소스코드 보유 정도는 국별 차이가 있는데 특히 미국과 유럽이 우위를 점하고 있음
- (서방권) 미국의 프린스턴 플라즈마 물리학연구소 (PPPL, Princeton Plasma Physics Lab), 미국의 General Atomics와 공식 MOU를 맺었고, 유럽, 영국과도 협력관계 유지 속에서 진단데이터를 분석
- (동북아) 일본, 중국과는 정기적인 연례 핵융합조정관회의(JCM, Joint Coordinator's Meeting)에 과기정통부가 참여하여 협력의제를 구성하고 공동연구를 진행하며, 이러한 국가간 협력의 토대 위에서 데이터 공유 및 분석이 이루어짐

□ 핵융합연구데이터 표준과 접근성

○ 핵융합연구의 출발점은 1960~70년대 미국이 주도했던 기초원천기술에 근거하고 있고 이것이 지금까지 발전해오고 있어 미국 중심의 데이터 구조, SW 포맷, 프로토콜이 주류임

- KSTAR 역시 미국과의 협력을 통해 만들어진 결과물로서 기술개발과 서비스 단계에서 미국 기준을 준용하고 있으며 KSTAR에서 저장하는 데이터 포맷 역시 미국에서 만들어진 규약을 따름

○ 다만, 현재 건설 중인 ITER는 국제 표준에 근거하고 있음

- 전 세계 7개국이 공동으로 진행하고 있는 국제열핵융합실험로(ITER)는 프랑스에 현재 건설 중에 있으며 2026년 완공을 목표로 하고 있음

- 2026년 가동을 시작하면 진단데이터를 생성할 것으로 기대되며 데이터의 분석은 각국이 가진 SW분석 역량에 따라 국별 폐쇄적으로 이루어질 가능성이 높음
- ITER는 건설에 필요한 모든 하드웨어 관련 특허는 공유하게 되어 있으나, 데이터 분석을 위한 소프트웨어 및 분석도구 등 분석환경 일체는 참여하는 각국이 준비하는 방식으로 운영될 예정

○ 핵융합장치에서 도출된 진단데이터 등은 그 자체가 주요한 공적 자산이자 보안 자료로 인식되어 국가·기관 간 자유롭게 공유하지 않음
- KSTAR 데이터를 기반으로 삼아 국내 혹은 해외 기관과 공동연구를 할 경우에도 보안 등의 요소를 점검해야 하며, 이러한 보호주의적 경향은 최근 더욱 심화됨

○ 최근에는 분석 SW와 같은 소프트웨어 개발 역량을 국가 차원에서 보호하는 경향이 뚜렷해짐에 따라 데이터뿐만 아니라 분석 SW 등에 대한 접근도 점차 폐쇄적으로 이루어지는 경향이 있음
- 미국 중심으로 발전한 일반적 데이터 구조, 저장 및 접근 프로토콜과 소프트웨어를 활용하면 타 장치 데이터를 해석·활용하는데 큰 무리가 없었으나 점차 각 장치마다 고유기술을 활용하여 별도로 생산·관리하는 데이터를 두고 해당 데이터는 자체 소프트웨어 없이 활용이 어려운 경구가 증가하면서 데이터뿐만 아니라 소프트웨어까지 전문적으로 다룰 수 있는 파트너, 전문가 수요가 능가하고 있음

[참고] 국제열핵융합실험로(ITER)와 한국형핵융합장치(KSTAR)

○ (ITER) 핵융합에너지의 대량생산 가능성을 실증하기 위해 전 세계 7개 국가가 공동으로 “국제열핵융합실험로 (ITER, International thermonuclear Experimental Reactor)”를 개발 중에 있음

- ITER 건설은 역사상 가장 큰 국제공동 연구개발사업으로서, 사업비만 약 79억 유로에 달함. 미국, 러시아, 유럽연합, 일본 4개국이 초기참여국으로서 1988년 ITER 설계에 착수했고, 한국은 2003년에 중국과 함께 참여. 2005년에는 인도가 합류하여 현재와 같이 총 7개 국가 체제가 됨.
- 참여 7개 국가가 각자 할당받은 장치 및 부품을 제작[32]하여 프랑스 남부 카다라쉬 현지에 현재 건설 중에 있으며 2025년 경 첫 실험데이터 생성을 목표로 하고 있음
- 전 세계 7개 국가가 ITER의 설계, 구축, 운영, 해체비용을 분담하고 각종 실험데이터를 비롯한 실험결과와 그로부터 창출되는 지식재산권 역시 공유하는 것에 합의했음[33]

[그림] ITER 건설 참여국가

자료: 한국핵융합에너지연구원 홈페이지, https://www.nfri.re.kr/kor/pageView/92(2022.4.22.접속)

○ (KSTAR) 한국형핵융합장치 “KSTAR(Korea Supercomputing Tokamak Advanced Research)” 사업에는 미국, 중국, 일본 등 6개국의 10개 연구소*가 참여하고 있으며, 국내에서는 POSTECH, KAIST, 서울대, 한국원자력연구원(KAERI) 등 KSTAR 공동연구거점센터 중심으로 15개 기관이 참여하여 연간 40TB 규모의 실험데이터를 생산하고 있음

– 국내외 연구자들이 KSTAR 클러스터에 접속하여 각 기관으로 개별적으로 복사하는 불편함을 해소하고자 KISTI에서 NFRI 차원의 보안성검토가 끝난 데이터를 KISTI에 보관하여 외부 접속사용자들의 접근 및 사용을 지원 중에 있음(KISTI, 2020)[34]

* 미국의 General Atomics, PPPL, ORNL, Columbia U., 중국 HUST, 일본 QSNIFS, 호주 ANU, 국제기구 ITER, 헝가리 WRCP 등

○ (데이터망) 테라바이트 규모의 대용량 데이터를 전 세계에 분포한 연구소들과 함께 분석하려면 안정적인 전송망 및 인프라 구축과 연구소 간 협력이 전제되어야 함.

– ITER, KSTAR와 같은 대형연구시설에서는 하루에만 테라바이트(TB) 규모의 대용량 실험데이터가 생성됨. 한국형 인공태양인 KSTAR 장치에서는 하루 생산되는 실험데이터가 1TB에 이르며, 국제핵융합실험로(ITER)의 경우 ‘26년 완공 시점에서 그간의 설계, 건설 등을 통해 누적된 데이터가 45 PB에 이를 것으로 예상됨.

– 이에 한국에서는 KISTI 주도로 미국 오크리지국립연구소(ORNL), 페르미국립가속기연구소(FNAL), 프린스턴플라즈마물리연구소(PPPL)과 같은 미국소재 핵융합 연구기관들과 실시간 데이터전송이 가능한 고속 데이터망 환경을 구축함[35]

32) 각국이 맡은 몸체, 부품 등을 위탁제조하여 프랑스에서 용접하여 합치는 방식이라, 국가 차원의 ITER 제작에 참여한 국내 업체는 현대중공업을 비롯하여 1,2차 협력업체까지 224곳에 달함. 부품을 각국이 제작하여 현지 조달해야 하므로, 각국의 긴밀한 공조 하에 국별 민간부문 기술력이 뒷받침되어야 성공 가능한 프로젝트임.

(3) 지질자원연구 기관사례

□ 지질자원연구의 넓은 데이터 스펙트럼

○ 국토지질, 해저지질, 지질환경, 기후변화, 지진/재해, 광물자원 등 다양한 대상을 연구하며 데이터를 획득하므로 연구데이터의 종류가 다양하고 스펙트럼이 넓은 편임
- 디지털 데이터부터 야외에서 채취하는 아날로그 형태의 시추코어, 암석샘플까지 다양
- 공간적으로 육지, 바다(해저)를 넘나들며 데이터 수집·분석
- 시계열적으로도 20세기 전후 작성되어 문화재적 가치를 지니는 조선지질도, 대한지질도부터 최근의 국토기본지질도, 탄전지질도, 해저지질도, 지질주제도 등을 수록한 최근의 웹 GIS 기반 지질도까지 취급

○ 데이터 스펙트럼이 넓은 만큼 데이터의 규모와 가치, 접근성 역시 편차가 큼
- 일반적인 지구관측 데이터는 특정 국가의 데이터만으로 연구가 어렵기 때문에 전세계적으로 공개·활용하는 사례가 많으며 온라인 상에서 쉽게 다운로드받아 활용 가능

* 지질학 국제 공동프로젝트인 'OneGeology'에서는 각국의 지질도를 하나의 온라인 포털에서 접속·활용할 수 있도록 하고 있음

** 한국지질자원연구원도 아시아지질자원위원회(CCOP, Coordinating Committee for Geoscience Programmes in East and Southeast Asia)와 데이터 관리 관련 공동프로젝트를 진행 중

- 반면 탐사선을 동원하여 해저에서 시추한 데이터는 대량의 자원과 시간이 투입된 희귀 자산으로 취급하며 상업적 활용도도 높아 쉽게 공개·공유하지 않음
- 데이터 규모 측면에서도 편차가 크며, 지진데이터*와 해저(석유)탐사데이터는 대표적인 대용량데이터임

* 지진데이터 등은 국가적으로 관리·활용되며 기상청 등 유관 기관과 고속망을 구축하여 데이터 공유

□ 기관 차원의 연구데이터관리(RDM) 지원시스템

○ 한국지질자원연구원 차원에서 생산되는 데이터의 체계적 관리 및 활용을 지원하기 위하여 2014년부터 데이터기반 연구플랫폼 구축을 위한 기획작업을 시작하여

33) 핵융합에너지연구원 홈페이지, https://www.nfri.re.kr/kor/pageView/92(2022.4.3.접속)
34) KISTI(2020), 대용량 실험데이터 활용 및 육성에 관한 법제적 지원 방안, 한국정보기술연구원.
35) 매일경제(2017.8.2.), "한국형 인공태양 실험 결과 미국 슈퍼컴퓨터로 본다." https://www.mk.co.kr/news/it/view/2017/08/517964/(2022.4.3.접속)

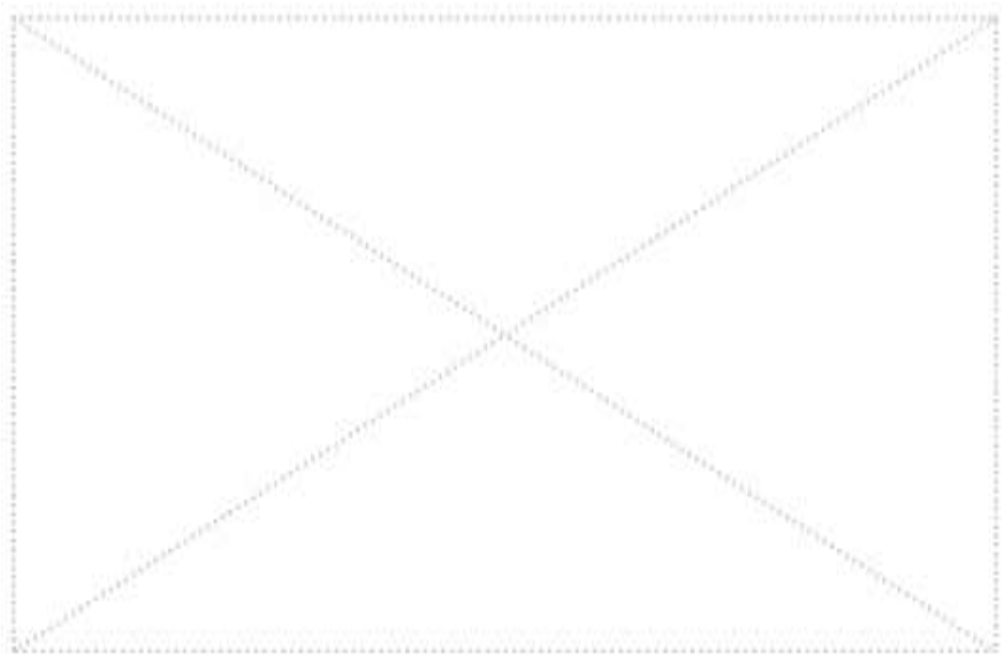
2015년 기관 내부용 지질자원 연구데이터 리포지터리, GDR(Geoscience Data Repository)을 구축함

- 한국지질자원연구원의 GDR은 KISTI에서 개발한 연구데이터 리포지터리(NaRDA)를 기본으로 삼아 기관 데이터서비스 수요에 맞게 변용한 것으로, 원내 연구자가 전산상으로 데이터관리계획의 기본 사항을 작성하고 연구과제에서 생산된 연구데이터를 등록, 검색, 열람할 수 있는 정보시스템임
- 자체 관리 시스템을 갖춘 지진데이터, 해저탐사데이터 등의 대용량 데이터를 제외한 연구원내 전체 연구데이터에 대해 GDR 활용 권고

○ 연구원 내 구성원은 GDR에 접속하여 연구데이터 생성 과제명을 선택하고 기관 차원에서 마련한 데이터 유형에 따라 일반적인 데이터 유형을 선택한 후 기본적인 메타데이터 정보를 입력할 수 있음

- 가능한 보편적으로 활용되는 포맷으로 데이터를 등록할 것을 권고하고 있으며, 그렇지 못하더라도 데이터 포맷과 활용 소프트웨어 정보 등을 제공하도록 권고하고 있음
- 또한 GDR에서 등록한 연구데이터에 대해 데이터 공개여부(원내공개, 원외공개, 비공개, 엠바고 기간 설정), CC 라이선스를 포함한 라이선스 설정 정보 등을 추가할 수 있음

[그림 3-7] 지질자원데이터 리포지터리 (GDR)



자료: 지질자원데이터센터, https://www.kigam.re.kr/menu.es?mid=a10405020000(2022.5.11.접속)

○ 연구원 차원의 체계적 데이터 관리 및 활용을 위하여 GDR 구축·운영 이외 제도적·조직적 정비가 이루어짐

- 2017년부터 연구원의 기본사업에 대해 데이터관리계획(DMP)을 시범 적용하고 2019년 기관 차원의 규정을 마련하여 전체 기본사업에 대해 DMP 적용(기관 기본사업 외 국가연구개발사업이나 타 부처 수탁과제에 대해서는 DMP 미적용)
- 기관 차원에서 과제계획단계에서 DMP를 함께 요청하며 과제심사단계의 부속서류로서 한 장짜리 DMP 검토, 과제수행 및 최종보고 단계에서 DMP 갱신 가능
- 기관 단위 GDR의 구축·운영을 촉진하기 위하여 '지질자원데이터센터'를 운영 중이며, 시설사업으로 별도의 데이터센터 건물 건축 중임
- 연구데이터의 공개 여부/방식은 연구책임자 결정을 존중하며 원내 연구데이터관리위원회가 존재하여 주요한 기관차원의 의사결정, 연구데이터 공개 전 실무차원에서 기본적인 사항을 사전 점검함

□ 지질자원데이터의 표준과 프로토콜

○ 지질자원 데이터 스펙트럼이 넓어서 분야별로도 용어, 분류, 프로토콜에 편차가 있음

- 지질도의 경우 국제적으로 이미 표준화되어 있고 데이터 인용 체계 정착

  * 미국의 지질조사국(USGS, US Geological Survey)은 각 주 마다 지질조사소가 있어 연방정부 차원에서 데이터, 용어 등을 표준화하는 조직을 별도로 운영. 영국의 지질조사소(BGS, British Geology Survey)와 호주 등도 데이터 관리가 체계적으로 이루어지는 사례

- 지화학데이터의 경우 국제학회 등에서 표준을 만들고 있는 상황이지만, 진행형이며 연구현장에서 해당 동향을 체감하지 못하는 경우 존재
- 분야에 따라 상용화된 소프트웨어를 활용하기도 하고 자체 제작한 코드와 포맷을 활용하는 경우도 있어 보편적인 데이터 포맷과 양식을 규정하기 어려움
- 한국지질자원연구원 차원에서 데이터 생산방식, 포맷, 활용 관련 고려사항 등을 담은 분야별 연구데이터 표준 가이드라인을 준비 중임

□ 지질자원데이터의 공개·출판·활용

○ 연구원의 지질자원데이터 외부 공개를 위하여 오픈데이터플랫폼, Geo Big Data Open Platform 구축·운영[36]

- 지질도에 대한 활용 수요가 높은 편이며, 연구자뿐만 아니라 매장문화재 발굴·조사나 하천정비계획을 담당하는 지자체 관계자, 역사와 지리 등을 교육하는 교사와

36) 지오 빅데이터 오픈 플랫폼, https://data.kigam.re.kr/#(2022.5.11.접속)

교육기관 등 다양한 그룹에서 공개되는 지질자원데이터 활용
- 외부 공개 시 디지털 고유식별자*를 등록하여 제공
* 국제지질시료식별번호(IGSN), 디지털매체식별자(DOI) 부여

○ 국내 유관기관과 협력하여 2019년 데이터저널로서‘Geo Data’창간[37)]
- 한국지질자원연구원을 비롯하여 국립생태원, 한국해양과학기술원, 극지연구원, 항공우주연구원 등이 협력하여 2019년 데이터저널‘Geo Data’창간
- 논문의 부속 자료가 아닌 연구데이터 자체를 출판하는 모델로서 22년 현재 연 4회 발간

(4) 소재연구 기관사례

□ 기관의 디지털 전환(DT) 전략과 연구데이터 DT

○ 한국과학기술연구원(KIST) 차원의 디지털 전환(Digital Transformation, DT) 논의가 최근 가속화되는 가운데 기관장 중심의 DT 전략 마련
- 크게 세 가지 – (1) Lab DT (스마트랩, 실험실 무인화 준비), (2) 연구지원 DT (연구행정 및 허가절차, 장비 및 서버 관리절차 전산화/간소화/자동화), (3) 연구데이터 DT 차원에서 추진 중임
- 기관 차원의 DT를 추진하기 위하여 기존 분산 운영되던 연구자원·연구데이터조직을 모아 연구자원·데이터지원본부 신설
- 연구현장의 실질적인 DT를 위하여 “연구자 행동 중심”지원 시스템 개발 지향

○ 연구데이터 DT는 연구과정 중 수집되는 데이터와 정보, 워크플로우 등을 체계적으로 수집, 관리, 공유, 활용하기 위한 조치로서 소재연구(에너지환경소재) 분야부터 적용 중임
① 비정형 데이터 포함 전자연구노트의 집단적 관리·활용 및 표준화된 워크플로우 템플렛 개발·공유·활용
② 주요 실험장비·시설에서 생성, 측정, 분석된 정형 데이터를 자동 수집, 처리하여 활용할 수 있는 절차와 시스템 개발·설치, 원내 보유한 계산 결과 DB와 외부 계산 DB (물성값)를 연동·활용 기반 마련(내부 구성원 우선 활용), 논문 DB, 학술 DB(Springer Nature, Elsevier, Royal Society of Chemistry, ACS Publications, Wiley, 그 외 OA 저널)에서 분야별 논문 자동수집 및 논문으로부터 데이터 추출(자연어 처리), 연계, 활용 계획

37) 지오데이터 홈페이지, https://geodata.kr/(2022.4.11.접속)

○ **(워크플로우-전자연구노트)** 연구자 행동기반 전자연구노트를 개발하여 연구과정이 자연스럽게 연구데이터로 수집될 수 있도록 '전자연구노트(KiRI)' 시스템을 구축·운영

- 수기로 작성하여 전자 시스템으로 옮기던 추가 부담을 줄이고* 연구노트 작성과 연구데이터 축적이 별개로 이루어지는 것이 아니라 연구노트를 작성하면서 동시에 연구자의 추가적인 노력 없이 연구데이터가 수집되고 메타데이터가 연결될 수 있도록 워크플로우 관점의 전자연구노트시스템 개발·활용 중

· 실험실에서 수기로 작성한 기록 pdf를 저장하면 머신러닝을 이용하여 해당 기록을 텍스트자료로 전산화 지원, 작성시점인증이 가능해서 법적효력을 갖는 연구노트로 인정

- 반복되는 과제수행체계와 워크플로우를 표준 템플릿화하여 기관 차원에서 공유, 활용할 수 있도록 함으로써 연구과업 수행과정에 반영된 암묵지와 노하우 계승·갱신

○ **(소재데이터-플랫폼)** 소재연구데이터플랫폼을 구축하여 실험, 컴퓨터 시뮬레이션, 문헌 등에서 생산되는 데이터를 수집하고 머신러닝 적용을 통한 데이터 추가 생성·활용

- 원내 연구장비를 통해 생성한 실험데이터와 슈퍼컴퓨팅 시뮬레이션에서 나온 가상실험데이터 등을 공유, 연계, 활용할 수 있는 기반 확대
- 논문 및 웹저널 등 문헌에서 자연어 처리 등으로 얻을 수 있는 데이터도 추출하여 연계, 활용할 수 있도록 지원

[그림 3-8] KIST 소재데이터플랫폼 구축을 통한 실험데이터 수집

자료: KIST 홈페이지, https://csrc.kist.re.kr/bbs/board.php?bo_table=m01_03&wr_id=11(2022.5.12.접속)

(5) 바이오헬스연구 사례

□ 암 다중오믹스연구

○ 다중오믹스 데이터 기반 암연구

- 암 다중오믹스연구는 암과 연관된 유전/후성유전 변이들을 집중 분석하는 연구로서 DNA 서열정보인 게놈(유전체)데이터[38] 이외에도 RNA 시퀀싱 데이터 등 다양한 OMICS 데이터를 포괄적으로 활용
- 컴퓨팅 자원과 기술이 발전하면서 점차 동일 검체의 유전체 데이터뿐만 아니라 이미지 데이터 등 이종데이터를 연계·분석하거나 대규모 DB와 연계·검증하는 추세 확대
- 대형실험장비에서 나온 대용량 데이터를 연구단계에 따라 분산 처리하는 거대과학과 달리, 암유전체연구는 환자로부터 얻는 오믹스 데이터를 모으고 연계하여 활용하는 분산적 데이터 프로세스로 이루어짐

○ 연구를 위한 암 다중오믹스 데이터를 확보하기까지 환자 샘플의 확보, 시퀀싱을 통한 유전체데이터로의 변환, 비식별화를 통한 개인정보 삭제 등 여러 단계를 거쳐야 함

- 국내 암환자에게서 샘플을 채취해야 하므로 대학 연구자는 병원 내 임상전문가들과 공동연구를 통해 필요한 다중오믹스 데이터를 생산·확보할 수 있음
- 인체유래물 연구이므로 IRB 심의를 거치고 환자들의 동의·서명을 득한 후 샘플을 채취하고 익명화·비식별화한 후 시퀀싱 회사 등을 통해 (유전체/전사체/후성유전체의) 다중오믹스 데이터를 추출하면 해당 데이터만 전달받아 분석
- 분석결과에 대해 병원 임상팀과 협의하여 임상증상과 다중오믹스 분석결과 비교·대조분석하는 형태로 공동연구 진행

○ 게놈데이터는 여러 개의 레이어가 있으며 이런 데이터 층위 간 표준화까지는 아니더라도 분석 방법이나 방향에 관해 학계에서 공통적으로 인식하는 표준 프로토콜이 어느 정도는 마련되어 있음

- 시퀀싱 프로토콜이 대체로 표준화되면서* 대체로 동일한 데이터 포맷을 이용하여 게놈데이터 간 호환 및 연동이 수월한 편

* 시퀀싱 기술이 사실상 반독점 상태로 시퀀싱 설비의 80-90%는 일루미나에서 제공하여, 샘플링하는 준비과정은 다양할 수 있으나 시퀀싱 프로토콜을 대체로 표준화된 상황임

· 유전체 분야는 미국 국립생물공학정보센터 (NCBI), 유럽 생물정보학연구소(EBI)를 중심으로

38) 유전체는 생물이 가지고 있는 유전정보로서 A-T-G-C로 구성된 서열정보. 시퀀싱 기계로 유전체 서열 정보를 읽어내게 되는데 이를 게놈데이터라 함

국제표준이 비교적 잘 정립되어 있음

- 그럼에도 불구하고 표준이 정해진 것은 아니라 연구자와 연구 특성에 따라 데이터 해석에 있어 다양한 선택지가 있음
- 시퀀싱 기술이 사실상 표준화되어 있으나 연구 고도화를 위한 이종데이터와의 연계·결합·분석은 쉽지 않음[39)]

○ 데이터와 함께 분석에 활용할 알고리즘이나 소프트웨어를 보편적으로 공유하고 있으며, 분석 방식이 독창적이거나 새로운 경우 데이터와 함께 분석 알고리즘까지 공개 검증 후 이용하는 것이 일반적임

- 특정 분석을 할 때는 어떤 프로그램/알고리즘을 쓸지 다수가 쓰는 것들이 몇 가지 있기는 하지만 분석 방식에 따라 직접 연구자가 만들어서 적용하기도 함
- 이 경우, 연구자가 데이터 분석·해석 과정에서 어떤 알고리즘을 썼는지에 관한 논문을 먼저 써서 공개 검증을 우선적으로 받는 것이 일반적이며, 검증받은 툴이나 알고리즘으로 데이터를 해석하게 됨

○ 자체 수집·분석한 데이터를 공개된 대규모 DB와 비교·검증하거나 공개된 해외 데이터를 활용한 연구도 진행함

- 데이터 검증과 논문출판 단계에서 미국, 유럽 등지 데이터 리포지터리 주로 활용
- 대표적 공개 DB로서 미국의 대규모 암유전체 프로젝트, The Cancer Genome Atlas (TCGA)[40)]가 있음

[참고] 암유전체 아틀라스(TCGA)와 GDC 유전체 데이터 포털

○ (TCGA 사업개요) 암 유전체 아틀라스 프로젝트(The Cancer Genomics Atlas)는 2006년 미 국립보건원 산하 암연구소(NCI)와 인간유전체연구소(NHGRI)가 암에 관한 유전체 변이를 파악하기 위해 공동으로 진행한 대규모 사업

- 20,000여개의 주요 암에 대한 분자적 특성을 파악하여 2.5페타바이트가 넘는 데이터를 생성하였고, 암에 대한 유전적 이해, 진단, 치료기술 개발에 활용될 수 있도록 모두 공개하고 있음
- 2008년 데이터 분석툴을 탑재한 데이터 포털을 개시하였고, 2012년에는 데이터

39) 이종데이터 결합시 개인 식별 수준에 따른 데이터 공개 정도가 상이하고, 동일 검체에 대한 이미지 데이터와 유전체 데이터 등을 결합·분석하고자 할 때에도 데이터 포맷과 분류, 데이터 처리 프로토콜과 식별자 관리 등의 편차로 데이터 연계·결합 분석이 쉽지 않음

40) https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

리포지터리 “Cancer Genomics Hub (CGHub)” 구축[41]

○ (GDC 데이터 포털) TCGA에서 구축한 데이터에 암 관련 타 사업에서 축적한 유전체 데이터들을 통합하여 “유전에 데이터 통합포털(GDC, the Genomics Data Commons Data Portal)*”을 구축

- GDC 데이터 포털에서는 데이터 종류(category), 형태(type), 저장 형식(format), 공개여부(controlled/open) 등 검색조건을 세분화하여 연구자 수요에 맞게 검색 및 다운로드가 가능함
- 데이터 시각화 도구, 분석 도구** 등을 탑재하고 있어, 페타바이트에 달하는 데이터를 연구자가 모두 다운로드 받는 부담 없이 웹 페이지 상에서 여러 분석이 가능함
- 전 세계적으로 진행 중인 유전체 관련 프로젝트 현황을 한 눈에 파악할 수 있고, 개인 연구자가 구축한 유전체 관련 데이터를 GDC 데이터베이스와 비교·검증하고 쉽게 저장·업로드할 수 있음

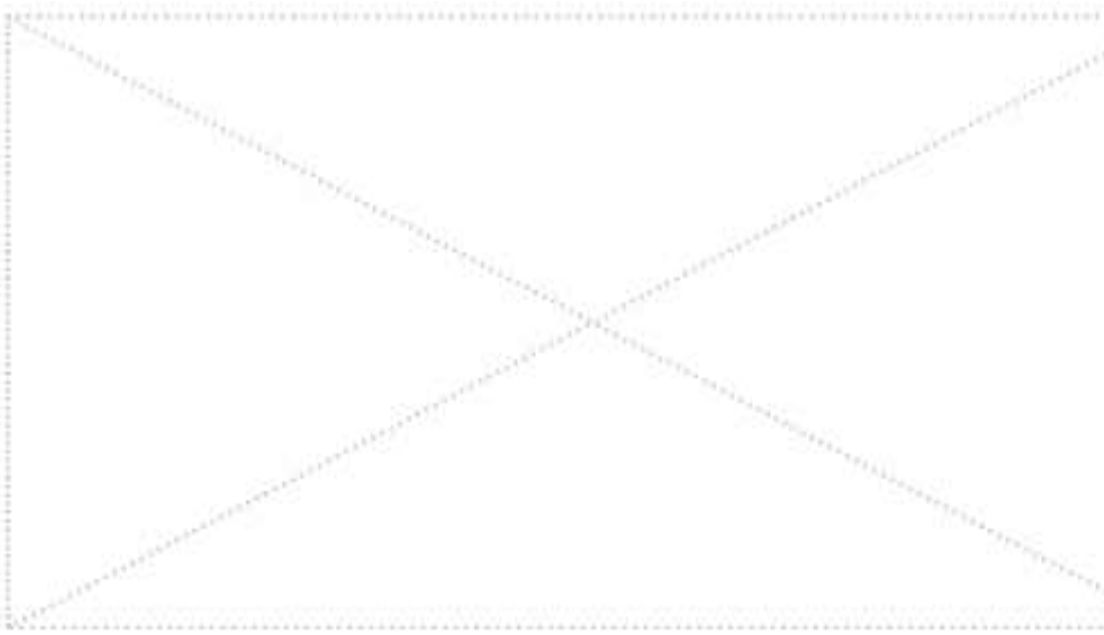
* Genomic Data Commons Data Portal, https://portal.gdc.cancer.gov/

** TCGA Compuational tools (분석도구) 별로 개별 포털웹사이트를 운영 중[42] – The Cancer Imaging Archive(TCIA), The Cancer Proteome Atlas Portal (TCPA), cBioPortal for Cancer Genomics 등

[그림] NCI의 유전체 데이터 통합 포털 홈페이지

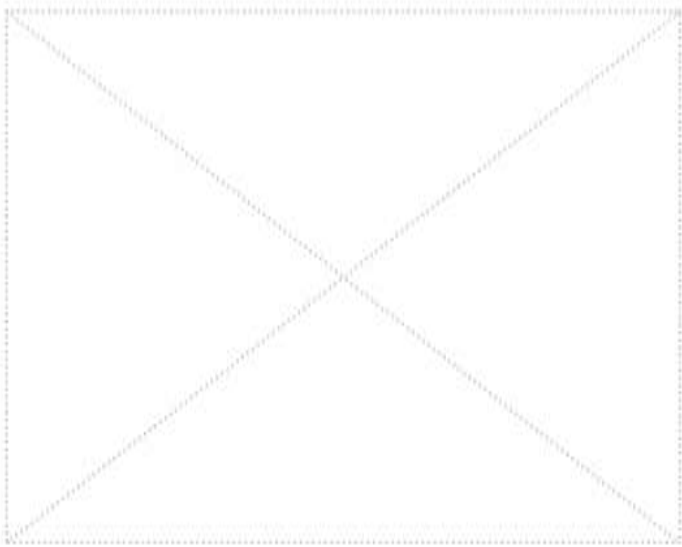
41) NCI TCGA 사업 홈페이지, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/t

[그림] NCI의 유전체 데이터 통합 포털의 Data repository



자료: GDC Data Portal 홈페이지 https://portal.gdc.cancer.gov/(2022.4.2.접속)

[그림] 암유전체 아틀라스 데이터 현황



자료: NCI TCGA 홈페이지,
https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history(2022.4.2.접속)

○ TCGA 공동연구팀은 구축한 데이터를 토대로 팬캔서아틀라스(암지도아틀라스, Pan Cancer Atlas) 프로젝트를 추진하여 사람들이 가장 많이 걸리는 33개 암의 DNA와 유전체를 분석했고, 그 성과들을 Cell과 같은 유명 저널에 게재하는 성과를 이룸[43]

○ 이러한 가시적 성과 이외에도 TCGA 프로젝트와 GDC 데이터 포털은 암 유전체 분야의 연구혁신을 주도하고 있다는 점에서 의미있음(Hutter and Zenklusen, 2018)[44]

- 첫째, 공동연구(Team science) 의 보상체계로서 새로운 가능성을 보여줌. TCGA 프로젝트에는 각기 다른 다양한 전문성을 가진 연구자들이 대규모 팀을 구축하여 성과를 냈는데, 일반적으로 논문의 1저자가 학계의 인정을 받는 것과 달리 TCGA 프로젝트 팀/컨소시움 형태로 연구 성과를 게재하여, 신진 연구자들이 암유전체 학계의 인정을 받을 수 있는 발판이 됨. 이는 연구자 개인이 아니라 과학자들이 공동연구로 노력을 결집할 수 있는 보상체계로서 가능성을 보여주는 사례임.
- 둘째, 데이터셋 구축 단계에서부터 알고리즘 개발자들이 분석 툴 개발에 참여하여 생물정보학(computational biology)과 같은 새로운 학문 분야의 성장과 미래 유전체 연구의 데이터 자원을 확보하게 됨. TCGA 프로젝트에서 개발한 도구들이 유전체 분석의 표준으로 자리하게 되어 데이터 사이언스와 computational biology 분야의 혁신을 가져왔고, 모든 TCGA 프로토콜이 공개되어 있어 메타데이터 구축이 가능해짐

□ (일반인코호트데이터 활용) 웰케어·웰니스연구

○ 바이오헬스분야에서는 건강한 사람을 대상으로 신체측정 데이터, 임상데이터, 유전체 시퀀싱 데이터, 디지털 헬스케어 기기정보, 라이프로그 데이터, 설문조사 및 문진 데이터 등 한 개인(검체)에 대한 다양한 데이터를 종합 분석하여 맞춤형 건강관리법을 모색함

- 웰케어·웰니스연구는 병원을 중심으로 환자 동의에 따라 수집하는 의료데이터와

cga(2022.4.2.접속)

42) NCI TCGA Compuatational Tools 목록 https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/tools(2022.4.2.접속)

43) Cell 2018년 4월호에 암게놈아틀라스 공동연구팀의 팬캔서아틀라스 성과를 27개나 발표함

44) Hutter and Zenlusen(2018), "The Cancer Genome Atlas: Creating Lasting Value beyond Its Data," Commentary, Cell 173, 283-285.

이미지/영상, 연구진을 중심으로 검체에 대한 실험과 분석을 통해 생성되는 분석 데이터, 시퀀싱 업체에서 제공하는 오믹스데이터를 종합적으로 활용함

- 기본적으로 임상 중심 병원과 연구 담당 기관/인력이 함께 협업하는 구조의 연구팀이 구성되어 웰케어·웰니스연구 진행(PI는 병원에서 맡는 경우가 다수)

○ 개인 맞춤형 건강관리를 증진시키기 위한 연구인만큼, 연구데이터 수집 및 처리, 활용 과정에서 개인정보보호를 위한 조치가 필수적임

- 바이오헬스분야에서는 개인정보를 제거하는 비식별화 과정이 이루어져야 하므로 국가공인인증과 같은 제3자 기관을 통한 비식별화 과정이 제도적으로 정착
- 병원 임상의와 연구담당 대학 연구자가 연구팀에 함께 참여하더라도 각각 구축한 데이터는 직접 공유되지 않고 제3의 기관을 통한 비식별화 조치를 통해 결합된 후 연구목적에 따라 부분적으로 공유·활용되는 구조임
- 각각의 이종데이터별로 수집·활용 목적이 상이한 경우 이종 데이터 결합을 하더라도 동의서에 기재된 내용 이외의 목적(예: 상업적 활용)으로 활용하기 어렵기 때문에 데이터를 구축해도 실제 활용으로 이어지지 못하는 경우 발생

○ 기술적으로도 이종데이터 간 포맷 및 표준이 상이하여 연계·결합하여 분석하기 어려움

- 유전체 데이터, 영상 등 임상데이터, 라이프로그 데이터 등 동일 검체에 대한 다양한 이종데이터가 존재하지만 실제 이종데이터간 포맷, 표준 등이 모두 상이하므로 연계가 쉽지 않음
- 병원이나 연구기관 자료 이외에도 국민건강보험공단, 건강보험심사평가원 등 공공부문이 보유한 데이터(국가건강검진데이터, 약물처방데이터 등)나 안산/안성코호트 데이터, 국민건강영양조사, 국가암빅데이터, 보건의료빅데이터, 울산 만명게놈프로젝트 등의 데이터와 결합하려는 노력이 이어지고 있지만 기술적·제도적 장애로 인하여 실제 데이터를 결합·분석하기 어려움

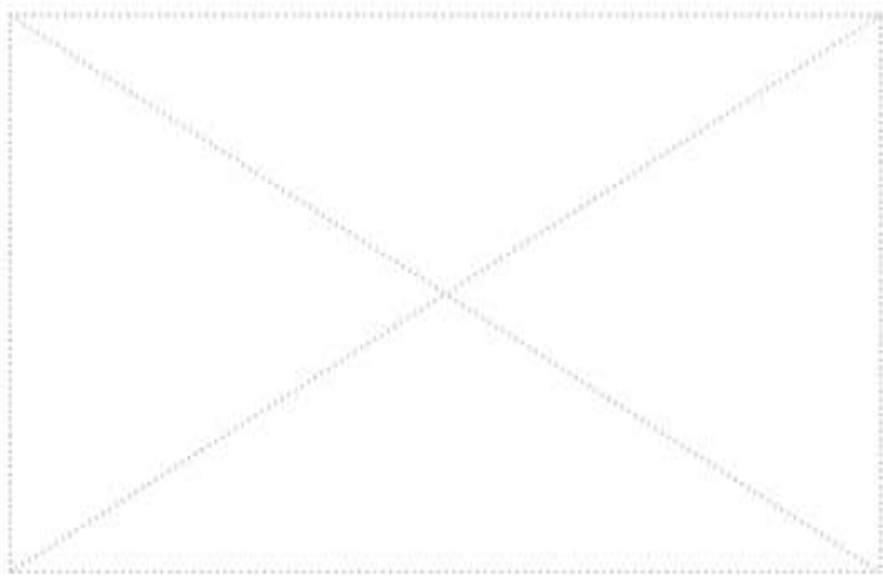
[참고] 미국 All of Us 연구 프로그램

○ 미국 All of Us 연구 프로그램(All of Us research program, 이하 All of Us)은 데이터 가시성이 높은 시스템 설계를 통해 데이터 제공 서비스 고도화

- All of Us는 개인의 생활습관, 환경 및 생물학적 구성이 건강과 질병에 어떻게 영향을 주는지 탐구하고 정밀의학을 발전시키기 위해 미국 국립보건원(NIH) 주도로 2018년부터 시작된 데이터 구축 프로그램

– 2018년부터 최소 10년간 100만 명의 지표 변화를 추적하여 데이터베이스를 구축하고 지원자들의 동의하에 유전자, 인종, 성별, 진료 기록, 직업, 생활습관 등의 정보를 수집·분석하는 것으로 2022년 4월 기준으로 48만 명 이상이 참여

[그림] 미국 All of Us, 데이터 제공자 현황



자료: All of Us research program 홈페이지 (접속일 : 2022.04.08.)

– All of Us의 이용자는 3가지 즉, ① 요약된 기술통계 데이터를 볼 수 있는 '공공 계층(public tier)', ② 설문조사, 의료정보(EHRs), 생체 지표 등을 비식별화된(de-identified) 형태로 제공하는 '등록 계층(registered tier), ③ 통제된 형태로 목적에 맞게 데이터가 제공되는 '통제 계층(controlled tier)'으로 구분
– All of Us는 '공공 계층(public tier)'에게 데이터 획득 전에 플랫폼 내에서 쉽게 데이터 종류, 표본 수, 수집 경로 등을 확인할 수 있도록 '데이터 브라우저(Data Browser)' 서비스를 제공(NIH, 2020.11.12)

[그림] 미국 All of Us의 데이터 브라우저(Data Browser) 일부

자료: 정일영 외(2022), p.29

- All of Us 내 '등록 계층(registered tier)'*에게는 데이터에 접근하여 분석할 수 있는 클라우드 기반 플랫폼인 연구자 작업대(Researcher Workbench) 활용 가능

* 2021년 기준으로 1,000명 이상의 연구원이 등록하였고 660개 이상의 연구 프로젝트가 시작

[그림] 미국 All of Us, 등록된 연구자가 사용할 수 있는 'Researcher Workbench' 분석플랫폼

자료: 정일영 외(2021), p.43

(6) 데이터과학·머신러닝기반 융합연구 사례

○ 문제해결형 알고리즘 개발 및 검증, 갱신을 위해 다양한 데이터소스 활용

– 학계 및 공공부문에서 공개적으로 접근·활용가능한 데이터*를 기본적으로 활용하지만, 데이터 양이나 제공되는 부가 정보 등이 제한적인 편임

* 공공부문에서 생산한 교통·관광 데이터 접근성이 높은 편이며 보건의료데이터 접근성은 제한적

** AI 데이터 허브에서 제공하는 데이터의 규모나 품질도 아직은 제한적임

– 제조공장, 서비스/마케팅 부문 데이터는 구매하거나 데이터거래소를 통해 거래*

* 국내 거래 중인 데이터 가격은 대략 데이터 생성·가공에 투입된 인력 중심으로 특급/중급 등의 구분이 형성되고 있는 편, 다만 콘텐츠 개발자 인건비 산정의 편차는 매우 큰 편이며 가치산정 기준에 대한 합의도 아직은 부족한 편임

** 데이터거래소에서 필요한 데이터 구매를 지원하는 데이터 바우처 사업 진행 중

– 데이터 수집·관리·보존을 위하여 클라우드 서비스 주로 이용

○ 분야·부문별 문제해결을 위한 알고리즘 개발 수요가 주를 이루며, 데이터 품질 고도화를 위한 추가 관리 필요

– 특정 수요가 존재하는 분야·부문부터 데이터 품질 고도화 및 표준화 필요, 이종데이터 연계는 그 다음 과제

– 이종데이터 결합의 수요가 증가하는 추세이며 향후 더욱 증가할 것으로 전망됨

* 특별히 웹 자료, 메타버스를 포함한 온라인 가상공간에서 생성된 콘텐츠/데이터도 활용, 데이터–콘텐츠간 연계·활용 수요 증가

○ 데이터를 다각도로 수집하여 학습데이터화하고 이를 기반으로 개발한 알고리즘을 통해 다시 생성된 데이터는 경우에 따라 무상 공개하거나 유상 거래하기도 함

– NIA에서 구축 중인 AI 학습데이터 사업의 경우 민간 기업이 참여하더라도 생성데이터 공개

– 연구를 통해 논문을 작성한 경우 활용한 데이터는 논문과 함께 공개하는 것이 일반적

– 기업과 기술개발협력을 통해 생성한 데이터의 경우 사업화를 위한 엠바고 등을 설정하는 경우가 일반적임

제2절 연구데이터 구축·활용 관련 정책이슈[45]

1. 연구데이터정책의 조정 수요

(1) 연구데이터정책의 현황과 간극

○ 전 세계적으로 연구데이터의 가치가 부상하면서 관련 정책과 인프라가 마련되고 있으며, 국내에서도 「연구데이터 공유·활용전략」이 수립되고, 소재연구데이터플랫폼과 바이오데이터스테이션이 구축 중

<표 3-4> 주요국 연구데이터 정책 동향 개괄

구분	연구데이터 정책 동향
글로벌	• OECD, 공공지원 받은 연구데이터의 접근성 제고 지침 개정(2021) • 주요 출판사의 출판논문 데이터 공개 요청 증가 및 데이터출판/저널 등장 • 연구기관 및 대학 차원의 연구데이터관리정책과 서비스 확대
유럽	• 유럽데이터전략과 법제 정비, Horizon Europe 연구지원의 데이터 공개·관리규정 강화 • 데이터관리계획(DMP) 이행, 데이터의 FAIR 원칙, 신뢰성있는 리포지터리 기준 제시 • 유럽 전역을 서비스하는 유럽오픈사이언스클라우드(EOSC) 구축 중
미국	• 연방정부 차원의 데이터 전략 확대, 공공액세스정책(2013)에 따라 연방지원을 받는 연구(펀딩)기관 연구데이터정책 마련, DMP의 도입, 연구데이터관리(RDM) 체계 확립 • NIH, USGS, NIST 등 일부 연방연구(펀딩)기관은 자체적인 데이터센터 구축 및 연구데이터서비스 제공 • 연구데이터뿐만 아니라 코드, 소프트웨어 공개를 위한 CODE 이니셔티브 확대
일본	• 일본 정부의 포괄적 데이터 전략(2021) 및 제6기 과학기술혁신기본계획에서 연구의 디지털 전환(DX) 강조 • JST 등 주요 펀딩기관을 넘어서서 연구기관(대학, 국립연구개발법인 등)의 DMP. 연구데이터정책 이행 기반 확대 • 역내 기관 리포지터리 및 국가 차원 디지털 연구 인프라(NII 포함) 확충
중국	• '14.5 규획'(2021)에서 디지털 전환(DX) 정책 전면화, 13차 과학기술혁신 5개년 계획에 데이터·정보기반 연구 진흥정책 제시 • 국가 차원의 과학데이터관리법제 마련(2018) 및 국가데이터센터 확대·구축 중
한국	• 4차 산업혁명과 디지털 전환 대응 정책, 데이터기반 행정, 데이터산업 및 지식정보연계 제도 이행 단계, 2018년 연구데이터 공유활용전략 이후 국가R&D사업에 DMP 부분 도입 • 분야별 연구데이터플랫폼 - 소재연구데이터(2020) 및 바이오데이터스테이션(2022) - 사업 착수, 국가연구데이터플랫폼(DataOn) 사업을 KISTI 주요사업으로 시작(2020)

45) 본 연구를 위해 추진한 정책 부처 협의, 정책 자문 T/F 회의 등을 통해 도출한 의견을 종합

○ 연구데이터정책은 다른 분야 데이터 전략이나 정책과 달리, 규정과 서비스를 구체화하는 과정에서 연구기관(학회)과 연구자의 참여와 기여 활발

- 연구데이터의 생산, 관리, 활용 과정이 연구과정과 불가분의 관계에 있어, 적실한 데이터 관리, 활용 규정과 서비스를 개발할 때 연구과정에 대한 이해와 개별 연구분야에 대한 전문성 필요

○ 연구데이터가 연구과정과 밀접한 관계 속에서 발전·진화하고 있는 바, 데이터 인프라를 구축하고 운용하는 “데이터” 정책과 연구데이터를 생산하고 활용하는 연구·혁신활동을 진흥하는 “연구” 정책을 상호 연계 운용할 필요

- 연구데이터플랫폼의 구축은 플랫폼을 설계하고 운영하는 사업주관기관에서 담당할 수 있지만, 이를 통해 데이터중심 연구·혁신이 가속화되고 성과를 창출하도록 하는 것은 연구·혁신주체의 참여와 기여없이 달성하기 어려움
- 연구데이터플랫폼 구축을 넘어서는 효과적 운영, 연구·혁신과정과의 긴밀한 소통과 협력구조 속에서 플랫폼을 활용하기 위한 방안 등을 모색할 필요

○ 무엇보다 데이터의 산업적 가치나 공공적 가치를 제고하기 위하여 추진되고 있는 기존의 정책과 차별화된 연구·혁신 중심의 연구데이터정책을 모색할 필요

- 데이터플랫폼이 목적이 아니라 이를 통하여 새로운 연구와 혁신을 추동하기 위한 연구개발정책이 요청됨

(2) 분야별·기관별 데이터중심연구 및 지원 규범의 편차

○ 데이터중심연구의 잠재력과 중요성이 증폭되고 있지만, 앞서 사례에서 볼 수 있듯이 연구데이터의 생산, 관리, 활용 기반과 방식이 분야별·기관별로 상이함

- 국내외적으로 연구를 위한 공개 DB가 증가해왔으나 이를 활용하는 범위와 방식이 연구분야, 기관, 개인별로 차이를 보이며, 연구자(기관)가 직접 생산한 연구데이터를 관리·공유·활용하는 방식의 편차는 더욱 큰 편임

○ 분야별·기관별로 존재하는 데이터역량의 편차를 고려하여 일괄적으로 적용하는 정책이나 사업보다 관련 활동이 활발하여 성과 창출의 가능성이 높은 사례 중심으로 우선 지원하는 선택적·단계적 접근을 고려할 필요

- 데이터중심연구의 속성과 사회적·산업적 파급력이 큰 과제 중심으로 성과가 창출될 수 있도록 기존 연구사업의 정비 및 신규 연구사업의 추가 편성을 검토할 필요

(3) 디지털 플랫폼의 쌍방향성과 데이터 공개-비공개 이분법의 한계

○ 디지털 플랫폼에서 다양한 유형의 쌍방향 교류를 도모하고자 한다면, 연구데이터의 공개-비공개가 절대적이거나 항구적인 조건이 되지 않는 경우가 다수 존재함

- 일부 데이터는 불특정 다수의 대중이 쉽게 접근, 활용할 수 있도록 전면 공개할 수 있지만, 상당수의 연구데이터는 일반 대중에게 전면 공개되지 않지만 연구팀이나 연구커뮤니티 수준에서 유의미하게 공유·활용되고 있음
- 보건의료데이터의 경우 데이터의 다층적 속성과 접근성 통제 필요성을 고려하여 개인 식별이 불가능하도록 최종 가공한 데이터는 온라인상에 전면 공개(public access)하지만 원 데이터는 일종의 보안절차를 마련하여 통제된 환경에서 연구목적으로 활용할 수 있도록 제공(controlled access)하는 접근법이 확산되고 있음

○ 한편, 연구데이터에 접근·활용하기 위해서는 데이터를 설명하는 메타데이터, 워크플로우, 데이터를 분석·활용하는 데 필요한 소스코드와 S/W 역시 데이터 못지않게 중요

[그림 3-9] 연구데이터와 관련 매체

○ 실제 연구데이터에 대한 접근성과 활용도는 데이터뿐만 아니라 관련 매체의 접근·활용도를 함께 고려해야하는 고차방정식에 따라 결정됨

- 소프트웨어-데이터만 고려하더라도 데이터 공개를 허용하면서 소프트웨어나 소스코드 접근을 제한하는 경우나 반대로 소프트웨어를 개방하면서 데이터 자체에 대한 접근은 통제하는 경우 모두 존재

<표 3-5> 데이터-SW 접근성에 따른 실질적 접근성의 차별화

S/W 접근성 / 데이터 접근성	공개	비공개(접근 통제)
공개	연구팀(기관) 중심 공공연구데이터모델	개인연구자 주도 기초학술연구모델
부분 접근 통제	연구팀 데이터 기반 공동생산·분석·활용 모델	연구팀 데이터 기반 개별 분석·활용 모델
비공개(접근 통제)	개별 데이터 (자산) 기반 공동분석지원 모델	개별 데이터 기반 사업화·내부자산화 모델

- 이외에도 하드웨어/활용프로토콜은 공유하지만 데이터와 S/W 통제하는 경우나 데이터와 S/W는 공개하지만 관련 하드웨어를 비공개 혹은 시장 거래하는 경우도 존재
- 따라서, 실제 연구데이터 접근성·활용도는 가능 vs. 불가능의 단정적 판단보다 양자 간 연속스펙트럼 어딘가에서 결정

○ 그럼에도 불구하고 데이터플랫폼을 구축·관리하는 기관이나 관련 규정 마련 단계에서는 아직까지 데이터를 공개-비공개할 지에 대한 특정 시점의 결정 중심으로 데이터 접근 및 관리 기준을 마련하는 경향 존재

- 대체로 연구데이터 등록 시점에서 연구자가 연구데이터의 공개, 비공개, 엠바고 기간 설정 등에 관한 판단을 내리고 이를 기초로 관리
- 데이터 등록 전후 발생가능한 다양한 활용의 수요와 맥락, 활용의 방식에 대한 고려가 필요하며 플랫폼 관리자 중심의 단순 관리를 넘어서는 이용자 중심의 다층적 접근·통제 모델을 고려할 필요

2. 연구데이터플랫폼 구축·활용 현안

(1) 표준화 문제

□ 데이터 표준 현황과 한계

○ 연구·혁신활동을 통해 생산되는 데이터와 기술의 표준화 노력이 지속되고 있으나 여전히 표준화되지 않은 데이터 프로세스 존재

- 미국, 유럽 중심 데이터, S/W 표준이 일반화된 분야가 존재
  · 게놈데이터: 일루미나 시퀀싱 프로토콜 지배적으로 활용, 미국 NCBI나 유럽 EBI 포맷 일반화
  · 지질/지형데이터: 학계에서 인정하는 데이터 포맷과 양식 정착, 지화학데이터: 현재 국제학회 중심으로 국제표준 만드는 중

- 하지만 데이터나 소프트웨어, 관련 매체에 대한 공통의 분류, 포맷, 생산·관리·분석의 프로토콜이 존재하지 않는 분야와 활동이 여전히 다수임

○ 연구활동 전반을 살펴보았을 때 데이터 표준화가 보편화되지 않고 있는 바, 연구데이터의 연계·결합, 데이터 센터간 연계·운영, 데이터플랫폼의 통합 등을 고려할 때 표준화 문제는 주요한 걸림돌로 지적됨

□ 선택적 표준화 전략의 필요성

○ 다만 연구과정과 연구데이터 속성이 불가분 관계에 있고 데이터-S/W-H/W-소재 등이 연계된 상황을 고려한다면 표준화로 해결할 수 있는 현안은 매우 제한적임을 알 수 있음

- 모든 데이터의 양식과 포맷을 일원화하려는 접근이 아니라 데이터 표준화의 범위와 수준을 합리적이고 전략적인 차원에서 설정할 필요가 있음

○ 다양한 형태의 데이터가 생산·공유되는 상황에서 데이터 표준화보다 먼저 선행될 것은 관련 메타데이터를 적실하게 제공하는 것일 수 있음

- 일차적으로 지질자원연구사례에서 알 수 있듯이 데이터와 함께 데이터에 대한 설명 정보(readme 파일) 제공을 권장하는 것이 데이터 활용도를 높이는데 도움이 됨
- 실제 연구데이터에 관한 메타데이터를 상호호환 가능한 형태로 제공·활용하기 위해 메타데이터 표준 개발·확산 추세 강화(RDA 메타데이터 디렉토리 참고)

[참고] RDA 메타데이터 디렉토리 – 표준

○ RDA(Research Data Alliance) 워킹그룹에서 집계한 현재 연구데이터 관련 메타데이터 현황은 다음과 같음

구분	메타데이터 표준
연구데이터 일반	• Registry Interchange Format - Collections and Services • CERIF (Common European Research Information Format) • Data Package • DataCite Metadata Schema • DCAT (Data Catalog Vocabulary) • Dublin Core • MODS (Metadata Object Description Schema) • OAI-ORE (Open Archives Initiative Object Reuse and Exchange) • Observations and Measurements • PREMIS (Preservation Metadata: Implementation Strategies) • PROV • RDF Data Cube Vocabulary • Repository-Developed Metadata Schemas • RO-Crate • schema.org
자연과학·공학	• ABCD (Access to Biological Collection Data) • AgMES (Agricultural Metadata Element Set) • AVM (Astronomy Visualization Metadata) • CF (Climate and Forecast) Metadata Conventions • CIF (Crystallographic Information Framework) • CIM (Common Information Model) • CSMD (Core Scientific Metadata Model) • Darwin Core • DIF (Directory Interchange Format) • EML (Ecological Metadata Language) • FGDC/CSDGM (Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata) • FITS (Flexible Image Transport System) • Genome Metadata • International Virtual Observatory Alliance Technical Specifications • ISA-Tab • ISO 19115 (ISO 19115-1:2014, ISO 19115-2:2009, ISO/TS 19115-3:2016) • MIBBI (Minimum Information for Biological and Biomedical Investigations) • NeXus • Observations and Measurements • Observ-OM

	• ODAM Structural Metadata • OME-XML (Open Microscopy Environment XML) • Open Standard for Particle-Mesh Data (openPMD) • PDBx/mmCIF (Protein Data Bank Exchange Dictionary and the Macromolecular Crystallographic Information Framework) • Protocol Data Element Definitions • Repository-Developed Metadata Schemas • RO-Crate • SDAC (Standard for Documentation of Astronomical Catalogues) • SPASE Data Model • UKEOF

주: 전체 목록에서 연구데이터 일반 표준과 자연과학·공학분야 표준만 발췌·요약
자료: RDA 메타데이터 표준 목록, https://rd-alliance.github.io/metadata-directory/standards/ (접속일: 2022.5.20.)

○ 데이터 연계·결합 활용의 수요가 높은 경우부터 표준화 노력이 이루어질 필요
- 데이터중심연구가 활발하여 연구데이터 표준화 수요가 증가하고 있는 분야부터 메타데이터를 포함한 데이터 표준화 활동을 장려할 필요가 있음
- 머신러닝 활용 연구가 증가하면서 기존 분과학문을 넘어 학문간 이종데이터를 연계·결합하려는 수요가 존재하지만 실제 수요보다 과장되거나 실재하지 않는 경우도 존재함
- 전 연구 분야에 대한 일괄 표준화보다 실제 연구데이터 수요가 있는 연구커뮤니티 중심의 데이터 표준화를 통해 선도 사례를 창출할 필요가 있음
- 이용자(수요자) 입장에서 데이터 접근·활용도를 높이는 방향으로 표준화를 추진할 필요

(2) 플랫폼 연계 문제

□ 데이터 규모에 따른 플랫폼 연계 수요 차별화

○ 연구데이터 생산·관리 및 연계·활용 특성을 결정하는데 있어 데이터의 크기와 생산 방식은 주요한 변수
- 다수의 연구자나 연구기관에서 상대적으로 소규모 데이터를 생산·공유하기 위해 운영되는 데이터 리포지터리나 플랫폼은 웹상 클라우드* 활용

* 주로 민간클라우드(아마존의 AWS, 기타 구글, MS 자사 클라우드나 네이버, KT 클라우드) 사용 중이며 정부에서 구축한 G클라우드 등의 이용률은 높지 않은 편임

- 하지만 취급하는 데이터의 용량이 증가하면서 데이터 전송 시간 지체, 누적 사용료 체증 등의 이유로 클라우드보다 자체 컴퓨팅 시설을 갖추고 저장·관리하려는 수요 발생

* 수십명 대상 게놈데이터 분석을 하는 경우 클라우드를 사용하는 비중도 상당하지만, 울산만명게놈프로젝트와 같이 데이터 수집·분석의 규모가 증가한 경우 자체 전산시설의 필요성 대두

- 대형연구시설장비나 특정 실험에서 생산하는 데이터의 양과 내용이 개별 기관이 저장·처리할 수 있는 수준을 넘어서는 경우 여러 기관이 참여하여 해당 데이터를 분산 저장·처리하는 시스템을 구축·운영함

* 대용량 데이터의 체계적 수집과 저장, 처리를 위하여 개별 기관은 자체 H/W와 S/W가 연계 기관과 상호호환될 수 있도록 운용하고 안정적이고 신속한 데이터 전송을 위하여 기관 간 전용 초고속망을 구축하기도 함

○ 따라서, 데이터의 생산 규모와 방식에 따라 컴퓨팅 시스템의 운영·연계 수요가 달라질 수 있음

<표 3-6> 데이터 특성에 따른 데이터 관리 특성 차별화

관리 특성 / 데이터 특성	데이터 관리 형태	가용 인프라 서비스
다수 거점에서 생산되는 소규모 데이터 (~10TB/GB)	기본적 규범 아래 개인의 자발적 참여와 관리	클라우드 방식 연계 + 상용(지역) 인터넷망 사용
소수 거점에서 생산되는 빅데이터 (수백TB/PB~)	명확한 연구커뮤니티 내에서 관리 의무/역할 분담	자체 데이터센터 + 데이터센터간 초고속망 연계

○ 소규모 데이터를 생산하는 다수를 위하여 필요한 데이터플랫폼의 핵심 요건은 다수의 개방적이고 자발적인 데이터 등록·활용을 장려할 수 있는 이용자 편의적 S/W임

- 다양한 맥락에서 생산된 연구데이터가 FAIR(Findable, Accessible, Interoperable, Resuable)한 형태로 제공되어 관련 분야에서 지속가능하게 이용될 수 있도록 할 필요
- 데이터 연계·활용을 위한 S/W 연동이 중요하며, H/W에 대한 추가 투자 없이 비용효율적 측면에서 민간 클라우드 이용을 권장할 수 있음

○ 한편, 대용량 데이터의 지속적인 저장, 관리, 처리가 필요한 경우 자체 컴퓨팅시설을 보강하거나 KISTI와 같이 슈퍼컴퓨팅센터를 보유한 기관에서 일부 전산자원을 분

양받아 H/W를 운영할 필요가 있으며, S/W뿐만 아니라 H/W, 더 나아가서는 초고속망까지 고려한 플랫폼 연계 방안이 필요

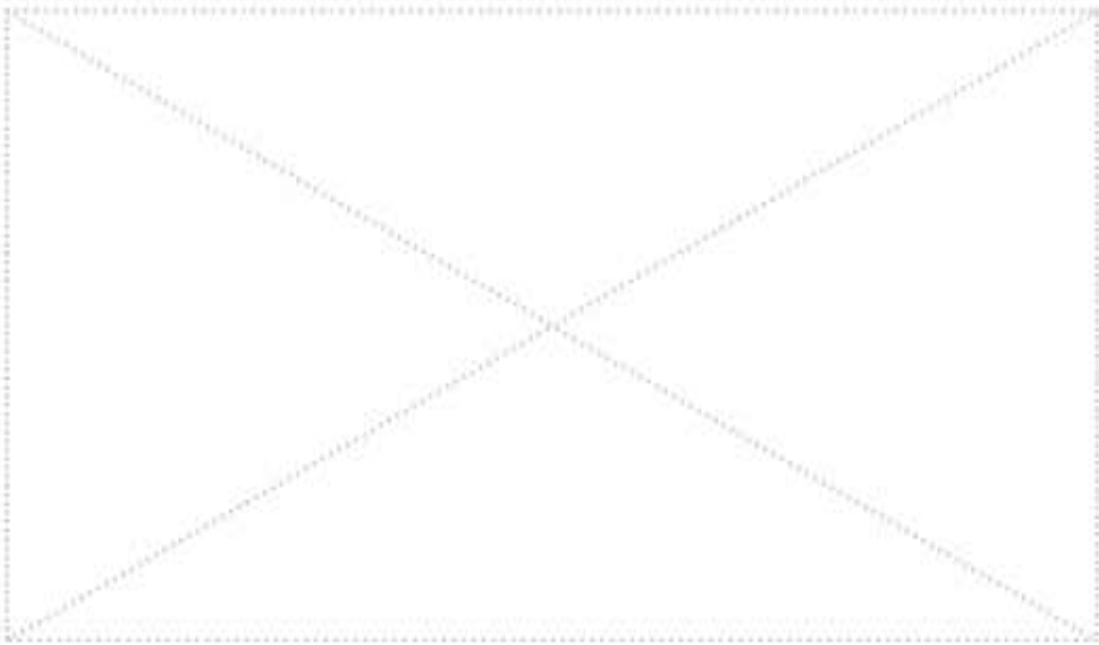
- 개별적으로 구축한 컴퓨팅 시설간 연계를 통해 통합적인 연구데이터 플랫폼 서비스를 구현하기 위해서는 API 전송시스템을 갖추는 것만으로 부족
- 개별 시설에서 구축한 S/W가 상호 연동될 수 있도록 시스템 언어, 구조, 설계 등의 측면도 상호호환 가능해야 하며, 대용량 데이터서비스의 지체를 막기 위해 시설간 초고속망 필요
- 대용량 데이터의 경우 역시 데이터플랫폼의 최종 목적은 이용자 편의적 서비스를 제공하는 것이지만 S/W의 시스템적 측면과 함께 초고속망 연계까지 고려할 필요가 있음

□ 데이터플랫폼과 개별 데이터센터와의 연계 문제

○ 개별 분야·기관별로 데이터센터를 구축·운영하기 시작하면서 개별 센터를 넘어서는 플랫폼 방식의 연계 필요성 제기

- 종적으로 연구자·연구팀의 데이터 프로세스가 기관이나 분야 데이터플랫폼과 연계되고 이러한 분야 플랫폼이 통합 플랫폼과 연계되는 구조
- 플랫폼의 열린 구조상 개인 연구자나 연구팀, 기관에서는 이러한 종적 연계 구조 이외에도 직접적으로 통합 플랫폼에 접근할 수 있고 필요시 연계·활용체계를 구축할 수 있음
- 따라서 개별 데이터센터의 연계 방식이 위계적 구조 이외 다양한 방식으로 확장될 수 있는 상황임

[그림 3-10] 연구데이터플랫폼의 다층적 구조와 기능



○ 개별 데이터 센터와 연계·확장된 플랫폼의 장·단점으로 상호 보완적 역할 필요

- 연구현장과 밀착 운영되는 개별 분야·기관 데이터센터의 경우 연구과정과 상호작용하며 데이터를 생산, 관리, 갱신, 제공할 수 있어 데이터 손실(loss)이 상대적으로 적고 메타데이터 정보가 상대적으로 풍부하지만 소규모 관리 자원과 역량으로 인한 효율성 문제 봉착
- 전담기관을 보유한 분야별 연구데이터플랫폼이나 국가연구데이터 통합 플랫폼의 경우 안정적인 운영 인력과 컴퓨팅 자원을 기 확보하고 있을 수 있으나 연구현장과의 거리감으로 연구과정에서 실시간 생산되는 연구데이터의 변천과 맥락을 이해하고, 데이터 관리 및 활용 방식을 유연하게 제공하기 어려운 한계 존재

□ 국가통합플랫폼과 기타 플랫폼과의 연계 문제

○ 국가 차원의 통합플랫폼과 분야별 데이터플랫폼이 동시 구축·발전하고 있는 가운데, 각각의 기능과 역할이 상호 경합하지 않고 상호 보완될 수 있도록 조정할 필요

- 분야 전문성에 기반하여 개별 데이터센터나 분야별 데이터플랫폼에서 실질적 데이터 관리와 분석을 지원하고 통합 플랫폼에 메타데이터만을 연계하여 통합 검색 및 활용을 지원하는 사례가 확대되고 있는 상황임
- 통합플랫폼과 분야별 데이터플랫폼의 연계, 분야별 데이터플랫폼이 개별 연구팀/연구기관의 데이터센터와 연계되는 방식의 효과성에 대한 점검과 개선 필요

○ 연구데이터관리와 연구현장의 연구관리, 데이터출판 활동의 간극을 좁히고 연구현장의 데이터 프로세스 전 과정이 데이터플랫폼에 효과적으로 연계될 필요

– 연구기관이나 사업단 차원에서 이루어지는 연구관리 기능과 데이터관리 기능이 상호 연계되어 있음에도 불구하고 담당조직과 인력이 달라 별도로 이루어지고 있음

* 연구사업관리는 연구기관/사업단 담당자와 연구사업을 지원한 정부 부처·연구관리전문기관간 이루어지는 반면, 연구데이터관리는 연구과제/사업단 담당자와 데이터플랫폼을 관리하는 데이터센터·플랫폼 운영기관 담당자와 이루어지고 있음

– 연구데이터를 논문, 특허 등의 연구성과 창출시 활용한 경우나 연구데이터 자체를 연구성과로서 출판한 경우, 연구성과정보를 함께 연계·제공할 때 데이터 활용도가 배가될 수 있으나 연구성과정보를 제공하는 플랫폼과 연구데이터 플랫폼이 연동되어 있지 않음

– 전체적으로 연구데이터의 접근성과 활용도를 제고하기 위하여 연구데이터플랫폼과 연구관리정보*, 연구성과정보, 출판데이터 등의 연계를 강화할 필요

* 연구관리정보, 연구과제정보와 연구데이터의 상호연계를 위하여 국가과학기술지식정보서비스(NTIS) 이외 범부처통합연구지원시스템(IRIS)과의 연계 운영방안을 확대할 필요

(3) 국내 플랫폼의 인지도·활용도 이슈

□ 국내 연구자의 해외 리포지터리 선호

○ 미국, 유럽 등을 중심으로 연구데이터를 공개·공유하는 리포지터리가 증가하면서 국내 연구자도 해외의 공개 데이터를 일상적으로 활용

– 일정 기간에 걸쳐 연구커뮤니티 중심으로 분야별 연구데이터 공유가 활성화되면서 데이터의 신뢰성과 최신성 등을 인정받은 주요 리포지터리*는 글로벌 차원의 연구데이터 소스로 성장

* 바이오분야 NIH의 NCBI, EBI, 소재분야 NIST의 MGI, 천문분야 SDSS 등은 글로벌 연구커뮤니티에서 인지도와 신뢰성을 확보하여 글로벌 차원의 연구데이터 서비스 제공

– 우리나라에서도 해외의 주요 리포지터리 이용이 활발하며 대용량 연구데이터의 생산·활용을 위한 글로벌 컨소시움에 개인 혹은 기관 차원에서 참여하는 사례도 다수

<표 3-7> 국내 연구자가 활용 중인 해외 리포지터리 데이터

구분	연구에 활용하는 주요 데이터 소스
해외 리포지터리 데이터	[바이오] NCBI, EBI, KEGG, NIH의 Human Microbiome Project 데이터, TCIA, TCGA, ICGC, Human Protein Atlas, GEO, PDB, DrugBank, PubChem, 계산신경과학 CRCNS 자료, Allen Brain Atlas 데이터 등 [소재] 미국 NIST의 MGI, 캠브리지 CSD, ICSD 무기결정구조 데이터, Materials Project 데이터 등 [우주/천문] NOAA 위성자료, 유럽 EISCAT, 미국 빅베어 태양천문대의 FISS 자료, SDSS 자료, LSST 천체망원경 자료, 가상천문대(VOC) 자료, 허블우주망원경 데이터 등 [물리] CERN 실험데이터, KEK 실험데이터 등

자료: 한국연구재단 기초연구본부 주관 데이터기반연구 현황 조사자료(2022.1.19.) 요약

○ 최근 주요 저널에 논문을 출판하는 과정에서 동료심사 전후로 논문에 활용된 연구데이터를 제출·공개 검증하도록 하는 추세가 강화되면서, 국내 연구자들이 논문 출판 전 해외 주요 리포지터리에 논문데이터를 등록하는 경우가 더욱 증가

- 주요 저널에서 논문데이터 등록을 위해 특정 리포지터리를 지정한 경우도 있지만 대부분 리포지터리의 기본 요건만 제시하고 저자가 리포지터리를 선택할 수 있도록 하고 있음
- 그럼에도 불구하고, 국내 데이터 리포지터리의 글로벌 인지도가 낮아 혹여나 발생할 데이터 접근·검증의 어려움과 이로 인한 논문심사 불이익에 대한 우려가 존재
- 이로 인하여 국내 연구자가 논문 투고시 데이터 제출을 요청받을 경우 국내보다 해외 리포지터리를 우선적으로 활용하는 경향이 지속되고 있음

□ 국내 리포지터리의 인지도·활용도 저조

○ 분야별 연구데이터플랫폼을 비롯하여 기관별 데이터센터 등을 구축 중이지만 대부분 구축 초기 단계인 바, 국내외 인지도와 이용률이 높지 않음

- 전 세계 연구데이터 리포지터리 현황을 목록화하고 있는 re3data에 등록된 국내 리포지터리는 소수에 불과함

○ 연구데이터를 축적하고 관리 신뢰성을 제고하는 데 시일과 자원이 소요되는 바, 중장기적으로 국내 리포지터리의 인지도와 신뢰성을 제고하기 위한 조치 필요

* 포스트게놈유전체사업도 파편화되어있다는 비난을 많이 받았지만 후반부에는 꽤 일관성있는 데이터를 모으고 있는 것으로 평가되며 향후 국내에서도 활용 가능한 데이터가 증가할 것으로 기대

○ 국내 데이터플랫폼 구축시 분석 도구의 다변화와 같은 부가 서비스를 개발하기 이전에

글로벌 수준에 부합하는 연구데이터의 관리 체계와 품질 고도화를 우선할 필요

□ 리포지터리의 신뢰성과 서비스 품질을 강화하는 추세

○ 미국, 유럽 등지에서 연구데이터 리포지터리의 신뢰성을 검증하는 추세 확대
- CoreTrustSeal, nestor Seal DIN31644, ISO16363 등과 같이 신뢰성있는 연구데이터 리포지터리를 인증하는 체계들이 생겨나고 있으며, 미국과 유럽연합 등지에서도 연구데이터 리포지터리의 신뢰성을 점검하는 정책 방향 도입
- 연구데이터 리포지터리가 증가하면 증가할수록 서비스 품질과 안정성, 신뢰성 등을 바탕으로 리포지터리의 등급이나 수준을 검토하고 선별적으로 사용하려는 움직임을 더욱 확대될 전망
- 민감데이터의 관리와 보호, 데이터 보완 관련 요건의 강화 추세 또한 이러한 리포지터리 인증체계 확대를 촉진시키고 있음

○ 연구데이터 리포지터리의 글로벌 신뢰성 기준에 걸맞는 데이터서비스 요건을 갖출 필요성 증가

3. 연구데이터 구축·활용기반 관련 이슈

(1) 데이터 인력 및 전문역량의 부족

○ 연구데이터관리 수요는 증가하고 있으나 학계나 산업계에서 관련 경력자를 양성·계발하기 위한 투자가 부족했던 바, 연구현장의 전문 인력과 역량이 부족한 상황임
- 전반적으로 데이터 프로세스를 이해하는 분야별 전문가나 개별 분야 전문성을 갖춘 데이터과학자·공학자가 부족함
- 데이터센터나 플랫폼에서 데이터 등록·제공과 같은 단순 행정서비스를 넘어서는 데이터 큐레이션, 데이터 서비스까지 담당할 수 있도록 전문인력 충원 및 재교육 필요

○ 연구데이터관리(RDM)가 단순 데이터의 수집·보관을 넘어서는 기술적·행정적·제도적 역량을 요구하는 전문연구활동으로 발전하고 있는 만큼, 이러한 연구현장의 역량이 전반적으로 개선될 수 있도록 정책적으로 진흥할 필요
- 연구기관이나 대학 단위 RDM 서비스를 확충하고 있는 주요 선진연구기관 사례와 달리 국내에서 조직·분야단위 RDM 서비스를 제공하는 사례가 미비하여 연구책임자와 연구팀의 데이터관리 부담이 여전히 큰 편임

– 연구데이터관리가 연구과정 중 일상적으로 일어나야하는 만큼 연구현장에서 연구데이터관리를 밀착 지원할 수 있는 지원 조직과 인력, 역량의 확충이 필요함

(2) 연구데이터에 관한 권리와 책임, 보상제도 불분명

○ 연구데이터의 생산, 관리, 활용 과정에 대한 추적이 어렵고 생산된 데이터가 인용·활용되는 과정에서 데이터 생산자나 제공자에게 보상·인정을 부여하는 시스템 미비

– 데이터 생애주기 관점에서 연구데이터에 디지털 고유식별자를 부여하고 논문과 같이 데이터 자체를 출판하거나 인용하도록 장려하는 체계와 기반이 부족함

* 데이터 출판은 국내에서 시작 단계이고 지질도 인용 등을 제외하면 데이터 인용 또한 초기 단계

– 데이터플랫폼을 운영하는 기관이나 사업단 차원에서 데이터 등록증이나 마일리지와 같은 인센티브를 고안하고 있으나 국가 R&D 사업 전체나 학계에서 통용될 수 있는 보상체계는 여전히 부족함

○ 연구데이터의 등록, 이전, 공유, 활용 등의 과정을 거치면서 다양한 이해관계자가 관여하게 되지만, 이들간 책임과 권리를 명확하게 규정하거나 합의하지 못하여 당사자가 불확실성과 거래비용을 떠안게 되면서 데이터 프로세스를 위축시키는 요인이 되고 있음

– 다만 데이터 이전 및 활용에 관한 데이터관리기관의 정책 구체성이 부족하고 데이터 라이선스나 당사자간 협의를 지원하는 모델계약서 등에 관한 합의와 가이드라인 부족

– 데이터 이전, 공개, 활용 등에 관한 기관이나 사업단의 규정이 마련된 곳이 많지 않음

– 데이터 생산자와 이용자, 플랫폼 운영자간 데이터 이전 및 활용의 범위와 방식, 책임과 보상에 관한 협약이나 계약이 일반화되지 않았으며 이를 위한 모델 계약서, 협약 지침서, 가이드라인 등도 부족함

– 한편으로 데이터 공개와 개방을 강조하면서 다른 한편으로는 데이터 사업화, 데이터 구독, 데이터 거래가 시도되고 있지만, 관련 시장이 성숙되고 있다고 판단하기 이르며 데이터 가치의 산정, 적절한 공개·공유의 범위와 절차 등에 관한 협의를 이루는 절차 미비

* 데이터거래소, 데이터바우처사업 등을 통한 거래나 특정 기업/산업 데이터에 대한 거래가 산발적으로 진행되지만 전체적인 데이터 시장이 형성·규모화되는 단계는 아니며 연구데이터를 통해 수익을 창출한 사례는 매우 드뭄

- 오픈라이선스를 포함하여 데이터 특성을 고려한 라이선스 설정 모델도 부족
  * 연구데이터 공개를 위하여 CC0 라이선스를 준용하는 정도이며 데이터/소프트웨어/메타데이터 등에 관한 다층적 라이선스 검토 및 부여 사례 부족

○ 연구데이터를 생산, 관리, 활용하는 당사자간 협약과 계약, 불특정 다수를 대상으로 적용할 수 있는 연구데이터 라이선스, 기타 연구데이터 인용·출판의 장려 등을 통해 데이터 관련 책임과 권리를 구체화하고 현실화할 수 있는 제도적 보완책이 필요함

(3) 민감데이터 관리·활용의 문제와 데이터 보안

○ 바이오헬스연구에서 개인정보보호법, 생명윤리법 등의 규제로 인하여 동일 환자 중심 데이터 결합이 여전히 제한적임
- 2020년 데이터3법(개인정보보호법, 신용정보법, 정보통신망법) 개정으로 연구목적 개인정보활용 가능성을 넓혔으나, 생명윤리법이나 의료법 등 바이오헬스분야에서 적용되는 법률의 개정은 이루어지지 않고 있음
- 개인 연구자나 민간기업 등에서 분쟁의 요소를 회피하려는 경향으로 인하여 민감데이터의 결합·연계 연구 위축
- 개인정보 보호 등을 위한 조치를 유지·발전시키되 부가가치가 큰 의료기술개발사업에 대해서는 필요한 연구데이터가 보다 효과적으로 연계·활용될 있도록 법률적·제도적 보완이 필요함

○ 민감데이터 등은 개인정보 보호, 정보보안 등의 사유로 클라우드 사용에 어려움이 있으며, 이를 개선하기 위한 기술적·제도적 보완책이 필요함

제4장. 연구데이터 구축·활용 촉진을 위한 정책(안)

1. 연구데이터 구축·활용 정책의 방향성

(1) 정책이슈 기반 정책개선(안) 도출

○ 앞서 살펴본 연구데이터 구축·활용 현황과 정책 이슈를 기반으로 삼아 향후 개선할 정책의 방향과 과제(안)를 다음과 같이 제안

[그림 4-1] 정책이슈에 대응하는 정책개선과제(안)

자료: 저자 작성

(2) 연구데이터 구축·활용 정책의 방향성

○ 국가 차원의 연구데이터플랫폼 구축 사업을 데이터중심연구를 진흥하기 위한 연구개발정책의 수단으로 바라보고 근간이 될 데이터중심연구의 진흥 정책 본격화

- 연구 과정의 일부로서 연구데이터 관리를 이해하고 연구협력이 활성화되고 실질적인 연구성과가 창출될 수 있는 방향으로 데이터 공유·활용, 플랫폼 연계, 데이터 표준화
- 연구데이터뿐만 아니라 관련 디지털 매체로 확대하여 연구 분야별 차별화된 데이터 수요와 역량을 살피고 이에 적합한 연구데이터의 구축, 공유, 활용 모델 확산

○ 구체적 실천과제로서 적어도 다음의 세 가지 방안 – 데이터중심연구 진흥 사업의 추진, 연구데이터플랫폼 등의 정비, 데이터 인력·제도의 확충 – 제안

– 데이터 관리, 플랫폼 운영 중심 사업을 넘어서는 데이터중심연구 진흥 사업의 추진 (성과 중심, 연구협력 중심, 선도사례 중심 데이터중심연구 활성화)

– 차등화된 데이터 접근 모델의 구현, 연구·혁신 중심 플랫폼 연계, 수요 기반 데이터 표준화 등 연구와 혁신을 활성화하는 차원의 연구데이터 인프라·서비스 구현

– 지속가능한 연구데이터 인력과 제도, 인센티브의 확충

2. 연구데이터 구축·활용 촉진을 위한 정책과제(안)

(1) 데이터중심 연구·혁신을 촉진하는 연구데이터 정책의 추진

○ 국가 차원의 연구데이터정책 추진 체계의 마련 및 연구데이터–R&D 정책의 연계

– 데이터중심 연구·혁신의 수월성, 선도성을 촉진하고 연구데이터 및 데이터서비스의 신뢰성을 제고하며 연구협력과 성과 창출을 활성화

– 범부처 연구데이터정책 총괄 행정조직의 정비 및 R&D 정책과의 연계·협력 체계 마련

– 민간주도 데이터중심 연구·혁신을 위한 연구데이터정책위원회의 구성 및 운영

(2) 데이터중심연구의 활성화

○ 데이터중심연구사업은 실질적 연구협력과 성과를 창출하는 방식으로 지원하되, 연구데이터커먼즈사업과 연구데이터플랫폼 고도화를 통해 데이터 공유·공개 활용 저변 확대

– 기존 연구데이터 플랫폼 구축 중심의 사업계획에 데이터중심연구 활성화 방안을 추가하여 실질적인 연구협력과 성과 창출 도모

□ 데이터기반 선도연구 프로젝트를 통한 우수성과의 창출

○ 바이오, 소재, 에너지, 기후변화 등 증가하는 도메인 중심 연구데이터를 통해 선도적 연구와 탁월한 연구성과를 창출하는 R&D 과제 지원 프로그램 신설·추가

– 전 세계적으로 연구계의 디지털 전환과 데이터중심연구가 확대되는 상황에서 우리나라의 연구 경쟁력과 파급력을 제고하기 위한 선도연구 프로젝트 선정·지원

* 극지, 우주, 천문, 입자물리, 핵융합 등의 대형연구프로젝트와 다수의 바이오/의료, 소재, 지질자원, 환경연구에서 데이터집약형 연구 성과와 협력 활성화

– 연구성과 창출을 목적으로 특정 분야에 국한되지 않는 다학제연구나 산학협력, 융합연구를 추진하는 사례를 별도로 지정하여 연구초기단계에서부터 향후 연구성과의 사업화, 데이터 거래 등이 가능할 수 있도록 지원

– 선도연구 프로젝트를 통해 연구 성과의 영향력과 파급력을 높이는 데이터 관리·공유 방안 고도화

<표 4-1> 데이터중심연구 활성화를 위한 선도연구사업(안)

구분 / 특징	데이터기반 선도연구사업 (성과창출형)	
		협력성과 창출형
사업목표 (평가기준)	연구 수월성 및 파급력, 국가 전략성	연구 수월성, 실용성 및 파급력, 국가 전략성
데이터 공개 특성	연구팀내 혹은 협력기관간 공유, 데이터 출판·인용 장려	연구팀내 혹은 협력기관간 공유, 데이터 거래·인용 장려
기타 사항	–	필요시, 산업데이터정책과 연계

□ 데이터 연계·공유를 장려하는 연구데이터커먼즈사업의 추진

○ 연구과정에서 연구팀을 넘어서는 다른 연구기관과 데이터 공유 협력을 도모하려는 시도를 장려하고 연구자 주도 데이터 연계·공유를 활성화하는 상향식 연구데이터커먼즈사업의 기획 및 추진

○ 국가 차원에서 개별적 노력을 넘어서는 공동의 데이터 구축·활용 노력이 필요한 분야를 전략적으로 선정하여 해당 분야 연구데이터커먼즈를 확대하는 하향식 사업 동시 추진(목적중심 연구데이터커먼즈 구축 참고사례: 인간게놈프로젝트)

– 연구분야에서 기 구축된 데이터의 연계·활용을 촉진하려는 효과성 중심 접근이나 이미 공개적으로 활용되는 데이터를 재가공한 데 대한 공개 의무를 강화하는 공공 액세스 접근뿐만 아니라, 국가 차원에서 데이터중심 연구·혁신을 촉진하기 위하여 연계·공유가 시급한 분야에 대한 전략적 탐색과 선정 절차 고안·마련

<표 4-2> 데이터중심연구 활성화를 위한 연구데이터 커먼즈사업(안)

구분 / 특징	연구데이터 커먼즈사업 (데이터성과 공유형)
사업목표 (평가기준)	공공성, 효과성 및 파급력, 국가 전략성
데이터 공개 특성	대국민 공개서비스 지향하되 연구커뮤니티 단위 커먼즈 공유 허용
기타 사항	필요시, 공공데이터정책/국가지식정보시책과 연계

(3) 수요자 중심 플랫폼 연계 및 고도화

□ 수요를 반영하는 프로세스 설계 및 고도화

○ (기획 단계) 수요자 중심의 플랫폼을 설계하기 위해 기획 초기에 수요 조사 등을 통해 연구자의 요구를 반영하는 프로세스 설계

– (참고 사례) 미국 'All of Us'는 2018년 350명 이상이 참여하는 전문가 워크숍을 개최하여 1,100여 개의 활용사례*를 수집함으로써 연구자의 니즈를 플랫폼 설계에 반영

* 활용사례에는 ① 연구 질문, ② 데이터 형태(data type), ③ 수집방법(method), ④ 수집 주기(specification) 등이 포함되어 있어 연구자들의 니즈를 구체적으로 파악 가능

○ (운영 단계) 수요 반영 프로세스는 초기 기획단계 뿐만 아니라 플랫폼 운영 과정에서도 필요한 것으로 빠르게 변화되는 데이터 환경에 대응하기 위해 연구자의 요구를 지속적으로 조사하고 관리하는 프로세스 수립 및 고도화

□ 수요중심 선택적 표준화와 차등화된 데이터 접근 모델의 확산

○ (선택적 표준화) 연구과정과 연구데이터 속성이 불가분 관계에 있는 상황을 고려할 때, 연구과정 전체를 표준화할 수는 없으므로 연구 수요 중심으로 선택과 집중
- 국내 연구 수요와 글로벌 연구·혁신 동향을 고려하여 파급력이 큰 분야의 데이터 표준화 우선, 이종데이터 연계를 위한 표준화도 철처히 연구 수요에 기반하여 진행

○ (차등화된 접근) 일괄 등록-공개의 논리를 넘어서서 연구커뮤니티 수요에 대응할 수 있는 다층적 공개·공유 모델 도입
- 국가보안이나 개인정보보호, 산업재산권 보호 등의 이유로 단계별 “통제된 접근(controlled access)”을 제공하는 방식 고안·제공
  * 데이터 등록 이후 일정기간 데이터 등록자/제공자만 활용할 수 있는 “데이터 안심구역” 제공도 검토 가능
- 데이터뿐만 아니라 데이터 접근·활용을 위하여 필요한 디지털 매체 등에 대한 접근성을 종합적으로 고려한 다원적 데이터 접근 모델의 구현

□ 연구자(수요자) 필요에 따른 연구데이터 플랫폼 확충과 연계

○ 연구데이터플랫폼의 핵심 고객인 연구자의 서비스 수요에 부응하기 위한 자원 확충
- 데이터중심 연구를 위하여 필요한 컴퓨팅 자원이나 클라우드의 공유, 데이터 저장/백업 서비스의 제공 등에 대한 지원 확대
  * 클라우드 vs. 자체 컴퓨팅시설의 수요가 분야·기관마다 다른 바, 연구분야와 수요를 고려하여 추진

○ 수요자 중심 연구데이터플랫폼-데이터센터 간 연계
- 국가연구데이터 플랫폼과 분야별 연구데이터 플랫폼과의 연계, 이들과 개별 연구기관 혹은 사업단에서 운영하는 데이터센터와의 연계 효율화
- 수요자 입장에서 원스톱(one-stop) 서비스를 구현하기 위한 플랫폼/시스템간 기술적·조직적 연계(시스템 언어, 구조, 설계 등의 호환 및 외부 이용자 중심 FAIR 원칙 구현) 강화
  * 데이터 연계를 넘어선 플랫폼 연계를 통해 실시간 데이터 전송·분석이 가능한 환경의 구현

○ 연구데이터와 연구관리/연구성과 시스템과의 연계 서비스 강화
- 국내 연구데이터플랫폼을 통해 공개·공유되는 연구데이터의 상당 부분이 국가

R&D사업을 통해 산출·등록되고 있는 바, 등록·공개된 연구데이터를 지원한 연구과제에 대한 정보와 해당 과제에서 함께 산출된 기타 연구성과에 대한 정보가 함께 제공·서비스될 수 있도록 연구개발정보시스템의 연계 확대

* 연구데이터와 연구관리정보 연계 서비스를 위하여 연구데이터플랫폼과 연구관리전문기관, 범부처통합연구지원시스템(IRIS)과 국가R&D 연구성과물 관리전담기관, 국가과학기술지식정보서비스(NTIS) 등의 시스템 연계 확대

- 연구데이터플랫폼(리포지터리)과 주요 검색엔진(학술검색엔진뿐만 아니라 일반국민 이용 상업검색엔진 포함) 연계를 강화하여 연구데이터와 이를 활용한 학술논문, 기타 연구성과간 연계 검색·활용이 가능한 기반의 조성

□ 신뢰받는 글로벌 데이터 리포지터리·플랫폼의 육성

○ 국내 연구데이터플랫폼과 기관 리포지터리, 데이터센터의 글로벌 인지도와 신뢰성, 접근성을 제고하기 위히여 신뢰받는 리포지터리(trusted repository) 플래그쉽 사업의 추진

- 국내 데이터 리포지터리와 플랫폼이 글로벌 데이터 리포지터리 연합활동에 지속 참여하며 데이터 표준 및 플랫폼 운영·연계 관련 글로벌 동향을 파악·대응하도록 장려
- 전 세계적으로 데이터 보안, 품질관리 등에 관한 기준이 구체화되고 리포지터리 인증제도 등이 확대되고 있는 바, 국내 데이터 리포지터리의 글로벌 평가와 신뢰도를 제고하기 위한 조치 확대
- 데이터뿐만 아니라 플랫폼(리포지터리) 또한 글로벌 인증 기준 등에 부합할 수 있도록 운영 프로세스 고도화

○ 국내 연구데이터 플랫폼(리포지터리)이 글로벌 학계에서 신뢰성과 인지도를 확대하여 국내 연구자들이 해외유수저널에 출판할 때에도 국내 플랫폼(리포지터리)를 자발적으로 활용할 수 있도록 유도

(4) 연구데이터 구축·활용을 위한 기반 강화

□ 연구데이터 전문인력 및 데이터 역량 육성

○ (플랫폼 서비스) 단순 행정 지원을 넘어선 데이터 큐레이션을 지원할 수 있는 연구데이터 전문인력 확충·육성

○ (연구분야별) 도메인 지식뿐만 아니라 데이터 역량을 갖추도록 분야별 데이터 역량 교육 확대 및 융합형 데이터 인재 육성

○ (현장지원 강화) 데이터 품질 관리, 플랫폼 운영 등의 기술적 역량과 함께 연구현장에서 데이터의 이전·유통과정에서 발생하는 제도적 이슈를 취급·처리·지원할 수 있는 지원 역량 확충

□ 데이터 인용 및 출판 활성화

○ 연구데이터에 디지털고유식별자를 부여하여 연구데이터의 활용시 데이터 인용을 할 수 있는 기반을 확대하고 데이터 출판, 데이터 저널이 성장할 수 있도록 장려
- 연구데이터와 메타데이터 등이 생성, 진화하는 과정을 추적할 수 있는 디지털 고유식별자의 부여·활용체계를 구축해 나가는 한편, 이를 활용하여 데이터가 인용·유통·평가되는 생태계 조성

○ 기존 출판산업의 지형이 급변하고 있는 가운데, 학술정보 및 연구데이터산업에서 국내 관련 기업이 성장할 수 있는 발판 마련

□ 연구데이터 특수성을 고려한 제도·지침 마련 및 적극 행정

○ 연구데이터의 특수성을 고려하여 데이터 소유, 보상, 거래, 이전, 사업화 등을 추진할 수 있는 제도적 기반과 가이드라인 마련
- 연구데이터 거래의 경우 기본적으로 산업데이터와 같이 당사자간 협의를 우선으로 할 수 있도록 하나, 이 과정에서 고려할 사항을 담은 모델계약서·협약서 등의 정보를 제공하여 불공정계약 사전 방지
- 연구데이터 특성을 고려한 데이터 라이선스 지침이나 연구성과 속성을 가질 경우 데이터 관련 책임과 권리 관계 등에 대한 상세 지침 제공
- 이외에도 플랫폼 혹은 기관 단위에서 추진되는 보상제도(마일리지제도, 데이터 구독 등)나 책임 규정 등의 운영 기반에 대한 제도적 프레임워크 제공

○ 연구데이터 공유·활용 촉진을 위한 법제 해석 지원 및 적극 행정 강화
- 민감데이터의 취급과 결합, 이전 및 활용을 계획하고 실행할 때 필요한 법적·윤리적 조치에 대한 지원, 지식재산권과 연구데이터 공개·이전 간 발생할 수 있는 갈등 조정, 기타 보안 및 이해상충 관련 제도적 불확실성을 제어할 수 있는 법해석 지원 서비스의 확대
- 연구데이터의 공유 및 활용을 촉진하기 위한 적극적 행정 지원

<표 4-3> 정책과제 제안 종합(안)

1. 데이터중심연구를 진흥하는 국가 차원의 연구데이터 정책 추진
° **데이터사업과 제도를 총괄하는 연구데이터정책의 기획·이행** – 국가 차원의 정책 추진 방향과 기준 마련, 연구데이터정책의 기획 및 이행 체계 구축
2. 데이터중심연구를 통한 실질적 성과 창출
2-1. **데이터 기반 선도프로젝트를 통한 우수모델 도출** – 바이오, 소재, 기후변화 등 데이터중심연구가 활발하고 실질적 성과 창출이 기대되는 분야를 중심으로 데이터중심연구 선도모델 구축 – (협력 연구) 데이터 공동 구축보다 데이터 기반 공동연구가 핵심 활동
2-2. **데이터 연계·공유를 위한 연구데이터커먼즈사업의 추진** – 상향식 및 하향식 연구데이터 공유·연계 활동 동시에 장려 – (가칭)연구데이터 센터를 통해 연구자들의 의견을 수렴하고 데이터 연계·활용 현장거점 및 연구단계별 연구데이터 협력연구 플랫폼 구축·운영 – (데이터 공동 구축) 공동연구를 통한 양질의 연구데이터 구축·공유 포함
3. 수요자 중심 플랫폼 연계 및 표준화, 활용도 제고
3-1. **연구중심 데이터 플랫폼 연계 및 확충** – 수요 기반 데이터 표준화 및 플랫폼 자원(데이터관리, 컴퓨팅자원 공유 등) 확충 – 수요자 중심 one-stop 서비스 구현을 위한 플랫폼간 역할·기능 정비, 데이터센터와 플랫폼, 연구관리/성과정보와 플랫폼간 연계
3-2. 수요자 중심 **연구데이터 표준화** – (국제 표준화) 연구데이터 관련 국제적인 표준화 활동 지속 대응·참여 – (국내 표준화) 분야별 연구데이터 표준화 전략 수립 – (전문가 포럼) 연구데이터 표준화 포럼, 국제표준화 대응위원회 등 운영
3-3. **신뢰받는 글로벌 데이터 리포지터리(trusted repository) 육성** – 국내 연구데이터 리포지터리와 플랫폼의 인지도와 신뢰성 제고 플래그쉽 추진
4. 지속가능한 연구데이터 인력과 제도, 인센티브의 확충
4-1. **데이터중심연구 지원 핵심 인력 양성** – 데이터 플랫폼 및 연구분야에서 데이터 큐레이션 전문인력·자원·교육 확충
4-2. **데이터 인용 및 출판 활성화** – 논문을 넘어 데이터를 직접 인용하거나 출판하는 학계 활동 및 산업 육성
4-3. **연구데이터 특수성을 고려한 제도와 지침 마련 및 행정지원** – 데이터 권리/책임 등을 구체화할 제도 마련 및 데이터 공유·활용 촉진 행정

3. 연구데이터 구축·활용 촉진을 위한 정책과제 추진전략

○ 제안한 정책과제는 크게 두 가지 – 신규 투자예산을 동반한 사업성 과제와 신규 정책이나 제도의 도입, 기존 제도와 조직의 개편을 동반한 제도 개선 과제 – 로 구분

<표 4-4> 정책과제(안)의 추진

	사업 예산 투자	제도 정비 개선	국제 협력
1. 데이터중심연구를 진흥하는 국가 차원의 연구데이터 정책 추진			
° 데이터사업과 제도를 총괄하는 연구데이터정책의 기획·이행 – 국가 차원의 정책 추진 방향과 기준 마련, 연구데이터정책의 기획 및 이행 체계 구축		○	
2. 데이터중심연구를 통한 실질적 성과 창출			
2-1. **데이터 기반 선도프로젝트를 통한 우수모델 도출** – 바이오, 소재, 기후변화 등 데이터중심연구가 활발하고 실질적 성과 창출이 기대되는 분야를 중심으로 데이터중심연구 선도모델 구축 – (협력 연구) 데이터 공동 구축보다 데이터 기반 공동연구가 핵심 활동	○		△
2-2. **데이터 연계·공유를 위한 연구데이터커먼즈사업의 추진** – 상향식 및 하향식 연구데이터 공유·연계 활동 동시에 장려 – (가칭)연구데이터 센터를 통해 연구자들의 의견을 수렴하고 데이터 연계·활용 현장거점 및 연구단계별 연구데이터 협력연구 플랫폼 구축·운영 – (데이터 공동 구축) 공동연구를 통한 양질의 연구데이터 구축·공유 포함	○		△
3. 수요자 중심 플랫폼 연계 및 표준화, 활용도 제고			
3-1. **연구중심 데이터 플랫폼 연계 및 확충** – 수요 기반 데이터 표준화 및 플랫폼 자원(데이터관리, 컴퓨팅자원 공유 등) 확충 – 수요자 중심 one-stop 서비스 구현을 위한 플랫폼간 역할·기능 정비, 데이터센터와 플랫폼, 연구관리/성과정보와 플랫폼간 연계	△	○	△
3-2. 수요자 중심 **연구데이터 표준화** – (국제 표준화) 연구데이터 관련 국제적인 표준화 활동 지속 대응·참여 – (국내 표준화) 분야별 연구데이터 표준화 전략 수립 – (전문가 포럼) 연구데이터 표준화 포럼, 국제표준화 대응위원회 등 운영	△	○	△
3-3. **신뢰받는 글로벌 데이터 리포지터리(trusted repository) 육성** – 국내 연구데이터 리포지터리와 플랫폼의 인지도와 신뢰성 제고 플래그쉽 추진	△	○	△
4. 지속가능한 연구데이터 인력과 제도, 인센티브의 확충			
4-1. **데이터중심연구 지원 핵심 인력 양성** – 데이터 플랫폼 및 연구분야에서 데이터 큐레이션 전문인력·자원·교육 확충	○		△
4-2. **데이터 인용 및 출판 활성화** – 논문을 넘어 데이터를 직접 인용하거나 출판하는 학계 활동 및 산업 육성	△	○	△
4-3. **연구데이터 특수성을 고려한 제도와 지침 마련 및 행정지원** – 데이터 권리/책임 등을 구체화할 제도 마련 및 데이터 공유·활용 촉진 행정		○	△

○ 국가차원의 연구데이터 정책 수립 및 데이터 특수성을 고려한 제도 개선 과제가 대표적인 제도 개선 과제이며, 데이터중심연구 활성화를 위한 선도프로젝트와 연구데이터커먼즈사업 기획은 대표적인 사업성 과제임

– 플랫폼 연계 및 표준화, 활용도 제고를 위한 제반 활동 등은 사업예산 소요와 제도개선 소요를 모두 동반

○ 사업성 과제의 경우, 국가 정책에 선행하여 파일럿 형태로 시작하는 방안과 국가정책 수립 후 체계적으로 규모 있게 추진하는 방안 모두 검토·이행

– 국가 연구데이터 정책의 거버넌스를 정립하고 중장기 계획을 수립하는데 상당기간이 소요될 것으로 예상되는 바, 이를 선행하여 데이터중심연구사업의 파일럿 시행도 가능할 것임

○ 기 제안·추진 중인 연구데이터사업 및 정책과 연계한 과제의 조속한 이행

– 바이오와 소재분야 연구데이터플랫폼 사업이 시작된 만큼, 해당 사업 등에 대해서는 수요자 중심 플랫폼 연계, 표준화, 활용도 제고 전략을 검토·도입

– 신정부 국정과제로 제시된 국가연구데이터플랫폼의 구축 과정에서 제시된 플랫폼 연계·활용 전략을 적극 수용, 반영

○ 중장기적으로 연구데이터 생태계 조성을 위한 제도 개선 및 시장 형성 노력 지속

– 연구데이터 특수성을 고려한 관련 법제도의 개선, 데이터 표준 등의 주요 관리체계 정비, 데이터 인용 및 출판 시장의 형성 등은 중장기 과제로서 지속적으로 추진

참고 문헌

1. 국문자료

관계부처 합동(2018), 「데이터산업 활성화전략」

관계부처 합동(2019),「-혁신성장 전략투자- 데이터·AI경제 활성화 계획」

과기정통부 보도참고자료(2020.4.19.), “소재 연구데이터 플랫폼 구축방안 수립(소부장 특위)”

과기정통부 보도자료(2021.6.1.) “국가지식정보 연계 및 활용 촉진에 관한 법률 제정”

국가과학기술심의회 운영위원회(2018),「연구데이터 공유· 활용 전략」

김지현(2014), “대학도서관의 연구데이터관리서비스에 관한 연구: 미국 연구중심대학도서관을 중심으로,” 한국비블리아학회지 25(3), 165-189.

동아사이언스(2018.2.2.), “중국, 둘레 100km 입자가속기로 ‘과학 굴기’”, https://www.dongascience.com/news.php?idx=21341(2022.5.1.접속)

매일경제(2017.8.2.), “한국형 인공태양 실험 결과 미국 슈퍼컴퓨터로 본다.” https://www.mk.co.kr/news/it/view/2017/08/517964/(2022.4.3.접속)

신은정 외(2018), 오픈사이언스를 통한 공공연구 효과성 제고 방안, 과학기술정책연구원.

신은정 외(2019a), 과학기술 연구데이터 공동 활용 기반 마련을 위한 연구, 과학기술정보통신부.

신은정 외(2019b), 공공부문 연구데이터 소유·활용제도 개선방향, 과학기술정책연구원.

신은정 외(2021), 국가연구개발사업 연구데이터 성과인정체계 마련연구, 과학기술정보통신부.

신은정(2021), “연구데이터 접근성 제고를 위한 글로벌 규범의 형성과 과제,” Global Legal Issues 21-17-5, 한국법제연구원.

신은정·정원교 (2017) 오픈사이언스정책의 확산과 시사점, STEPI Insight 제216호, 과학기술정책연구원.

제20대 대통령인수위원회 (2022) 「윤석열정부 110대 국정과제」, https://www.evaluation.go.kr/web/page.do?menu_id=143

정일영 외(2021), 헬스케어 데이터 공공 플랫폼의 활성화를 위한 통합적 전략 연구, 과학기술정책연구원.

정일영 외(2022), “헬스케어 데이터 플랫폼 활성화의 핵심과제와 정책 제언,” STEPI

Insight, 과학기술정책연구원 (발간 예정)

한국연구재단(2020), 연구과제 신청, 평가와의 연계 및 사후관리를 위한 절차

한국연구재단 기초연구본부 주관 데이터기반연구 현황 조사자료(2022.1.19.)

황석원 외(2020), 데이터 기반 연구개발 관리 혁신 방안, 과학기술정책연구원.

행정안전부, 「제1차 데이터기반행정 활성화 기본계획('21년-'23년)」, 행정안전부.

KIAT 산업기술정책센터(2022.5.9.) "디지털 정책 및 제도 동향(일본)", 참고 (원 자료: デジタル関連先進技術の制度政策動向調査レポート(디지털 관련 첨단기술의 제도 정책 동향 조사보고서), (日) 독립행정법인 정보처리추진기구 사회기반센터 혁신추진부)('22.2)

KIAT(2022), "디지털 정책 밍 제도 동향(중국)," KIAT 산업기술정책센터 정책자료 공유(2022.5.9.)

KISTI (2017), 데이터 기반 과학기술 R&D를 위한 클라우드 서비스 기획연구, 과학기술정보통신부.

KISTI(2019), 국가연구데이터플랫폼 구축 및 운영, 과학기술정보통신부.

KISTI(2019), 출연(연) 연구데이터 관리 활용 방안 연구, 국가과학기술연구회.

KISTI(2020), 대용량 실험데이터 활용 및 육성에 관한 법제적 지원 방안, 한국정보기술연구원.

KISTI(2021), 연구데이터 공유확산체제 구축, 한국정보기술연구원.

KOBIC(2022), K-BDS 홍보자료 및 내부자료.

2. 해외자료

일본 閣議決定(2021), 「第６期 科学技術・イノベーション基本計画(令和３年３月２６日)」, 제6기 과학기술혁신기본계획 전문, https://www8.cao.go.jp/cstp/kihonkeikaku/index6.html

村山泰啓(2018. 3. 17), 「Research Data Aliance(RDA)とは何か」, NICT

중국 国务院办公厅关于印发 科学数据管理办法的通知 - 国办发 [2018] 17号, http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm (접속일: 2020.4.1.)

Altman,M, Crosas, M.(2013), The evolution of data citation: from principles to implementation. IAssist quarterly, 37(1-4), 62-70.

Briney, K. & Goben, A. & Zilinski, L.,(2015), "Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies", Journal of Librarianship and Scholarly Communication 3(2), p.eP1232. doi: https://doi.org/10.7710/2162-3309.1232

Clauset, A., D. Larremore, and R. Sinatra(2017), "Data-driven predictions in the science of science," Special Issue Essay, Science 355(6234): 477-480.

EC(2018.4.15.), Elements of the European data economy strategy 2018, https://digital-strategy.ec.europa.eu/en/library/elements-european-data-economy-strategy-2018(접속일: 2022.5.1.)

EC(2020) Exploitation & Open Science in Horizon Europe, https://ec.europa.eu/research/participants/data/ref/h2020/other/events/2020-10-09/3_exploitation-ipr-open_science_en.pdf?msclkid=4862130abad211ecb6c1f7c64bf68939 (접속일: 2022.3.1.)

EC(2021.12), Horizon Europe General Model Grant Agreement.

EC(2021.11), EU Grants AGA - Annotated Model Grant Agreement: EU Funding Programmes 2021-2027.

EC(2022), Staff Working Document on Common European Data Spaces, https://digital-strategy.ec.europa.eu/en/library/staff-working-document-data-spaces(접속일:2022.5.1.)

ERC(2022), Open Research Data and Data Management Plans - Information for ERC grantees, European Research Council.

Hayashi(2020.9.15.), UNESCO 아태지역회의 발표자료.

Hey, T., Tansley, S. and Tolle, K. (2009), The Fourth Paradigm: Data-intensive Scientific Discovery, Microsoft Research.

Huang Y, Cox AM, Sbaffi L. (2021), Research data management policy and practice in Chinese university libraries. J Assoc Inf Sci Technol. 2021;72:493-506. https://doi.org/10.1002/asi.24413

Hutter and Zenlusen(2018), "The Cancer Genome Atlas: Creating Lasting Value beyond Its Data," Commentary, Cell 173, 283-285.

Jones, L. Grant, R. Hrynaszkiewicz, I.(2019), “Implementing publisher policies that inform, support and encourage authors to share data: two case studies,” UKSG Insights 32(1): 11, http://doi.org/10.1629/uksg.463

Kitchin, Rob(2014), “Big Data, new epistemologies and paradigm shifts,” Big Data & Society 1(1).

Lipton, Vera(2020), Research Data Management at CERN, Open Scientific Data: Why Choosing and Reusing the RIGHT DATA Matters.

Murray, M., et al.(2019), A Survey of the State of Research Data Services in 35 U.S. Academic Libraries, or “Wow, what a sweeping question”, Research Ideas and Outcomes, 5:e48809, https://doi.org/10.3897/rio.5.e48809

Ma, Juncai(2020.9.16.), 유네스코 오픈사이언스권고(안)에 관한 아시아태평양지역 온라인 컨설팅회의, 발표자료.

NIH(2020.11.12), All of Us Researcher Onramp.

NITRD(2021), SUPPLEMENT TO THE PRESIDENT’S FY2022 BUDGET.

NSTC(2022) Desirable Characteristics of Data Repositories for Federally Funded Research, The National Science and Technology Council.

OECD(2021), Recommendation of the Council concerning Access to Research Data from Public Funding, OECD Legal Instruments, OECD.

OS4OS(2022) Where is Open Science in Horizon Europe, https://openscience.eu/Open-Science-in-Horizon-Europe (접속일: 2022.6.1.)

OSTP(2021) 2021 Public Access Congressional Report, https://www.whitehouse.gov/wp-content/uploads/2022/02/2021-Public-Access-Congressional-Report_OSTP.pdf

Sarker, I.H.(2021), Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science. 2, 160. https://doi.org/10.1007/s42979-021-00592-x

Science Europe(2020), Implementing Research Data Management Policies Across Europe: Experiences from Science Europe Member Organisations

Tenopir, C., Talja, S., Horstmann, W., Late, E., Hughes, D., Pollock, D., Schmidt, B., Baird, L., Sandusky, R., & Allard, S. (2017), “Research Data Services in European Academic Research Libraries,” LIBER Quarterly: The Journal of the

Association of European Research Libraries, 27(1), 23-44. https://doi.org/10.18352/lq.10180

University of California Berkeley Library,“Elsevier, Springer Nature, and AAAS: Publisher Research Data Policies”, https://update.lib.berkeley.edu/2017/05/04/elsevier-springer-nature-aaas-publisher-research-data-policies/ (작성일: 2017.5.4., 접속일: 2022.4.5.)

UK Biobank(2015), UK Biobank ethics and governance framework.

U.S. Department of Commerce(2021), Commerce Data Strategy : Fiscal Years 2021-2024.

U.S. Federal Data Strategy Team(2019), Federal Data Strategy, https://strategy.data.gov/

U.S. Federal Data Strategy, https://strategy.data.gov/2021/action-plan/ (접속일: 2022.5.1.)

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.(2016), Th FAIR Guiding Pridnciples for scientific data management and stewardship, Scientific Data 3, 160018.

Yoon, A. and Schultz, T.(2017), “Research Data Management Services in Academic Libraries in the US: A Content Analysis of Libraries’ Websites,” College & Research Libraries, 78(7), https://crl.acrl.org/index.php/crl/article/view/16788/18346

Zhang, L, Downs, RR, Li, J, Wen, L and Li, C. (2021) A Review of Open Research Data Policies and Practices in China. Data Science Journal, 20: 3, pp. 1-17. DOI: https://doi.org/10.5334/dsj-2021-003

3. 국내외 법령

(국내)

국가연구개발혁신법, 법률 제18645호 (2020.6.9. 제정, 2021. 12. 28. 일부개정)

국가연구개발사업의 관리 등에 관한 규정, 대통령령 제31297호 (2001.12.19. 제정, 2020.12.129. 폐지)

과학기술분야 연구개발사업 처리규정, 과학기술정보통신부훈령 제193호 (2013.6.10. 제정, 2022.1.14. 일부개정)

국가연구개발정보처리기준, 과학기술정보통신부고시 제2020-102호 (2020.12.21. 제정)

공공데이터의 제공 및 이용 활성화에 관한 법률, 법률 제17344호 (2013.7.30. 제정, 2020.6.9. 타법개정)

데이터기반 행정 활성화에 관한 법률, 법률 제17370호 (2020. 6. 9. 제정)

국가지식정보의 연계 및 활용 촉진에 관한 법률, 법률 제18197호 (2021. 6. 8. 제정)

산업 디지털 전환 촉진법, 법률 제18692호 (2022. 1. 4. 제정)

데이터 산업진흥 및 이용촉진에 관한 기본법, 법률 제18475호 (2021. 10. 19. 제정)

생명연구자원의 확보·관리 및 활용에 관한 법률, 법률 제16016호 (2009.5.8. 제정, 2018. 12. 24. 일부개정)

생명윤리 및 안전에 관한 법률, 법률 제17783호 (2004.1.29. 제정, 2020. 12. 29, 일부개정)

개인정보보호법, 법률 제16930호 (2011.3.29. 제정, 2020. 2. 4. 일부개정)

(해외)

OECD (2021) Recommendation of the Council concering Access to Research Data from Public Funding (공공자금으로 산출된 연구데이터의 접근성에 관한 권고) 개정안

UNESCO (2021) Recommendation on Open Science (오픈사이언스권고)

유럽연합 PE 85 2021 REV 1, REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ON EUROPEAN DATA GOVERNANCE AND AMENDING REGULATION (EU) 2018/1724 (DATA GOVERNANCE ACT)

유럽연합 COM(2022) 68 final, 2022/0047 (COD), Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on harmonised rules on fair access to and use of data (Data Act)

미국 Office of Science and Technology Policy (Feb. 22, 2013) Memorandum for Heads of Executive Departments and Agencies, "Increasing Access to the Results of Federally Funded Scientific Research"

미국 Office of Management and Budget, Circular A-110, "Uniform Administrative Requirements for Grants and Other Agreements with Institutions of Higher Education, Hospitals and Other Non-Profit Organizations (11/19/1993) (further amended 09/30/1999, Relocated to 2 CFR, Part 215

프랑스 디지털 공화국법(Loin° 2016-1321 pour une République Numérique)

일본 부정경쟁방지법(不正競争防止法)
https://elaws.e-gov.go.jp/document?lawid=405AC0000000047

4. 웹사이트

(국내)

과학기술정보서비스플랫폼 Science On. https://scienceon.kisti.re.kr/main/mainForm.do (2022.5.31.접속)

국가연구데이터플랫폼 Data On, https://dataon.kisti.re.kr/ (2022.5.31.접속)

지오 빅데이터 오픈 플랫폼, https://data.kigam.re.kr/#(2022.5.11.접속)

지오데이터 홈페이지, https://geodata.kr/(2020.4.11.접속)

지질자원데이터센터,
https://www.kigam.re.kr/menu.es?mid=a10405020000(2022.5.11.접속)

한국핵융합에너지연구원 홈페이지,
https://www.nfri.re.kr/kor/pageView/92(2022.4.22.접속)

KIST 소재연구데이터플랫폼 구축사업 웹페이지,
https://csrc.kist.re.kr/bbs/board.php?bo_table=m01_03&wr_id=11(2022.5.12.접속)

K-BDS 홈페이지, https://kbds.re.kr/web/guest/version(2022.5.12.접속)

(해외)

All of Us research program 홈페이지, https://allofus.nih.gov/(접속일 : 2022.04.08.)

CENDI, https://cendi.gov/projects/Public_Access_Plans_US_Fed_Agencies.html (접속일: 2017. 7. 12.)

CERN Open data portal
http://opendata.cern.ch/search?page=1&size=20&type=Software&type=Dataset(2022.4.26.접속)

Center for Open Science, “ Top Guidelines”,
https://www.cos.io/initiatives/top-guidelines

CODE, 미국 국가기술표준원(NIST) 오픈소스플랫폼, https://code.nist.gov/ (접속일:2026.2.)

DOE CODE, 미국 에너지부(DOE) 오픈소스플랫폼, https://www.osti.gov/doecode/ (접속일:2022.6.2.)

Datacite 블로그 (2016.4.13. 작성), https://blog.datacite.org/re3data-reaches-a-milestone-and-begins-offering-badges/#:~:text=re3data.org%20has%20reached%20a,disciplines%20from%20around%20the%20world. (접속일: 2022.3.31.)

GDC Data Portal 홈페이지 https://portal.gdc.cancer.gov/(2022.4.2.접속)

Getting Started with ALICE Open Data, https://opendata.cern.ch/docs/alice-getting-started#start(2022.4.21.접속)

Fairsharing, 통계, https://fairsharing.org/summary-statistics (접속일: 2022.3.31.)

JISC 홈페이지, https://v2.sherpa.ac.uk/view/funder_visualisations/1.html, 접속일: 2022.4.6.

Model Catalogue, 미국 지질조사국(USGS) 과학모델 공유 플랫폼, https://data.usgs.gov/modelcatalog/ (접속일:2022.6.2.)

NITRD 웹사이트, https://www.nitrd.gov/about/ (접속일: 2022.4.1.)

NCI TCGA 사업 홈페이지, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga(2022.4.2.접속)

NCI TCGA 사업 홈페이지, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga(2022.4.2.접속)

NCI TCGA Compuatational Tools 목록 https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/tools(2022.4.2.접속)

NSF 홈페이지, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1946932&HistoricalAwards=false (접속일: 2022.4.1.)

Our World in Data 웹 자료, https://ourworldindata.org/technological-progress (접속일: 2022.3.28.)

Research Data Alliance 홈페이지, rd-alliance.org (접속일: 2021.5.6.)

Research Data Alliance 홈페이지, https://www.rd-alliance.org/about-rda (접속일: 2022.3.28.)

RDA 메타데이터 표준 목록, https://rd-alliance.github.io/metadata-directory/standards/ (접속일: 2022.5.20.)

Real-time Linked Dataspaces - Common European Data Spaces, http://dataspaces.info/common-european-data-spaces/#page-content (접속일: 2022.4.13.)

re2data coref 블로그 (2022.3.15. 작성), https://coref.project.re3data.org/blog/mapping-the-global-repository-landscape (접속일: 2022.3.31.)

ROARMAP 홈페이지(http://roarmap.eprints.org/, 접속일: 2020.4.6.)

SPARC, http://researchsharing.sparcopen.org/data (최종 작성일: 2016.2.2., 접속일: 2018.9.10.)