데이터 기반 R&D 지원을 위한 연구자의 학술정보 및 데이터 요구 분석 연구*

A Study on the Scholarly Information and Data Requirements of Researchers for Data-Driven Research and Development

이 석 형 (Seok-Hyoung Lee)**, 이강산다정 (Kangsandajung Lee)***
김 재 훈 (Jayhoon Kim)****, 이 혜 진 (Hyejin Lee)*****

목 차

초 록

본 연구에서는 연구자의 데이터 기반 R&D를 효율적으로 지원하기 위해 새로운 학술정보유형과 데이터셋을 발굴하고, 학술정보서비스의 방향을 제시하기 위한 선행 연구로서 연구자가 필요한 학술정보와 데이터 요구사항을 분석하였다. 이를 위해 관련 연구자 5인의 탐색적 사례 연구와 ScienceON 이용자의 온라인 설문을 통해 데이터 기반 R&D 행태 및 정보·데이터 요구사항을 도출하였다. 그 결과 데이터 기반 연구를 수행하는 연구자들은 학술논문을 많이 활용하며 데이터셋이나 소프트웨어 정보 또한 학술논문이나 학술회의자료로부터 참조하는 것으로 나타났다. 또한 주제 분야별로 활용하는 데이터 확보 방법, 획득 경로와 활용 데이터 유형이 차이가 있으며, 연구자들은 필요한 데이터셋이나 학습모델과 같은 소프트웨어가 어디에 있고 어떻게 확보해야할지 모르는 경우가 많아 연구를 수행하는데 애로사항이 많은 것으로 나타났다. 향후 데이터 기반 R&D를 지원하기 위해 주제별로 데이터셋을 체계적으로 구축해야할 필요가 있으며, 학술논문과 연계하여 데이터셋과 관련 소프트웨어 정보를 별도로 추출·요약해서 제공하는 방안을 고려해야 할 것으로 분석하였다.

ABSTRACT

In this study, as a preliminary research to effectively support data-driven R&D of researchers, we analyzed the academic information and data requirements for researchers to discover new types of academic information and datasets, and to propose directions for academic information services. To achieve the research objectives, we conducted an exploratory case study involving five researchers and administered an online survey among ScienceON users to glean insights into data-driven R&D behaviors and information/data requirements. As a result, researchers relatively referred to academic papers, datasets and software information from academic papers or conference materials. Moreover, the methods and pathways for acquiring data, as well as the types of data, varied across different subject areas. Researchers often faced challenges in data-driven R&D due to difficulties in locating and accessing necessary datasets or software such as learning models. Therefore it has been analyzed that for future support of data-driven R&D, there is a need to systematically construct datasets by subject. Additionally, it is considered necessary to extract and summarize dataset and related software information in conjunction with academic papers.

키워드: 학술정보요구, 데이터요구, 데이터 기반 R&D, 이용자 요구 분석, 탐색적 사례 연구
Academic Information Requirement, Data Requirement, Data-driven R&D, User Requirement Analysis, Explorative Case Study

* 이 논문은 2023년도 한국과학기술정보연구원(KISTI)의 기본사업과제 "지능형 과학기술정보 큐레이션체제 구축" (K-23-L01-C01-S01)으로 수행한 내용의 일부를 재구성한 것임.
** 충남대학교 문헌정보학과 부교수(skyi@cnu.ac.kr / ISNI 0000 0004 6771 8906) (제1저자)
*** 한국과학기술정보연구원 디지털큐레이션센터 선임연구원(lksdj@kisti.re.kr / ISNI 0000 0004 7775 6308) (공동저자)
**** 한국과학기술정보연구원 디지털큐레이션센터 책임연구원(jay.kim@kisti.re.kr / ISNI 0000 0004 6478 844X) (공동저자)
***** 한국과학기술정보연구원 디지털큐레이션센터 책임연구원(hyejin@kisti.re.kr / ISNI 0000 0004 6490 0147) (교신저자)
논문접수일자: 2024년 1월 22일 최초심사일자: 2024년 1월 26일 게재확정일자: 2024년 2월 7일
한국문헌정보학회지, 58(1): 255-283, 2024. http://dx.doi.org/10.4275/KSLIS.2024.58.1.255

1. 서 론

1.1 연구의 필요성 및 목적

4차 산업혁명 시대가 도래한 이후 과학기술은 물론 사회경제 전 분야가 인공지능과 빅데이터 기반으로 빠르게 변화하고 있으며, 정부부처, 공공기관, 민간기업 등은 데이터에 기반한 디지털 전환(Digital Transformation)을 통해 업무 프로세스와 비즈니스 모델의 혁신을 도모하고 있다(유주현, 조상민, 김동현, 2017). 이러한 흐름 속에 연구개발(R&D) 패러다임도 과거 실험, 이론, 컴퓨팅 자원 중심의 연구에서 데이터 중심의 연구로 변화가 가속화되고 있다. R&D 패러다임의 변화에 따라 연구자들이 데이터 기반 R&D 수행이 증가하면서 이를 효율적으로 지원할 수 있는 학술정보 및 데이터에 대한 요구 또한 증가하고 있다.

정보센터나 도서관에서 연구자의 R&D활동을 효율적으로 지원하기 위해 이용자의 요구사항을 분석하고 서비스 개발에 반영하기 위한 대표적인 연구는 권나현, 이정연, 정은경(2012)의 생명공학 및 나노 분야 연구자를 대상으로 질적연구방법론을 적용한 R&D 라이프사이클 규명에 대한 연구가 있다. 이 연구에서는 연구개발의 시작에서 종료, 결과 확산까지 각 단계별 세부 활동에 대한 이해를 바탕으로 R&D 라이프사이클 단계별로 필요한 학술정보 유형을 정의하고 향후 학술정보서비스에 고려해야 할 기능을 도출하였다. 또한, ScienceON의 경우 연구자의 R&D를 효율적으로 수행하기 위해 필요한 각종 학술정보와 컴퓨팅자원, 분석 도구 등의 과학기술 지식인프라를 통합적으로 제공하기 위해 이용자 로그 정보와 설문을 통한 이용자 행태 분석을 수행하고 서비스 기능을 도출한 사례(아주대학교 산학협력단, 2018)가 있다.

그러나, 현재 대규모 데이터를 처리하고 분석할 수 있는 소프트웨어를 활용하여 R&D 성과 목표를 달성하는 형태의 데이터 기반 R&D를 효율적으로 지원하기 위한 이용자 분석 사례는 없는 것으로 판단된다. 앞서 언급한 기계학습데이터 활용서비스, 분류기반 R&D분석, DataON의 분석플랫폼 등의 서비스 기능들은 연구자의 데이터 기반 R&D 활동을 지원하기 위해 데이터의 제공 범위를 확대하거나 분석 인프라를 체계적으로 구축한 노력의 결과물이라고 볼 수 있다. 그러나 실제 데이터 기반 R&D를 수행하는 연구자가 연구 단계별로 필요한 정보나 데이터가 무엇이고 어떠한 절차나 방법으로 데이터를 획득하는지에 대한 요구 분석 없이 서비스를 제공하고 있어 연구자들이 효율적으로 R&D 활동을 수행하기 어려운 측면이 있는 것으로 판단된다.

이에 본 연구에서는 데이터 기반 R&D 활동을 효율적으로 지원하기 위해 연구자의 정보・데이터 요구 분석을 통해 향후 학술정보・데이터 구축 범위의 확장과 학술정보서비스 방향을 제시하기 위한 선행 연구를 수행하였다. 데이터를 활용한 연구개발 과정에서 주로 활용하거나 부족하다고 판단되는 학술정보의 유형, 데이터셋과 인공지능 학습모델, 분석도구 등과 같은 소프트웨어를 인지하고 획득하는 방법과 경로 등을 파악하여 시사점을 도출하는 것이 본 연구의 목적이며 이를 통해 앞으로 신규로 발굴해야 할 학술정보 유형과 정보・데이터원을 설정하고 학술정보서비스에서 고려해야 할 요

소를 도출할 것이다.

1.2 연구 방법

본 연구의 목적을 달성하기 위해 우선 데이터 기반 R&D를 수행하는 연구자 5명을 선정하여 질적사례연구를 진행하였다. 5명의 연구자를 대상으로 탐색적 사례 연구를 통해 R&D 과정에서 활용하는 학술정보 유형과 활용 범위를 파악하고 데이터 기반 R&D 행태 및 정보·데이터 요구사항을 파악하였다. 그 결과를 토대로 한국과학기술정보연구원에서 운영하고 있는 ScienceON 이용자를 대상으로 온라인 설문을 실시하여 데이터 기반 R&D 행태 및 요구 사항을 분석하고 시사점을 도출하였다.

단, 본 연구는 전체 연구분야중 일부 연구분야(생리학, 전기전자, 기계, 컴퓨터, 행정학)의 데이터 기반 R&D를 수행하는 연구자만을 대상으로 탐색적 사례 연구를 수행하였으므로 그 결과를 해당 연구분야나 전체 연구분야에까지 일반화하기는 어려운 한계점이 있다.

2. 선행 연구

본 연구와 관련된 선행연구들은 주로 연구자의 정보이용행태를 분석하여 정보서비스의 개발 방향을 도출하기 위한 목적으로 수행되어 왔다.

권나현, 이정연, 정은경(2012)은 국내 생명 및 나노과학기술 연구자를 중심으로 과학기술 분야 R&D전주기를 규명하고 그 과정에서 나타나는 과학자들의 연구와 정보행동을 조사하였다. 국가 R&D전주기를 아이디어 생성 및 개발, 연구지원비 확보, 실험 및 분석, 성과 창출, 평가 등 총 5단계로 모형화하였으며 각 단계별 주요 연구활동과 특징적 정보행동을 정의하였다. 이를 통해 R&D전주기별로 필요로 하는 정

〈표 1〉 연구절차 및 내용

단계	연구방법	세부연구내용
1단계 이론연구	문헌연구	• 연구자의 정보·데이터 이용행태 관련 선행연구 분석
↓	↓	↓
2단계 사례조사	탐색적 사례 연구	• 데이터 기반 R&D를 수행하는 연구자 심층인터뷰를 통한 데이터 기반 R&D 연구행태 분석
↓	↓	↓
3단계 사례조사	설문조사	• ScienceON 이용자 대상으로 데이터 기반 R&D 연구행태 설문조사
↓	↓	↓
3단계 시사점 도출	설문분석	• 문헌연구, 탐색적 사례 연구, 설문조사 결과의 심층 분석을 통한 시사점, 종합 결론, 향후 연구 주제 제안

보를 효율적으로 제공하기 위한 정보지원기관의 시스템 개발과 서비스 수월성 제고 방향을 제시하였다.

한종엽, 서만덕(2014)은 해양과학기술분야 연구자의 정보이용행태를 규명하여 연구자 개인적 특성에 따른 차별화된 정보서비스 수립과 전문도서관 서비스 고도화를 위한 기초자료를 확보하고자 하였다. 설문지기법을 통해 연구자가 선호하는 정보유형은 전자자료 유형의 해외 학술논문이며 인터넷정보원과 소속 도서관 이용을 통해 정보를 입수하는 것으로 확인하였다. 또한, 자료 수집시 겪는 문제점으로 소속도서관의 전자자원 다양성 부족과 유료정보에 대한 이용 부담으로 나타났다. 향후 중점적으로 고려해야 할 서비스로 맞춤형 정보검색서비스, 프로젝트지원서비스, 연구동향분석서비스를 도출하였다.

이란주, 김수진(2015)는 주제별 연구자의 정보이용행태에 관한 선행 연구 분석에서 과학기술, 사회과학, 그 외 기타 분야의 연구방법, 대상, 변인, 결과 등을 분석하여 다양한 학문 분야 및 전문분야 연구자들의 정보이용행태별로 이용자 요구를 반영한 정보서비스 구현을 위한 기초자료를 제공하고자 하였다. 특히 분석 결과 각 분야의 연구자들은 디지털 환경에 따른 정보이용의 변화를 보였으며 주제분야, 신분 및 직위 등에 따라 정보 요구 및 이용 행태가 다르게 나타나 이용자의 특성을 고려한 차별화된 정보서비스를 제안하였다.

심윤희, 김지현(2019)은 대학도서관 연구데이터 관리서비스 개발 방안에 관한 연구에서 연구데이터를 관리・공유・이용할 수 있는 연구데이터 리포지터리 서비스 기능을 정의하기 위한 자관 소속 연구자를 대상으로 요구 분석을 수행하였다. 요구분석을 통해 도출된 연구데이터관리서비스 방안은 크게 연구비제공기관 및 학술저널에서의 데이터 제출 의무사항, 데이터 기술, 데이터 관리에 대한 교육을 제공하는 교육서비스, 연구데이터 분석・활용・라이선스 관리 등을 지원하는 컨설팅 서비스, 연구데이터 공유와 관련한 큐레이션 기술 서비스 등을 구현하는 것이었다.

김나연, 정은경(2020)은 사회과학 분야 연구자의 데이터요구와 데이터 재이용 행위에 관한 연구에서 연구자 12명과의 심층면담을 통해 데이터를 재이용하는 요인을 개인적 차원에서는 연구의 용이성과 데이터의 접근성, 경제적 차원에서는 연구의 효율성, 기술적 차원에서는 정보기술 환경의 변화, 사회적 차원에서는 학술적 풍토의 변화와 연구분야 및 주제 특성에 의해 데이터가 재이용되는 특성을 도출하였다. 또한 데이터를 파악하고 습득하는 주요 경로로 웹 기반의 정보원과 학술공동체 내의 비공식적 커뮤니케이션임을 확인하였으며, 데이터 재이용시에 발생하는 학술연구자의 데이터 요구로 생산단위는 기관을, 언어는 영어, 국가는 미국을 선호하는 것으로 분석하였다.

나은엽(2023)은 자연과학분야 연구자들의 연구계획서 작성을 위한 정보추구행동의 탐색적 연구에서 연구자들이 펀딩을 획득하기 위해 연구계획서를 작성할 때 발생하는 정보요구 및 정보추구행동과 정보행동에 영향을 미치는 요소를 파악하고 연구자 펀딩 지원을 돕기위한 정보서비스 전략 수립의 근거를 마련하고자 하였다. 연구계획서 작성이라는 과업중심 정보추구행동을 바탕으로 자연과학분야 6인의 교수를

대상으로 질적연구방법인 심층면담과 다이어리 연구를 통해 연구계획서 작성과 관련하여 하위업무를 도출하였다. 그 결과 커뮤니케이션 관련 업무, 펀딩 지원 관련 연구설계 관련 정보요구가 발생하였고 정보요구 해결을 위해 인적자원, 문서, 온라인 데이터베이스, 웹사이트가 주된 정보원으로 이용되고 있음을 확인하였다.

심원식 외(2023)는 개방형 연구 커먼스에 대한 연구자 요구 분석에 관한 연구에서 개방형 연구 커먼스의 구현을 위한 연구데이터 플랫폼에 대한 연구자 요구를 분석하였다. 오픈사이언스의 오픈데이터를 지원하기 위한 연구데이터 플랫폼 구성을 위해 연구데이터의 공유활용, 분석도구의 공유활용, 컴퓨팅자원의 활용, 연구데이터 커먼스의 필요성 등 5개 영역에 대해 설문지법을 통해 요구사항을 조사하였다. 그 결과 연구자들은 연구데이터 커먼스 서비스 제공시 연구자가 보유하고 있는 연구데이터와 분석도구를 공유하고자 하는 의향(각각 약77%, 84%)을 보였으며 공유된 연구데이터와 분석도구, 컴퓨팅 자원을 활용하고자 하는 의향(각각 92%, 93%, 93%)도 매우 높았음을 확인하였다.

Zenk-Möltgen et al.(2018)은 사회학 및 정치학 분야 연구자들의 데이터 공유 행태에 영향을 미치는 요인 연구에서 해당 학문 분야의 주요 저널에 대한 특징을 분석하고 연구자들의 데이터 공유에 대한 동기를 조사하였다. 그 결과 저널의 영향력과 역사, 명시적인 데이터 공유 정책이 데이터 공유 가능성에 영향을 미치고 연구자가 과거에 데이터를 공유한 경험, 사회적 규범, 연구개발 가이드라인 등이 데이터 공유 의도에 영향을 미친다는 점을 확인하였다.

이상의 선행연구들의 특징은 학술정보서비스, 연구데이터플랫폼의 개발 및 고도화를 위한 이용자 요구 기반의 서비스 기능 설계와 개인형 맞춤정보서비스 개발 등을 목적으로 이용자 요구 분석을 수행한 점을 들 수 있다. 권나현, 이정연, 정은경(2012)은 R&D전주기별로 연구자가 필요한 정보 요구에 대한 분석을 수행하였으나 데이터 확보 및 활용과 관련한 요구분석 보다는 학술정보 요구에 대한 분석에 초점을 맞추고 있다. 김나연, 정은경(2020)은 연구자의 데이터 요구 분석 내용에 데이터 획득 경로에 대한 분석이 있으나 사회과학분야 연구자 대상인 점, 심원식 외(2023)의 연구는 개방형 연구데이터 커먼스 구축에 대한 필요성을 도출하기 위한 인프라 측면에서의 요구 분석이라는 점에서 국내외에서 데이터 기반 R&D를 수행하는 연구자의 요구 분석을 수행한 연구는 부족하다고 할 수 있다.

3. 사례 연구를 통한 정보·데이터 요구 분석

본 절에서는 데이터 기반 R&D를 수행하는 연구자 5명을 선정하여 연구자의 연구행태 및 콘텐츠 요구사항을 분석하기 위해 연구자들이 연구목적을 달성하는데 필요한 콘텐츠 유형 및 콘텐츠를 획득하기 위한 방법, 획득시 고려사항 등을 인터뷰를 통해 확인하였다.

3.1 조사분석 방법 및 절차

본 연구의 목표를 달성하기 위한 첫 번째 단

계로 질적 사례 연구를 수행하였다. 이를 위해 5명의 데이터 기반 R&D를 수행하는 연구자에 대해 탐색적 사례 연구(Explorative Case Study)를 진행하였다. 탐색적 사례 연구는 결과물이 불분명한 사례(주로 프로그램)가 평가 대상이 되었을 때나, 대상이 확실하지 않고 다양하게 영향을 미친다는 판단이 드는 요소가 가미된 사례를 탐색할 때 쓰이는 연구 방법이다. Merriam (1998)에 의하면 사례 연구는 구체적인 상황이나 현상이 갖는 특징과 의미에 초점을 두고 사례에 대한 풍부하고 심도 있는 서술을 제공하는 특성을 가지고 있다. 따라서 데이터 기반 R&D를 수행하는 연구자의 R&D 행태를 구체적으로 연구한 사례가 많지 않기 때문에 앞서 정의한 연구 문제에 대한 답을 발견해가면서 신규 콘텐츠 발굴 대상과 연구자 제공 콘텐츠 영역을 정의하는 것에 목적을 둔 탐색적 사례 연구로 진행하였다.

그리고 본 연구에서는 다양한 학문 분야 주제 영역을 고려하여 복수의 사례를 선택하는 다중 사례 연구(Multiple Case Study)로 진행하였다. 학문 분야나 주제 영역별로 활용되는 데이터의 유형이나 활용 방법이 다를 것이라고 판단하여 단일 사례 연구보다는 5명의 참여자를 통해 주제 영역별로 데이터 기반 R&D 사례를 분석하는 다중 사례 연구 방법을 사용하였다.

본 절의 질적 연구 수행을 위해 반구조화된 면담가이드에 의한 면담(semi-structured interview) 형식의 심층면접(in-depth interview) 유형을 채택하여 사용하였다. 반구조화 면담가이드에 의한 면담은 개방형 질문을 사용하여 더 깊게 피면담자의 반응을 이끌어내고 상황에 따라 질문의 순서나 속도, 폭과 범위를 유연하게 변화하는데 유용하다. 심층면접은 일대일 대화 방식으로 1시간 이상 지속되어 데이터 기반 R&D를 수행하기 위한 방법과 절차 등을 확인하였다.

피면담자는 특정한 특징이나 인구통계학적 특성을 가지는 사람들을 대상으로 하는 의도성(Purposive)에 기반한 표본을 설정하여 모집하였다. 한국과학기술정보연구원에서 매년 선정하는 지식창조대상(한국과학기술정보연구원, 2022) 수상자들의 성과물을 분석하여 데이터 기반 R&D를 체계적으로 수행하고 있다고 판단된 수상자들을 선별하여 유선 전화로 문의하여 조사에 참여하겠다는 의사를 밝힌 연구자를 대상으로 4인을 우선 선정하였다. 그리고, 4인의 연구자가 이공계열임을 고려하여 사회과학 분야에서 데이터를 분석하여 연구를 수행하는 빈도가 높은 연구자를 추가 섭외하여 총 5인의 연구자를 대상으로 면담을 진행하였다. 면담은 본 연구의 저자 중 2인이 참여의사를 밝힌 피면담자의 연구실로 직접 방문하여 진행하였고 면담시간은 평균 약 1시간 10분이 소요되었다.

일대일 심층면담을 위한 반구조화된 질문지는 〈표 2〉와 같이 정의하고 인터뷰가 진행되는 동안 개별적인 면담 상황, 피면담자의 답변 방향에 따라 질문이 조금씩 변경되었다. 질문영역은 크게 피면담자의 개인적 배경, 데이터 기반 R&D를 수행하는 연구적 배경, 정보와 데이터 등 콘텐츠의 추구 행태와 관련된 내용으로 구성되어 있다. 개인적 배경은 연구자의 소속 학과와 전공, 학력, 연구활동기간, 주요연구주제에 대한 내용이고, 연구적 배경은 최근 주요 연구내용, 연구주제 선정 및 연구추진 배경, 해당분야에서 연구 내용의 중요성과 의의, 학술

〈표 2〉 심층면담 질문내용

면담 요소	면담 질문 내용
개인적 배경	• 연구자의 소속 학과 • 연구자의 전공 및 학력 • 연구활동기간 • 주요연구분야
연구적 배경	• 최근주요연구내용 • 연구주제선정 및 연구추진배경 • 해당분야에서 연구내용의 중요성과 의의 • 학술커뮤니케이션 활동 여부
콘텐츠 추구 행태	• 연구단계별 혹은 연구과정에서 주로 이용하는 콘텐츠 • 콘텐츠의 특성 • 콘텐츠 입수 방법 및 절차 • 정보·데이터 활용 내용 • 정보·데이터 수집 및 활용시 고려사항

커뮤니케이션 활동 여부 등이 포함되어 있다. 콘텐츠 추구 행태는 데이터 기반 R&D를 수행하는 과정에서 주로 이용하는 콘텐츠의 특성, 콘텐츠 입수 방법 및 절차, 고려사항 등을 포함하고 있다.

3.2 피면담자 특성 분석

본 연구에 참여한 피면담자는 〈표 3〉과 같다. 개별 피면담자 각각을 A~E까지 기호를 부여하였으며 기호 순서는 인터뷰 순서와 동일하다. 피면담자는 모두 남성이었으며, 연령은 20대 1명, 30대 2명, 40대 1명, 50대 1명이고, B는 석사과정 연구원이며 나머지 피면담자는 박사학위를 보유하고 있었다. 피면담자의 전공분야 및 주요연구분야는 피면담자 A는 생리학분야의 나노독성 및 세포치료, B는 전기전자공학분야의 자율주행자동차, C는 기계공학분야의 이차전지, D는 컴퓨터공학의 언어처리 등 인공지능기술개발, E는 행정학의 공공행정이다. 피면담자의 전공 관련 연구기간은 7년부터 40년까지 분포되어 있으며 평균 약21년이었다. 피면담자 A와 C는 사립대 교수로 재직중이었으며 D와 E는 국립대 교수로 재직하고 있었다.

〈표 3〉 피면담자 정보

구분	성별	연령	학력	전공분야	주요연구분야	연구기간	직책
A	남	50대	박사	생리학	나노독성, 세포치료	40년	사립대 교수
B	남	20대	학사	전기전자공학	자율주행자동차	7년	사립대 석사과정
C	남	30대	박사	기계공학	이차전지	19년	사립대 교수
D	남	40대	박사	컴퓨터공학	언어처리	27년	국립대 교수
E	남	30대	박사	행정학	공공행정	15년	국립대 교수

3.3 연구적 배경 분석

피면담자 A는 DNA, RNA 분석 등을 통한 펩타이드 치료제 연구를 주된 연구로 수행하고 있었으며 치료제 개발에 필요한 DNA 분석시에 AI 기술을 활용한 데이터 분석이 필요하다고 하였다. B는 자율주행자동차에 탑재될 도로 장애물 인식 기술 개발을 위해 도로 주변 이미지 및 영상 학습데이터를 활용한 도로 장애물 학습모듈 개발 연구를 수행하고 있었다. C는 자동차 베터리 등에 활용되는 이차전지의 성능 향상 연구를 수행하고 있으며, D는 한국어 언어처리 기술 개발과 언어처리 기술 개발 관련 코퍼스 구축연구를 진행하고 있었다. E는 공공행정 및 지방행정의 재정 현상 분석과 재정 건전성, 운영 효율성 등을 분석하는 연구를 진행하고 있었다.

연구주제선정은 연구자의 학력과 전공분야, 관심 주제 등에 따라 이루어지며 특히 피면담자 D의 사례와 같이 최근에는 정부나 기업의 R&D 지원 방향, 사회적 이슈 등과의 관련성도 중요한 요소로 언급하였다. 피면담자 C의 경우 기계공학분야에서 전기자동차가 대두되고 기계공학의 기본 이론이라 할 수 있는 열역학이 이차전지와도 관련성이 높아 이차전지 연구를 본격적으로 진행한 사례로 볼 수 있다.

> “(기계공학과 이차전지의) 관련성이 아주 고전적인 개념인데, 가장 체감될 수 있는 예시가 자동차 엔진 있지 않습니까? 기계과의 꽃이 원래 엔진이었거든요. 엔진이라 하면 기계, 머신 딱 이런 느낌이었거든요. 그게 사라졌습니다. 옛날에는 그런 바운더리에 대한 컨셉들이 강하던 시절인데 지금은 많이 허물어졌죠. 기계공학 분야에서는 역학을 기반으로 한다는 이야기를 많이 해요. 그런데 전지에서도 그런 일들이 굉장히 많이 일어나고 지금 실질적으로 소재에서도 성능 감소가 일어나는 원인을 역학적인 부분에서 찾으려고 합니다. 그러니까 역학이론 이런것들이 이차전지나 소재분야에서도 적용 가능하구요.” (피면담자 C)

> “요즘 인공지능 키워드가 없으면 과제도 따기 어렵고 학생들도 안들어와요. chatGPT 얘기 많이 하잖아요. 네이버, 카카오도 연구하고 있고, 그런 분위기를 무시할 수 없죠. 저도 언어처리로 학위를 받았고 지금 대세가 언어모델이니까 특히 한국어 처리는 (연구)할게 많죠.” (피면담자 D)

피면담자 중 C,D,E는 R&D 라이프사이클의 아이디어 생성 및 개발, 펀딩 파악 및 확보, 실험 및 분석, 성과 창출 및 평가 과정에 따라 연구를 수행하는 것으로 조사되었다. 따라서, 연구추진배경에 연구과제 수주와 수주된 연구과제에 기반한 R&D활동이 이루어진다고 볼 수 있다.

> “요즘 국가차원에서 이차전지에 대한 연구지원이 많고 저도 대학원생들하고 과제를 제안하고 수행하면서 이 연구를 진행하고 있는거죠. (주제가 이차전지이니까 민간 업체랑 또 많이 하시겠네요. 이차전지 생산업체라던지) 네 뭐 기업과제도 많고 국가 과제도 있고 되게 다양하게 브로드하게 많습니다.” (피면담자 C)

반면 B는 석사과정 연구원으로 아이디어 생

성 및 개발과 펀딩 파악 및 확보는 연구실의 박사과정 혹은 지도교수가 주로 수행하며, 실험 및 분석 단계에서의 연구를 수행하는 것으로 나타났다.

"전체적으로 처음에 이제 과제 세팅은 교수님께서 하시는데 이후에 실제 과제가 진행되는 과정 중에는 파트가 또 여러 개로 나눠지는 부분이고, 그러면 파트별로는 석사 과정에 계신 분들이 나눠서 주로 연구 개발하는 형태로 진행이 됩니다. 그리고 과제 제안할때는 박사 과정분들이 같이 정리하는 것으로 알고 있습니다." (피면담자 B)

피면담자 A는 R&D과제를 수주하여 연구를 진행한다기보다는 기존 연구논문 동향을 파악하여 특정 주제를 설정하고 주제 관련 목표를 달성하기 위한 실험 및 분석, 성과 창출 과정에 초점을 맞추어 연구를 수행하는 것으로 나타났다.

"저 같은 경우에는 실제로 대형 과제를 가지고 있어 본 적이 없습니다. 그러면은 데이터를 어디에 썼냐 하면은 공개된 데이터, 논문에서 데이터를 공개하도록 되어 있거든요. 그걸 받아가지고 다시 다양한 머신러닝 알고리즘을 가지고 예측 모델을 만들었을 때 다른 사람들은 예를 들어 78%라면은 어떻게 해서든지. 아까 펩타이드라든지 DNA 규칙성이 강하니까 그걸 잘 추출하고 이러다 보면은 그거 조금 정확도 높은 모델을 만들고 이렇게 했거든요. 논문 연구 내용을 보고 새로운 걸 해보는 게 제 연구 방식이죠." (피면담자 A)

피면담자 E 또한 R&D과제를 수주하여 연구를 진행하고 있는데, 특정 지방자치단체별로 재정위기 현상 분석을 위해 국내외 정세관련 통계데이터와 신문, 보고서 등의 원문을 분석한 결과를 융합하여 재정위기 원인을 파악하고 해결책을 제시하는 연구하는 것으로 나타났다.

"지방분권시대에서 지방재정 환경의 변화는 사회적, 경제적, 정치적으로 다양하고 급격하게 진행되고 있습니다. 특히 코로나 19 이후로 경제활동이 제한되고 위축되는 상황에서 긴급재난지원금 소비 등으로 인해 방재정의 어려움이 커지고 있는 듯합니다. 지방재정은 경기침체, 소비활동 위축, 기업운용의 어려움 등 여러 요인에 의해 변동성이 커지고 있는데 이러한 대내외적 현상을 면밀히 분석하여 건전한 지방재정 운영방안을 제시하는 것이 매우 중요하구요. 그래서 경제지표, 사회현상 등을 데이터화 하여 분석하고 원인을 찾는 것이 연구 과정에서 필요합니다." (피면담자 A)

면담에 참여한 피면담자들 모두 현재 수행하고 있는 연구의 중요성과 필요성에 대해 앞서 언급한 바와 같이 해당 주제 영역에 대한 과제 공고와 연구자의 수가 증가하고 있다는 점, 학문 영역간 경계가 모호해지고 융합연구로 진화하고 있다는 점, R&D 정부정책 및 기업체의 수요가 증가하고 있다는 점 등을 언급하면서 해당 주제에 대한 중요성과 필요성을 강조하였다. 피면담자 A는 펩타이드 치료제 개발 연구가 코로나 19로 인한 펜데믹 시대 이후에 증가하고 있고 바이오 분야에서 특히 인공지능 학습모델에 기반한 DNA 실험으로 논증하는 방

식이 인정받고 있어 앞으로도 데이터 기반의 R&D가 확대될 것이라고 언급하였다.

연구적 배경의 마지막 요소인 학술커뮤니케이션 활동 여부는 오픈사이언스와 데이터 기반의 R&D의 주요 특성이라 할 수 있는 협업 연구 방식으로 연구활동을 수행하는지 여부를 판단하기 위한 것으로 피면담자 A를 제외하고 B,C,D와 E 는 국내외 학술커뮤니케이션 활동을 한다고 응답하였다. 피면담자 A는 참여자가 구성한 연구실을 활용하여 기존 연구결과를 기반으로 새로운 연구 아이디어와 관련 정보를 획득한다고 응답하였고, 연구자 B와 연구자 E는 국내 학회보다는 해외 학회 참석을 통해 연구 아이디어와 관련 정보를 획득한다고 응답하였다. 이는 활용하는 학술정보 및 데이터의 특성과도 관련이 있고 연구경력과도 관련이 있는 것으로 판단되었다.

> “저희가 연구하고 있는 컴퓨터 비전 이쪽은 국내 저널이나 학회가 아무래도 최신 연구보다 기존 연구를 활용할 케이스가 많기 때문에 딜레이가 있다고 해야할까요? 근데 저희 분야가 워낙 발전도 빠르고 변화가 빠른 분야이기 때문에 일단은 해외 컨퍼런스 쪽으로 발표를 하고 저널은 약간 정리된 형태로 해서 성과 위주로 고려한다고 보시면 되겠습니다.” (피면담자 B)

> “저는 학위를 해외에서 받았고 여기 임용되기 전까지 해외에서 있었으니까요. 지도 교수님이나 같이 연구하셨던 분들이 해외에 많이 계시니까 해외 컨퍼런스 중심으로 활동합니다. 그래도 국내로 들어왔으니까 국내 학회에서도 이제 활동을 하려고 합니다.” (피면담자 E)

피면담자들은 R&D 라이프사이클에서 전체 연구 활동 중 데이터 분석 및 실험의 비중이 70% 정도로 동일 분야 타 연구자보다 데이터 분석 연구 비중이 크다고 생각하고 있으며, 현재 주변 연구자들 중에 데이터 기반 연구를 하는 비율은 10명 중 2-3명 정도라고 추측하고 있었다. 그러나, 최근에는 해당 학문 분야에서 신진 연구자를 중심으로 데이터 기반의 R&D 비율이 점차 증가할 것이라고 모든 피면담자가 언급하였다.

피면담자들의 연구적 배경 특징을 종합해보면 학문 분야별 다양한 주제 영역에 대해 R&D를 수행하고 있었으며 모두 데이터를 기반으로 한 연구를 주로 수행하고 있었음을 확인하였다. 그리고, 피면담자 A를 제외하고 국가연구개발사업 및 과제를 수주하거나 향후 수주를 목표로 연구 활동을 수행하였으며, 피면담자 모두 해당 연구의 중요성과 필요성에 대해 인식하고 있음을 알 수 있었다. 또한, 학술커뮤니케이션 활동을 통해 연구 주제 아이디어와 관련 정보 및 데이터를 획득하는 것으로 파악되었다.

3.4 콘텐츠 추구 행태 분석

본 절에서는 데이터 기반 R&D를 수행하는 피면담자의 콘텐츠 추구 행태를 분석하고 시사점을 도출하여 학술정보서비스 이용자 요구분석을 위한 설문지 구성에 참고하고 향후 신규 콘텐츠 발굴 및 학술정보서비스 방향성을 설정하는 것을 목표로 하였다.

면담 결과 피면담자들의 R&D활동에서 참고하는 학술정보 유형은 학술논문과 학술회의자료로 나타났다. 단, 주제 분야별로 참고정보 유

형이 차이가 있었는데, 피면담자 A의 경우 국내학술논문보다는 해외학술논문을 주로 참고한다고 하였다. 그 이유는 생리학분야 주요 해외학술논문은 데이터 공개가 필수이기 때문에 연구 과정에서 데이터의 획득 과정이 국내학술논문보다 용이하다는 점을 언급하였다.

"데이터를 수작업으로 확보를 하는 것보다는 논문이 게재가 되면 특히 IF가 낮은 것보다 Briefings in Biology, BioInformatics라고 어찌 보면 AI에 관계되는 가장 관련성이 높은 분야거든요. (중략) 바이오 분야에서 실제로 AI 기반으로 예측 모델을 하는 거는 Briefings in Biology, BioInformatics이 저널이 임팩트 팩터가 작년에 한 13점, 9점까지 되었는데, 거기는 웹 서버를 만들어 가지고 데이터를 올려놓도록 하고 유저들이 2년간을 사용할 수 있도록 해줍니다." (피면담자 A)

반면 피면담자 B의 경우 학술논문보다 학술대회논문을 많이 활용한다고 하였다. 앞 절에서 컴퓨터 비전 영역은 발전속도와 변화주기가 빠르고 해외 컨퍼런스가 관련 데이터를 일정기간동안 해당 사이트에 올려놓고 공유할 수 있도록 하기 때문에 해외학술회의 논문을 활용할 수 밖에 없다고 하였다. 그 외의 타 피면담자들 모두 국내외학술논문이나 학술회의 자료를 연구과정에서 주로 참고한다고 응답하였는데 피면담자 C와E는 국내외학술논문을 유사한 비율로 참고하는 편으로, D는 국내학술논문과 국내학술회의자료를 많이 참고한다는 응답이었다.

"컨퍼런스 논문 자체가 상호 동료 평가를 거친 것이라 신뢰도가 있다고 보기 때문에 컨퍼런스 발표 결과물을 좀 더 신뢰합니다." (피면담자 B)

피면담자들이 활용하는 데이터 유형은 주제분야별로 다양했다. 피면담자 A는 펩타이드 치료제 개발에 필요한 DNA 정보를 실험데이터로 주로 활용한다고 하였으며, B는 도로 주변 이미지 데이터를 학습데이터로 활용한다고 하였다. 피면담자 C는 외부에서 생성되거나 제작된 데이터를 사용하는 것이 아닌 이차전지 소재 특성 향상 알고리즘을 적용하여 슈퍼컴퓨터를 활용하여 생성한 수치형 데이터를 활용한다고 하였다. 피면담자 D는 주로 말뭉치나 언어처리용 테스트컬렉션, 사전, 웹문서 등 텍스트 데이터를 활용하고, E는 설문데이터를 자체 생성하거나 웹 사이트 등에 게시된 수치형 통계데이터를 많이 활용한다고 하였다.

"제가 하는 실험은 소재 관련된 양자 시뮬레이션이고 소재는 2차전지 소재이거든요. 데이터를 생산해내는 메소돌로지는 이제 양자 계산을 통한 거고, 그 데이터를 생산해내는 하드웨어는 슈퍼컴퓨터를 쓰고 있습니다. 그러면 다양한 정말 무한한 데이터가 지금도 여전히 계속 생산되고 있는 상황인 거죠. 그게 첫 번째 데이터고 두 번째는 이제 실험 데이터입니다. 데이터를 활용해서 전지의 모든 목적은 성능을 높이고 싶어 하는 계산이 어떻게 보면은 가상 데이터거든요. 그냥 말도 안 되는 가상이 아니라 현실에 근사한 데이터이죠. 이 데이터가 그래도 간극이 사실 있거든요. 근데 간극을 줄이기 위해서는 실제 실험 데이터가 필요해서 실험 데이터를 생산해낸 데이터가 있고, 서두에서 말씀드린 첫

번째, 두 번째는 인하우스 자체적으로 생산해내는 데이터 자생형이고 이제 나머지 이제 세 번째는 오픈 DB를 씁니다." (피면담자 C)

피면담자들이 활용하는 데이터의 입수처 또한 다양하였는데, 피면담자 A는 해외학술논문에 언급되어 있는 데이터셋 URL을 참고하거나 논문에 언급되어 있는 데이터 생성 방법을 재현하여 데이터를 확보한다고 하였으며 PubMED에 게시된 데이터도 많이 참고하는 것으로 응답하였다. 특히, 상용 데이터셋에 대한 라이센스를 구매하여 활용하는데 만족도가 높다고 언급하였다.

"저는 진짜 데이터셋 구축 관련해서 롤 모델이 IPA라고 생각하거든요. 분기마다 300명 정도의 사람들이 발표된 논문을 계속 업데이트 시켜가지고. 제가 가장 대표적으로 했던 게 아마 생물학적인 실험을 하고, 제가 만든 빅데이터를 가지고 뭐 RNA 단백체, 대사체 모아가지고 인공지능으로 활성상수를 계산하고 추출하고. 어쨌든 데이터만 뽑아내면 인공지능으로 추출한 데이터를 가지고 변수 상수만 조정하면 복잡하지 않고 단순한 네트워크를 만들수 있죠. 저는 IPA라는 인제니터 패스 어낼리시스라는 프로그램이 가장 좋은 케이스로 생각합니다." (피면담자 A)

피면담자 B는 해외학술회의자료를 통해 데이터셋의 URL을 참고하거나 저자에게 직접 요청한다고 응답하였다. 또한 학회참석을 통해 데이터 소재를 파악하고 해당 웹 사이트를 통해 데이터셋을 다운로드 받아 사용하는데, 최근에 해당 연구분야 관련 데이터셋이 바이두 클라우드, 깃허브(Github), 페이퍼스 위드 코드(Papers with Code) 등에서 많이 게시가 되어 해당 사이트도 많이 참고한다고 하였다.

"데이터 자료를 찾는 방법은 컨퍼런스를 주최하는 주최 기관에서 운영하는 웹 페이지를 통해서 그 회에 공개됐던 자료들을 찾는데 웹 페이지 기반이다 보니까 과거 5년 정도 지난 데이터는 조금 유실이 되는 게 있어요. 그래서 저희 분야의 특징상 연구에 활용하였던 프로그램 코드나 관련 자료 혹은 딥러닝을 하니까 딥러닝 모델의 웨이트 파일이 논문의 성능을 대략적으로 판단하는 데 중요한 목적으로 활용이 되는데 이러한 파일들 같은 경우에는 구글 드라이브나 중국분들이 쓰시는 논문의 경우 바이두 클라우드를 많이 활용을 하는데, 중국 분들은 중국 서비스를 이용하다 보니까 내부에서 주로 이용하시는 웹 서비스를 이용하시더라고요." (피면담자 B)

반면 피면담자 D의 경우 응용 분야에 따라 데이터의 입수처가 다양하다고 하였는데 예를 들어 학술정보를 활용한 자동분류 학습모델을 개발하는 연구를 할 경우 국내외 학술정보의 초록과 분류 데이터를 확보하기도 하며 웹문서의 리뷰에 대한 의미를 분석할 경우 댓글 정보를 직접 수집하여 연구를 수행한다고 응답하였다.

"언어처리용 학습데이터 같은 경우는 직접 구축하거나 연구소 같은데서 공개한 데이터셋을 사용해요. 저희팀은 학술정보 데이터를 가지고 분

류라던지 요약기술을 개발하니까 논문을 크롤링해서 그 안의 초록을 추출하고 정제하고 문장 내용과 의미에 따라 태깅작업을 합니다. 태깅작업은 요즘 인공지능 학습데이터셋 구축도구들 많으니까 그걸 써서 하는 경우가 대부분이죠.. 분야별로 논문을 활용한 학습데이터셋들이 있는데 그런것들은 공개가 되어 있어서 바로 사용할수도 있고 약간 가공해서 사용할 수도 있습니다. 언어처리용이나 영상인식처리 같은 기술을 위해 요즘은 데이터셋 자체에 대한 논문을 많이 발표합니다. 데이터셋과 학습모델을 같이 다루기도 하고 데이터셋에 대해서만 발표하는 경우도 있습니다. 그런 논문들이 나중에 관련 연구를 수행하는데 있어서 활용데이터가 될 수도 있고요. 그런데 국내 논문 같은 경우 github 같은데 공개하는 경우는 많지 않습니다. 그냥 저자한테 연락해서 받거나. 그게 안되면 논문에 나와있는 내용대로 만들어서 적용하기도 합니다." (피면담자 D)

피면담자들이 데이터 기반 R&D를 수행하는데 주요 애로사항 혹은 건의사항으로 데이터셋의 확보문제, 데이터분석 전문 인력의 부족, 데이터셋의 공신력 등을 언급하였다. 피면담자들 모두 특정 주제 영역에 대해 처음 연구를 시작할 때 분석에 필요한 데이터를 확보하는 것이 매우 어렵기 때문에 연구 주제별로 많이 활용되었던 데이터셋과 소프트웨어 URL 혹은 관련 논문 정보를 학술정보서비스에서 제공해 주길 원하는 응답이 많았다.

"저는 연구주제에 대해 분석을 할 때 데이터가 어디에 있는지를 몰라서 연구를 진행하지 못하는 경우가 있습니다. 국가통계포털이나 사회과학자료원 같은 곳을 살펴보아도 제 주제와 관련된 데이터를 쉽게 찾을 수가 없습니다. 너무 오래된 데이터이기도 하고 막상 유사한 걸 찾아도 그 데이터가 쓸모가 없는 경우도 많아요. 제가 원하는 주제를 이야기하면 관련된 데이터 소재를 알려주는 서비스가 정말 필요합니다." (피면담자 E)

데이터셋의 공신력과 관련하여 피면담자들은 공공데이터포털이나 빅데이터포털 등에서 제공하는 데이터셋을 한번쯤은 확인해 보았는데 실제 연구주제와 관련된 데이터가 거의 없고 비슷한 내용의 데이터라도 그 품질이 낮아 활용성이 떨어진다는 응답이 많았다. 또한 상용 데이터셋의 경우 비용이 들기 때문에 과제가 없을 경우 활용할 수 없는 문제가 있기 때문에 학술정보서비스 외연 확대 관점에서 주제영역별로 품질이 보장되고 활용성이 높은 데이터셋 정보를 제공하길 원하는 의견도 많았다.

"근데 그걸 좀 싼 돈 들여가지고 그냥 일부 연구원들 몇 명 모아가지고. 이거 한번 해보자 그거는 저는 데이터의 질이 떨어질 가능성이 많다고 생각합니다. 그래서 앞으로 국가적으로 데이터 셋을 모으려고 하면 예를 들면, biology라든지. 그 원래는 논문을 읽고 그걸 잘 이해할 수 있는 사람이 데이터 셋을 모아요. (중략)..이제 그런 데이터셋을 좀 잘 모을 수 있는 우리나라에 지금 생물학을 많이 하고 고학력자들이 많이 있지 않습니까? 그런 사람들이 중심이 되어서 국가차원에서 데이터셋을 모으는 체계를 만들어야 해요. 우리나라

에 지금 데이터셋 모아주는 기관은 없더라고요.” (피면담자 A)

“모델과 모델이 선언된 파일, 모델에 대한 웨이트 파일에 대한 링크도 같이 잘 알려주는 뭔가가 있으면 좋겠네요. 그리고, 연구분야에 데이터 셋이 어떤 것들이 있고 연구 내용과 연구 데이터셋과 그거에 대한 어떤 성질이라든지 특징 이런 것들을 쭉 이렇게 종합적으로 제공해주는 데이터베이스가 존재한다면 해당 데이터베이스에서 약간 미러링의 형태로 해당 논문이 제공하였던 데이터를 다시 보존을 해줬으면 좋겠네요. 아무래도 논문 보존하는 웹 페이지에서 데이터도 보존해주면 개인이 공유하는 것보다 훨씬 안정적이기 때문에 그 부분이 제일 필요성이 클 것 같습니다.” (피면담자 B)

“ScienceON에서 논문 안에 나와있는 학습데이터 내용과 모델을 예를 들어 모델 유형별, 기본 알고리즘별, 데이터 유형별로 정리해서 보여주면 연구하는 분들이 좋아할 거 같은데요. 저는 이용할거 같아요. 데이터 찾기가 생각보다 쉽지 않아요.” (피면담자 D)

3.5 시사점 및 기타 의견

사례연구를 통한 데이터 기반 R&D를 수행하는 피면담자들은 연구 과정중에 학술논문을 주로 참고하는 것으로 분석되었으며, 주제 분야에 따라 국내외학술논문를 주로 참고하거나 학술회의자료를 참고하는 등의 차이가 있었다. 그리고, 피면담자들은 동일 혹은 유사주제를 다룬 학술논문의 본문안에서 데이터셋과 학습모델에 대한 소재정보를 확인하여 데이터셋을 확보하는 것으로 파악되었다. 공공데이터포털, 주제분야별 전문정보센터, 국가 빅데이터 포털 등 공공데이터를 활용하여 연구를 수행하는 경우는 거의 없었으며 그 원인은 실제 논문 등의 성과와 연계가 되지 않아 데이터의 신뢰성과 정확성을 보장하지 않기 때문이라고 응답하였다.

피면담자들은 데이터 기반 R&D를 효율적으로 수행하기 위해 가장 필요한 것이 주제영역별로 공신력 있는 데이터셋의 소재정보라고 응답하였으며, 관련하여 ScienceON 등 학술정보서비스에서 데이터셋의 소재정보와 학습모델을 주제별로 구분하여 요약 제공한다면 적극 이용할 것이라고 응답하였다. 연구에 필요하고 적합한 데이터셋을 찾는 과정이 생각보다 어렵고 오래 걸리기 때문에 학술정보서비스에서 데이터셋 안내 기능을 제공해 준다면 연구에 많은 도움이 될 것이라고 응답하였다.

기타 의견으로 데이터를 획득하거나 활용할 때 데이터의 최신성과 속보성 등에 우선순위를 두는 피면담자가 있는 반면 데이터의 정확성과 신뢰성이 중요하다는 피면담자의 의견도 있었다. 그리고 데이터의 정확성과 신뢰성을 위해 상용 데이터를 적극적으로 활용할 의사도 있는 것으로 확인되었다. 또한, 데이터의 정확성과 신뢰성을 공공기관이 보증해주는 절차나 방법이 필요하다는 의견도 제시되었다.

4. 설문기법을 통한 정보·데이터 요구 분석

본 절에서는 앞 절에서 수행한 인터뷰 결과

를 바탕으로 도출된 데이터 기반 R&D 행태 및 콘텐츠 요구사항 분석 결과를 토대로 한국과학기술정보연구원(KISTI)의 과학기술지식인프라 통합서비스 ScienceON 이용자를 대상으로 데이터 기반 R&D 행태 및 요구사항을 설문을 통해 분석하고 그 시사점을 도출하였다.

4.1 설문지 구성

본 연구 목적을 달성하기 위한 설문지 구성은 두 가지 요소를 고려하여 설계하였다. 우선 주요 학술정보서비스에서 서비스하고 있는 학술논문, 연구보고서, 특허정보 등 기존 학술정보에 대한 수요 및 요구사항 분석 측면에서의 설문 내용을 포함하였으며, 다른 하나는 데이터 기반 R&D를 수행하는 연구자가 필요로 하는 데이터에 대한 수요 및 요구사항 분석 측면에서 설문 내용을 포함하였다. 이에 따라, 본 연구에서는 설문지를 크게 연구자의 정보 추구 및 탐색 행태 특성, 연구자의 데이터 추구 및 탐색 행태 특성, 연구자의 정보원 이용 특성 등 3개 영역으로 구성하였다.

연구자의 정보 추구 및 탐색 행태 특성 설문 문항을 통해 ScienceON을 이용하는 연구자들의 정보 활용 목적과 주로 사용하는 정보 유형, 향후에 필요성이 높을 것으로 예상하는 정보 유형 등을 살펴보았다. 연구자의 데이터 추구 및 탐색 행태 특성 설문 문항을 통해 데이터의 활용 목적, 데이터 획득 경로, 데이터 활용 유형, 데이터 소재파악 방법, 데이터 활용시 고려사항 등을 확인하였다. 연구자의 정보원 이용 특성은 연구자 본인이나 주변 연구자들이 정보와 데이터를 탐색하고 획득하기 위해 이용하는 정보원의 유형과 기타 정보・데이터 획득 및 활용과 관련한 기타 제언 등을 확인하였다. 특히, 앞서 인터뷰 내용 결과에서 도출되었던 주로 사용하는 데이터 유형, 데이터의 활용 목적, 데이터를 획득하는 방법, 데이터셋의 소재정보를 파악하는 방법 등을 설문문항의 선택항목으로 반영하였다.

질문에 대한 대답은 R&D를 수행하는 과정에서 다양한 유형의 정보・데이터를 활용할 수 있다는 가정하에 정보 및 데이터 유형, 활용 목적, 참고정보원 등을 선택하는 질문에는 최대 3개까지 복수응답을 허용하였다. 설문조사를 실시하기 전에 앞 절에서 참여한 연구자들을 대상으로 설문 내용과 응답 항목에 대한 적절성을 피드백을 받았으며, 그 결과를 바탕으로 일부 설문 항목을 수정하여 최종 설문지를 구성하였다(〈표 4〉 참조).

4.2 자료 수집 및 분석

본 연구에 대한 설문조사는 한국과학기술정보연구원(KISTI)의 협조를 받아 사이언스온(ScienceON) 홈페이지에 베너를 설치하여 사이언스온 이용자 중 데이터 기반 R&D를 수행하는 연구자로 참여조건을 명시하고 해당 이용자가 베너에 직접 접근하여 응답하는 방식으로 진행되었다. 설문지는 온라인 설문조사 플랫폼인 '왈라(https://home.walla.my)'를 활용하여 제작・배포하였으며, 조사기간은 2023년 9월 12일부터 9월 20일까지 실시하였다. 설문은 총 52명이 최종적으로 응답을 완료하였으며, 수집된 자료는 왈라에서 제공하는 분석 도구를 사용하여 통계적 분석을 실시하였다.

〈표 4〉 설문 내용

영역	설문 문항 내용
응답자 일반 사항	• 연구자의 연구분야, 연구 경력, 소속기관 유형, 직위, 연령대, 최종 학위
연구자의 정보 추구 특성 분석	• 주로 수행하는 R&D 유형 • R&D 중에 많이 활용하는 정보 유형 • R&D 중에 필요하지만 부족하다고 생각하는 정보 유형 • 학술정보 이외에 필요한 정보 유형
연구자의 데이터 추구 및 탐색 행태	• 해당 연구분야에서 주로 사용하는 데이터의 유형 • 해당 연구분야에서 주로 사용하는 데이터의 활용 목적 • 해당 연구분야에서 데이터를 획득하는 방법 • 데이터셋의 소재정보를 파악하는 방법 • 주로 이용하는 데이터의 출처 국가 • 해당 연구분야에서 데이터를 선택하는 기준 • 데이터를 찾거나 활용하는 과정에서 겪는 문제점 • 해당 연구분야에서 데이터를 활용한 연구개발과제의 비중
연구자의 정보원 이용 특성	• 주로 사용하는 정보·데이터원 • 연구에 필요한 데이터셋을 학술정보서비스에서 제공할 경우 사용 의향

4.3 응답자의 일반적 특성

설문 응답자의 주요 연구분야는 공학이 59.6%로 가장 많았으며 사회과학과 자연과학이 각각 11.5%, 복합학(9.6%), 의약학(3.8%), 인문학(3.8%) 순으로 나타났으며, 연구 경력 분포는 6~10년이 38.5%로 가장 많았으며 5년 이하가 26.9%, 11~15년이 17.3% 20년 이상 11.5%, 16~20년이 5.8%로 나타나 주로 10년 이하의 경력을 가진 공학, 사회과학, 자연과학 분야 연구자들이 데이터 기반 R&D를 많이 수행한다는 것을 유추할 수 있었다.

다음으로 연구자의 소속기관 유형은 연구소가 34.6%, 대학교가 25.0%, 공공기관이 15.4%이며 기업체, 초/중/고등학교, 정부기관의 순으로 나타났으며, 직업군은 연구원이 57.7%로 그 비율이 가장 높았으며 대학원생, 교수 및 교사, 회사원, 학생의 순으로 응답자가 많았다. 응답자의 연령대는 30~39세가 53.8%로 가장 높았으며 40~49세가 26.9%, 19~29세가 11.5%의 순으로 나타났고, 응답자의 최종학위는 석사 42.3%, 박사 36.5%, 학사 19.2%로 나타났다(〈표 5〉 참조).

4.4 연구자의 정보 추구 특성 분석

4.4.1 학술정보 추구 목적

연구자들의 정보·데이터 요구의 발생 동기, 정보 추구 목적을 분석하기 위해 연구자의 연구활동 유형(복수응답)을 확인하였다. 연구자가 정보·데이터를 활용하는 주요 이유는 논문 또는 보고서 작성이 27.6%로 가장 높았으며 R&D 과제제안 및 수행(22.8%), 연구개발 중 모델 개발 등 실험 실습(17.1%), 특허출원 및 응용 소프트웨어 개발 등 기술개발 (12.2%), 강의 교재 및 발표자료 개발(10.6%)의 순으로 나타났다. 기타 학문 분야 이론 연구, 정책 발굴 및 정책 실행 계획 개발 등의 응답이 있었다.

〈표 5〉 설문응답자의 인구통계학적 특성

표본특성	분포내용	표본수	비율
주요연구분야	공학	31	59.6
	복합학	5	9.6
	사회과학	6	11.5
	의약학	2	3.8
	인문학	2	3.8
	자연과학	6	11.5
연구경력	5년 이하	14	26.9
	6~10년	20	38.5
	11~15년	9	17.3
	16~20년	3	5.8
	20년 이상	6	11.5
소속기관	공공기관	8	15.4
	기업체	7	13.5
	대학교	13	25.0
	연구소	18	34.6
	정부기관	2	3.8
	초, 중, 고등학교	4	7.7
직업	교사	4	7.7
	교수	4	7.7
	대학원생	7	13.5
	연구원	30	57.7
	학생	3	5.7
	회사원	4	7.7
연령대	19세 미만	1	1.9
	19~29세	6	11.5
	30~39세	28	53.8
	40~49세	14	26.9
	50~59세	2	3.8
	60세 이상	1	1.9
학위	학사	10	19.2
	석사	22	42.3
	박사	10	36.5
	기타	1	1.9

이러한 결과는 연구자들이 R&D 과제를 제안하고 수행하면서 데이터 기반의 실험・실습을 진행하고 그 결과를 논문이나 보고서를 정리하는 연구활동유형이 많고, 그 단계별 특성에 맞게 학술정보를 요구하는 것임을 의미한다고 볼 수 있다.

4.4.2 학술정보 활용도

설문 응답자들이 R&D과정 중에 주로 활용하는 학술정보유형(복수응답)은 해외학술논문

〈표 6〉 연구자의 주요 연구활동 유형 및 학술정보 추구 목적

분포내용	표본수	비율
논문 또는 보고서 작성 등	34	27.6
R&D 과제 제안 및 수행(자체사업, 연구재단 등)	28	22.8
데이터 기반의 실험 실습, 모델 개발	21	17.1
특허 출원, 응용SW 개발	15	12.2
강의 교재 및 발표 자료 개발	13	10.6
이론 연구	7	5.7
정책 발굴 및 실행 계획 개발	5	4.1

〈표 7〉 연구자의 학술정보 활용도

분포내용	표본수	비율
해외학술논문(SCI/SCOPUS급)	34	15.7
국내학술논문	28	12.9
국내연구보고서	23	10.6
해외학술회의/컨퍼런스논문	18	8.3
기술동향보고서	15	6.9
국내학술회의/컨퍼런스논문	16	7.4
강연자료 혹은 발표자료(한글/PPT로 작성된 문서)	14	6.5
데이터셋(실험, 조사, 시계열, 이미지, 영상 등)	14	6.5
국내학위논문	13	6.0
정책동향보고서	11	5.1
해외학위논문	7	3.2
국내특허	6	2.8
블로그, 웹페이지등검색포털(네이버, 구글 등)	6	2.8
해외특허	5	2.3
정책입안자료	4	1.8
실험실습용 소프트웨어(분석소프트웨어, 학습모델 등)	3	1.4

(15.7%), 국내학술논문(12.9%), 국내연구보고서(10.6%), 해외학술회의자료(8.3%), 기술동향보고서(6.9%), 국내학술회의자료(7.4%), 강연자료 혹은 발표자료(6.5%), 데이터셋(6.5%), 국내학위논문(6.0%) 등의 순서로 나타났다. 기타 유형으로 해외학위논문, 국내특허, 블로그, 해외특허, 정책입안자료, 실험실습용 소프트웨어 등을 활용하는 것으로 나타났다.

설문 결과를 살펴보면 전반적으로 연구자들은 학술논문과 학술회의자료를 많이 활용한다고 볼 수 있으며 연구보고서 유형도 많이 활용하는 것으로 보인다. 이는 국내외 학술논문은 연구결과에 대한 공신력과 신뢰성이 보장되기 때문이며, 연구보고서는 R&D과제와 관련하여 연구의 과정과 결과가 포함되어 있어 연구수행 시 참고할 내용들이 많기 때문으로 판단된다.

논문, 연구보고서 이외에 주로 참고하는 정보유형으로 데이터셋과 강연자료 혹은 발표자료와 데이터셋을 선택한 연구자도 다수로 파악되었다. 인구통계학적 특성에서 설문 응답자의 직업 유형에 연구원, 교수, 교사가 많은 점에서 발표자료 혹은 강연자료에 대한 활용이 많은 것으로 판단되며 데이터셋은 최근의 연구 방법이 데이터 기반으로 수행되는 경우가 많아 이에 대한 수요가 있는 것으로 예측할 수 있다.

4.4.3 학술정보 요구도

본 설문은 연구자들이 많이 활용하지만 부족하다고 생각하는 학술정보 유형(복수응답)을 파악하기 위한 것이다. 활용도는 높지만 학술정보서비스에서 찾기 어려워 서비스를 확대하기를 원하는 학술정보 유형으로 가장 많은 비율을 차지하는 것은 데이터셋(실험, 조사, 시계열, 이미지, 영상 데이터 등)이며, 그 다음으로 강연자료 혹은 발표자료, 기술동향보고서, 국내연구보고서, 실험실습용 소프트웨어, 해외학술논문의 순으로 나타났다. 반면, 국내학술논문, 해외학술논문 및 특허정보가 부족하다는 의견은 소수에 불과하였다.

설문 응답자들은 국내연구보고서, 기술동향보고서가 R&D 과정에서 실제 활용도 많지만 여전히 부족하다고 느끼는 것으로 나타났으며 데이터셋과 실험실습용 소프트웨어는 실제 활용 비율에 비해 향후에 필요한 정보·데이터 유형으로 생각하는 연구자가 많은 것으로 보였다. 이는 전통적으로 ScienceON, RISS, KCI 등의 클리어링 하우스 역할을 하는 정보서비스나 국립중앙도서관, 국회도서관 등은 오랫동안 학술논문, 연구보고서, 특허 중심의 학술정보 수집과 구축을 진행하였으며 이에 따라 연구자들은 쉽고 편리하게 이들 정보들을 관성적으로 획득할 수 있었다고 해석할 수 있다.

〈표 8〉 연구자의 학술정보 요구도

분포내용	표본수	비율
데이터셋(실험, 조사, 시계열, 이미지, 영상 등)	19	15.2
강연자료 혹은 발표자료(한글/PPT로 작성된 문서)	14	11.2
기술동향보고서	13	10.4
국내연구보고서	11	8.8
실험실습용 소프트웨어(분석소프트웨어, 학습모델 등)	11	8.8
해외학술논문(SCI/SCOPUS급)	11	8.8
정책동향보고서	9	7.2
해외학술회의/컨퍼런스논문	8	6.4
국내학술회의/컨퍼런스논문	7	5.6
정책입안자료	6	4.8
국내학술논문	5	4
해외특허	5	4
국내학위논문	4	3.2
해외학위논문	2	1.6
국내특허	0	0

반면 연구보고서는 학술논문에 비해 연구의 과정과 결과가 비교적 자세하기 기술되어 있기 때문에 데이터 기반의 연구를 수행하는 연구자들이 많이 선호하는 정보 유형이지만 최근 연구 트랜드에 비추어 여전히 부족한 정보원으로 인식되는 것으로 판단되며, 최근 데이터 중심의 연구개발이 활발한 상황에서 아직까지 연구에 직접적으로 도움이 되는 데이터셋이나 실험 실습용 소프트웨어 등의 제공은 미흡한 것으로 볼 수 있다.

4.4.4 기타 정보 요구도

본 연구에서는 학술정보 이외에 필요로 하는 정보・데이터 유형 중에 펀딩정보, 연구자정보, 연구기관정보, 용어사전, 데이터셋 소재정보, 소프트웨어 소재정보를 대상으로 R&D과정에서 필요한 정도를 리커트 5점 척도 방식으로 선택하는 설문을 하였다.

펀딩정보는 국가R&D과제 및 공공기관, 민간기관 등에서 발주 예정인 과제공고정보를 의미하며, 연구자정보는 연구자 이름, 소속기관, 성과물, 공저자 네트워크 등을 포함하고 있는 정보이다. 연구기관정보는 R&D를 수행하는 기관명, 기관유형, 소속 연구자정보 등을 포함하는 정보이고 용어사전은 분야별로 연구주제가 되는 단어와 그 뜻풀이 정보를 포함하는 정보이다. 데이터셋과 소프트웨어 메타데이터는 R&D에 필요한 데이터셋이나 소프트웨어의 URI(Uniform Resource Identifier) 혹은 URL(Uniform Resource Locator) 등 인터넷 상에서의 고유 리소스 주소를 의미하며 연구자가 해당 데이터셋이나 소프트웨어를 접근하여 다운로드 혹은 열람할 수 있는 위치를 가리킨다.

설문결과 연구자정보에 대한 요구가 제일 높았으며 소프트웨어 메타데이터, 펀딩정보, 연구기관정보, 데이터셋 메타데이터, 용어사전 정보의 순으로 요구도가 높았다. 실제 연구자정보와 연구기관정보는 한국과학기술정보연구원, 국립중앙도서관 등에서 구축하고 있는 정보이지만 연구자 요구가 많은 정보이므로 좀 더 양질의 정보를 구축하여 제공할 필요가 있으며 데이터셋과 소프트웨어의 위치정보를 포함하는 메타데이터를 연구자들이 효율적으로 제공받을 수 있는 구축 관리 방안이 요구된다.

4.5 연구자의 데이터 추구 및 탐색 행태 특성 분석

4.5.1 데이터 활용 목적

연구자 본인 혹은 해당 분야에서 연구목표를

〈표 9〉 기타 정보・데이터의 요구도

분포내용	필요성 평균	4점~5점 비율
펀딩정보	4.23	80.8
연구자정보	4.37	82.7
연구기관정보	4.23	80.8
용어사전	4.10	73.1
데이터셋 메타데이터	4.19	80.4
소프트웨어 메타데이터	4.29	80.4

달성하기 위해 데이터를 활용하는 목적(복수응답)은 〈표 9〉와 같이 다양한 통계데이터를 분석하여 특정 현상이나 경향을 파악하는 통계적 분석, 특정 문제나 상황을 해결하기 위해 실제로 적용할 수 있는 응용연구, 데이터를 기반으로 알고리즘을 학습하거나 AI 기술을 개발하는 기계학습/인공지능을 위해 데이터를 활용하는 것으로 나타났다.

미래의 특정 현상이나 경향을 예측하기 위한 예측모델링, 이론적 또는 개념적 배경을 탐구하고 정책 및 전략을 수립하기 위해 데이터를 활용하는 비율도 높은 편이었다. 즉, 연구자들은 현재 상황을 분석하여 특정 문제나 상황을 해결하거나 미래를 예측하기 위한 목적으로 데이터를 활용하는 경우가 많은 것으로 파악되었다.

4.5.2 활용 데이터 유형

연구자 본인 혹은 해당 분야에서 연구목표를 달성하기 위해 활용하는 데이터 유형(복수응답)은 〈표 10〉과 같이 텍스트로 표현된 텍스트 데이터, 수치로 표현된 숫자형 데이터, 특정 카테고리나 분류에 따른 범주형 데이터 순으로 활용하는 것으로 나타났다.

사진, 스캔된 문서 등의 시각자료로 구성된 이미지 데이터, 시간 순서에 따라 수집된 시계열 데이터, 동영상이나 CCTV 등의 영상데이

〈표 10〉 데이터 활용 목적

분포내용	표본수	비율
통계적 분석	27	18.2
응용 연구	24	16.2
기계학습/인공지능	24	16.2
예측 모델링	20	13.5
기본 연구	17	11.5
정책 및 전략 수립	14	9.5
시장 분석	9	6.1
시뮬레이션	8	5.4
공간적 분석	5	3.4

〈표 11〉 활용 데이터 유형

분포내용	표본수	비율
텍스트 데이터	29	18.7
숫자형 데이터	26	16.8
범주형 데이터	25	16.1
이미지 데이터	23	14.8
시계열 데이터	17	11.0
영상 데이터	13	8.4
지리/공간 데이터	8	5.2
센서/기기 데이터	8	5.2
소셜 네트워크 데이터	6	3.9

터 등 멀티미디어 데이터를 활용하는 연구자도 다수인 것으로 파악되었으나 소셜 네트워크 데이터는 상대적으로 활용이 많이 되지 않은 것으로 나타났다. 이는 설문 응답자의 인구통계학적 특성에서 공학 및 자연과학 연구자가 많고 소셜 네트워크 데이터 분석이 많을 것으로 예상되는 사회과학 연구자의 비율이 많지 않은 점 때문인 것으로 판단된다.

4.5.3 데이터 획득 경로

설문 응답자들이 데이터를 확보하는 방법은 공개 데이터셋을 이용하는 비율이 가장 높았으며 다음으로 직접 수집하거나 구축하는 경우, 논문에 수록된 데이터를 참고하여 추출하는 경우가 많았다. 공개데이터셋을 활용하는 비율이 25% 이상인 점은 앞 절에서의 인터뷰 결과와는 상이한 부분이지만 논문에 수록된 데이터를 추출하여 데이터를 획득하는 점이 상위권에 분포해있는 점은 인터뷰 결과와 유사한 획득 방법이라 할 수 있다.

4.5.4 데이터 소재 파악

설문 응답자들이 데이터 획득을 위해 데이터 소재정보를 파악하기 위한 방법은 〈표 12〉와 같다. 학술논문, 연구보고서를 통한 파악이 30%로 가장 높았으며 데이터의 참조/인용정보를 통한 파악과 Google, Github, Kaggle 등에서 검색해서 파악하는 방법의 순으로 나타났다. 반면 워크샵/세미나, 소셜미디어를 통한 데이터 소재 파악은 높지 않은 것으로 보였다.

〈표 12〉 데이터 획득 경로

분포내용	표본수	비율
공개 데이터셋 이용	32	26.4
직접 수집/구축	27	22.3
논문에 수록된 데이터 추출	22	18.2
컨퍼런스	14	11.6
연구데이터 플랫폼	9	7.4
상용 데이터 구매	7	5.9
스크래핑	6	5.0
자동 데이터 수집	3	2.5
파트너십/협력	1	0.8

〈표 13〉 데이터 소재 파악

분포내용	표본수	비율
학술논문, 연구보고서를 통한 파악	36	30.5
참조/인용 정보	25	21.2
Google, Github, Kaggle 등에서 검색	17	14.4
데이터 제공자/출처와의 직접적인 연락	16	13.6
학술지/컨퍼런스 사이트	15	12.7
워크샵/세미나	7	5.9
소셜미디어 및 블로그	2	1.7

4.5.5 데이터 선택 기준과 활용 과정에서의 문제점

설문 응답자들이 데이터 활용과 관련하여 가장 우선시 하는 점은 〈표 13〉과 같이 주제 관련성, 출처의 신뢰성, 데이터 정확성, 다빈도이용 데이터의 순이었다. 연구자들은 연구주제와 관련성 있는 데이터를 우선적으로 고려한다고 하였으며 출처의 신뢰성도 비중이 높은 것으로 나타났다. 즉 연구관련성과 출처의 신뢰성이 데이터 활용에 있어서 중요하게 생각하는 것으로 분석할 수 있다. 반면 데이터 가격이나 데이터 라이선스에 대해서는 크게 고려하지 않는 요소로 보였다.

연구자 본인이나 주변 연구자들이 데이터를 탐색하고 획득하여 활용하는 과정에서 겪는 문제점에 대해서 연구자들은 원하는 데이터를 찾는데 오래 걸리는 점과 원하는 데이터의 형식과 형태가 제한적이라는 점, 그리고 낮은 데이터 품질을 선택한 연구자의 비율이 높았다.

4.6 정보·데이터원 이용 특성 및 기타 의견

4.6.1 정보·데이터원 이용 특성

설문 응답자들이 연구자의 연구활동을 위해 주로 찾는 정보원을 확인하기 위해 응답자로 하여금 최대 3개 정보원을 선택하도록 하였으며 연구자들이 주로 찾는 정보원에 대한 조사

〈표 14〉 데이터 선택 기준

분포내용	표본수	비율
주제 관련성	39	29.1
출처의 신뢰성	29	21.6
데이터 정확성	17	12.7
다빈도이용 데이터	15	11.2
데이터 최신성	12	9.0
데이터 규모	11	8.2
메타데이터 유무	8	6.0
데이터 포괄성	2	1.5
데이터 가격	1	0.7
데이터 라이선스	0	0.0

〈표 15〉 데이터 획득 및 활용 과정에서의 문제점

분포내용	표본수	비율
원하는 데이터를 찾는데 오래 걸림	32	25.6
원하는 데이터의 형식이나 형태가 제한적	31	24.8
낮은 데이터 품질	22	17.6
데이터 라이선스 및 저작권 이슈	14	11.2
데이터의 출처나 원천에 대한 정보 부족	14	11.2
데이터 구조를 이해하기 어려움	6	4.8
공유 데이터의 피드백이나 리뷰의 부족	4	3.2
데이터 보안	2	1.6

〈표 16〉 연구자의 주요 정보원

분포내용	표본수	비율
학술정보 포털사이트(ScienceON, RISS, DBpia 등)	36	30.8
검색포털(예: 구글, 네이버 등)	37	31.6
국가도서관	11	9.4
동료 연구자	11	9.4
소속기관의 도서관	10	8.5
데이터제공사이트(공공데이터포털, AIHub 등)	8	6.8
개인 소장 정보	3	2.6
기타	1	0.9

결과는 〈표 16〉과 같다.

연구자들이 가장 선호하는 정보원은 ScienceON, RISS, DBPia 등 학술정보 포털사이트나 구글, 네이버 등 검색엔진으로 나타났다. 반면, 국가도서관, 동료 연구자, 소속기관의 도서관, 데이터 제공사이트를 참고한다는 응답자는 약 10% 비율로 나타났다.

연구자 본인이나 주변 연구자들이 데이터를 탐색하고 획득하여 활용하는 과정에서 이용하는 정보원의 유형은 〈표 16〉과 같이 ScienceON, DataON, NTIS 등의 정보・데이터를 동시에 제공하는 KISTI 서비스를 주요 정보・데이터원으로 인식하고 있었으며, 공공데이터포털, 국가통계포털, Github, AIHub 등의 순서로 정보・데이터 제공 서비스를 활용하는 것으로 나타났다. 참고로 〈표 15〉와 〈표 16〉의 설문에서 ScienceON, DataON, NTIS 등을 정보・데이터원으로 많이 활용한다는 결과는 설문 대상자가 KISTI 이용자임을 감안하여 해석할 필요가 있으며, 추후에 데이터원을 분야별로 세분화하여 많이 활용하는 정보・데이터원을 분석하는 것이 필요할 것이다.

〈표 17〉 주로 사용하는 정보・데이터원

분포내용	표본수	비율
ScienceON, DataON, NTIS 등	36	22.8
공공데이터포털	34	21.5
국가통계포털	20	12.7
Github/GitLab	16	10.1
AIHub	14	8.9
Google Dataset	13	8.2
Kaggle	7	4.4
data.gov	6	3.8
모두의 말뭉치	4	2.5
네이버 데이터랩	4	2.5
서울시열린데이터광장	3	1.9
AWS 클라우드	1	0.6

4.6.2 기타 의견

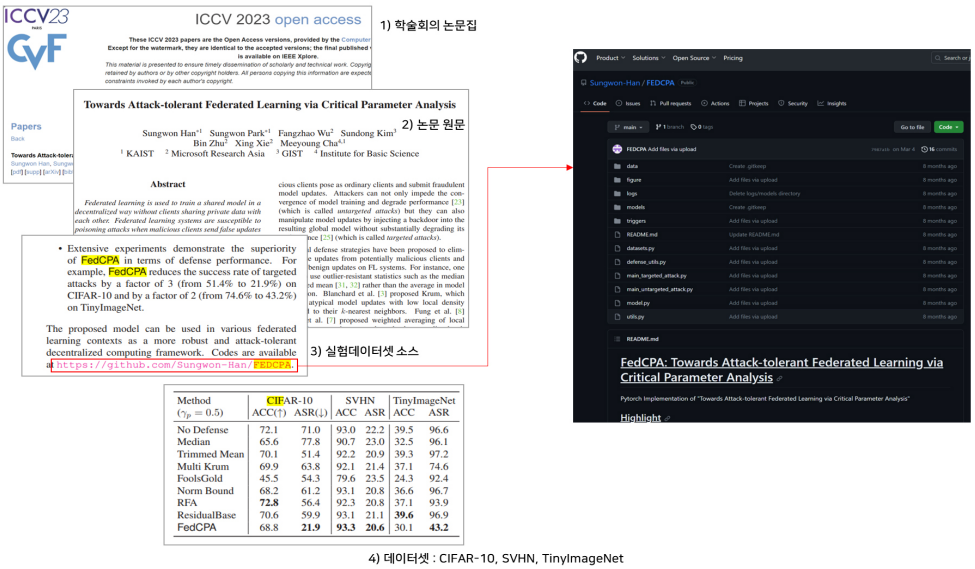
연구자들은 필요한 데이터셋이나 학습모델과 같은 소프트웨어가 어디에 있고 어떻게 확보해야할지 모르는 경우가 많아 데이터 기반 R&D를 수행하는데 애로사항이 많다는 의견이었다. 이에 분야별로 데이터셋이나 학습모델 관련 정보를 요약해서 제공하기를 원하는 의견이 있었으며, 향후에 데이터 기반의 R&D가 확대될 것이므로 적극적으로 데이터셋이나 소프트웨어를 수집하여 학술정보와 연계하여 서비스하기를 요구하는 의견도 포함되어 있었다.

ScienceON과 같은 학술정보서비스에서 데이터셋과 소프트웨어의 메타데이터를 별도로 제공할 경우 사용할 의사가 있는지에 대한 질문에 63%의 응답자가 적극적으로 사용할 의사가 있다고 답했으며 이용할 의사가 있다는 응답이 전체의 84% 비율에 달하는 것으로 나타나서 논문, 보고서 등에 포함되어 있는 데이터셋의 메타데이터를 분야별, 혹은 주제별로 모아서 제공하는 것을 요구하고 있음을 확인하였다.

4.7 시사점

데이터 기반 R&D를 효율적으로 수행하기 위해 주제 영역별로 공신력 있는 데이터셋의 소재정보가 제공될 경우 적극 이용할 것이라는 응답결과와 연구에 필요하고 적합한 데이터셋을 찾는 과정이 생각보다 어렵고 오래 걸리기 때문에 학술정보서비스에서 데이터셋 안내 기능을 제공해 준다면 연구에 많은 도움이 될 것이라는 의견에 따라 본 연구에서 데이터 기반 R&D를 지원하는 서비스 사례로 논문에서 데이터셋/소프트웨어 소재정보를 추출하여 제공하는 방안을 제시하고자 한다.

〈그림 1〉은 데이터셋과 소프트웨어 링크정



〈그림 1〉 데이터셋과 소프트웨어 링크정보를 포함하는 논문의 예

보를 포함하는 논문의 원문을 보인 것으로 학습모델 소스코드와 관련 데이터셋을 Github에 공유하고 논문에 링크정보를 게시하였다. 데이터셋과 소프트웨어 정보가 포함된 논문 원문을 분석하여 관련정보를 추출하고 메타데이터로 구축하여 주제분야별, 연구주제별 등으로 카테고리화 하여 학술정보 서비스로 제공한다면 연구자들이 데이터셋 소재를 파악하는데 유용하게 활용될 수 있을 것이다.

5. 결론 및 제언

본 연구에서는 연구자의 데이터 기반 R&D를 효율적으로 지원하기 위해 새로운 학술정보 유형과 데이터셋을 발굴하고 학술정보서비스의 방향을 제시하기 위한 선행 연구로서 연구자가 필요한 학술정보와 데이터 요구 사항을 분석하였다. 연구의 목적을 달성하기 위한 연구방법은 데이터를 이용하여 R&D를 수행하는 연구자 5명을 선정하여 탐색적 사례 연구를 통해 R&D 과정에서 활용하는 학술정보 유형과 활용 범위와 정보・데이터 요구사항을 파악하였으며, ScienceON 이용자를 대상으로 추가 온라인 설문을 실시하여 데이터 기반 R&D 행태 및 요구사항을 도출하였다.

연구자 면담에서 파악한 데이터 기반 R&D 행태의 특성은 첫째, 데이터 기반 R&D를 위한 연구주제선정이 연구자의 학력, 전공분야, 관심주제 등에 따라 이루어지며 특히 정부나 공공기관, 기업의 R&D 방향, 코로나 등과 같은 사회적 이슈에 따라 설정이 되며, 최근 인공지능과 빅데이터를 활용한 디지털 전환 노력 등으로 데이터 기반 R&D 과제 수요가 많다는 점이었다. 둘째, 피면담자들은 데이터 기반 R&D를 수행하는 과정에서 국내외 학술논문과 학술회의자료를 많이 참고하는 것으로 나타났다. 분야별 연구주제의 특성에 따라 국내정보와 해외정보를 선호하는 경향이 달랐으나 전체적으로 학술논문을 주로 참고하고 보고서 및 기타 자료를 참고하는 경우는 없었다. 셋째, 데이터나 학습모델등 소프트웨어를 획득하기 위해 선행연구자료인 학술논문 본문을 확인하는 경우가 대부분이었으며 공공데이터포털, 주제분야별 전문정보센터, 국가 빅데이터 포털 등 공공데이터를 활용하는 경우는 많지 않은 것으로 파악되었다.

또한, 데이터 기반 R&D를 효율적으로 수행하기 위해 주제 영역별로 공신력 있는 데이터셋의 소재정보가 제공될 경우 적극 이용할 것이라고 응답하였으며 연구에 필요하고 적합한 데이터셋을 찾는 과정이 생각보다 어렵고 오래 걸리기 때문에 학술정보서비스에서 데이터셋 안내 기능을 제공해 준다면 연구에 많은 도움이 될 것이라는 의견이 많았다. 그리고, 데이터 분석 시에 데이터의 정확성과 신뢰성를 보장할 수 있는 상용 데이터를 적극적으로 활용할 의사도 있는 것으로 확인되었는데 비용 부담을 고려하여 데이터의 정확성과 신뢰성을 공공기관이 보증해주는 절차나 방법이 필요하다는 의견도 제시되었다.

시례 연구를 통해 도출된 주요 시사점을 설문지에 반영하여 ScienceON이용자를 대상으로한 학술정보 및 데이터 요구사항 수렴 결과는 다음과 같다. 첫째, 연구자들은 논문과 보고서 작성, R&D 과제 제안 및 수행을 위해 학술

정보와 데이터를 추구한다고 응답하였으며 해외학술논문, 국내학술논문, 국내연구보고서, 해외학술회의자료, 기술동향보고서 등의 순으로 학술정보를 활용하는 것으로 나타났다. 둘째, 연구자들이 많이 활용하지만 부족하다고 생각하는 학술정보 유형으로 데이터셋, 강연자료 및 발표자료, 연구보고서의 순으로 나타나 데이터셋의 수요가 증가하고 있음을 확인할 수 있었다. 셋째, 연구자의 데이터 추구 목적은 통계분석, 응용연구, 기계학습/인공지능, 예측 모델링을 위한 것으로 나타났고, 주로 활용하는 데이터 유형으로는 텍스트 데이터, 숫자형데이터, 범주형 데이터, 이미지 데이터 등을 주로 활용하는 것으로 분석되었다.

연구자들은 공공기관이나 민간기관에서 공개한 공개 데이터셋을 다운받거나 연구 목적에 맞는 데이터셋을 직접 수집, 구축하는 방식, 그리고 논문에 수록된 데이터를 추출하는 방식 등으로 원하는 데이터셋을 확보하였으며, 데이터 소재를 파악하는 방법은 학술논문, 연구보고서를 통한 파악, 데이터 인용 정보를 확인하는 방식, Google, Github, Kaggle 등에서 검색하여 파악하는 방식으로 확인되었다.

연구자의 심층면담과 온라인 설문을 통해 최근 대부분의 학문 분야에서 데이터를 활용한 R&D 활동이 증가하고 있으며, 데이터셋과 소프트웨어의 확보를 위해 학술논문이나 학술회의자료 등 학술정보를 많이 참고한다는 점을 확인하였다. 주제 분야별로 활용하는 데이터 확보 방법과 획득 경로, 활용 데이터 유형이 상이하며 연구자들은 필요한 데이터셋이나 학습모델과 같은 소프트웨어가 어디에 있고 어떻게 확보해야할지 모르는 경우가 많아 데이터 기반 R&D를 수행하는데 애로사항이 많은 것으로 나타났다.

본 연구의 요구 분석 결과를 고려하여 앞으로 국내외 학술정보를 구축하거나 신규 학술정보원을 발굴하기 위해 학술정보 및 연구보고서 등의 본문에서 데이터셋과 소프트웨어 관련 내용을 인식하여 주제 분야나 주제 키워드 별로 활용되는 데이터셋 정보를 요약하여 구축하는 방법이 필요하다. 최근 데이터셋 제작 및 구축, 모델의 성능 비교 분석, 모델 개발 등을 다룬 논문들이 증가하고 있어 이들 논문에서 데이터셋, 소프트웨어의 이름, 특성, 위치정보 등 메타데이터를 추출・생성하여 주제별로 분류한 정보를 제공할 경우 연구자에게 유용할 것으로 판단된다. ScienceON과 같은 학술정보서비스에서 데이터셋과 소프트웨어의 위치정보를 별도로 제공할 경우 사용할 의사가 있는지에 대한 질문에 63%의 응답자가 적극적으로 사용할 의사가 있다고 답했으며 이용할 의사가 있다는 응답이 전체의 84% 비율에 달하는 점은 데이터셋, 소프트웨어 요약정보의 필요성을 의미한다고 볼 수 있다.

본 연구의 설문 결과에서 살펴본 바와 같이 데이터 기반의 R&D를 수행할 때 분야별, 연구목적별, 주제별로 활용하는 정보・데이터원이 다양할 것이라 점을 예측할 수 있다. 이에 향후에는 분야별, 연구목적별, 주제별로 연구자들의 정보・데이터 활용 행태를 면밀히 분석하여 서비스로 제공하는 방안이나 연구자별 지원 방안을 모색하여야 할 것이다.

참 고 문 헌

권나현, 이정연, 정은경 (2012). 과학기술분야 R&D 전주기 연구: 국내 생명 및 나노과학기술연구자를 중심으로. 한국문헌정보학회지, 46(3), 103-131. https://doi.org/10.4275/KSLIS.2012.46.3.103

김나연, 정은경 (2020). 사회과학 분야 연구자의 데이터요구와 데이터 재이용 행위에 관한 연구. 정보관리학회지, 37(4), 1-26. https://doi.org/10.3743/KOSIM.2020.37.4.001

나은엽 (2023). 자연과학분야 연구자들의 연구계획서 작성을 위한 정보추구행동의 탐색적 연구. 한국비블리아학회지, 34(1), 53-74.

심원식, 안혜연, 박규리, 송사광, 임형준 (2023). 개방형 연구 커먼즈에 대한 연구자 요구 분석에 관한 연구. 한국문헌정보학회지, 57(4), 209-232. https://doi.org/10.4275/KSLIS.2023.57.4.209

심윤희, 김지현 (2019). 국내 대학도서관의 연구데이터관리서비스 개발 방안에 관한 연구: 서울대학교 소속 연구자들의 요구 분석을 중심으로. 정보관리학회지, 36(3), 61-80. https://doi.org/10.3743/KOSIM.2019.36.3.061

아주대학교 산학협력단 (2018). 지능형 큐레이션 서비스를 위한 이용자 행태 분석. 대전: 한국과학기술정보연구원.

유주현, 조상민, 김동현 (2017). 4세대 R&D 패러다임 전환과 제도설계. 국정관리연구, 12(2), 1-24. https://doi.org/10.16973/jgs.2017.12.2.001

이란주, 김수진 (2015). 주제별 연구자의 정보이용행태에 관한 선행연구 분석. 한국비블리아학회지, 26(2), 129-153. https://doi.org/10.14699/kbiblia.2015.26.2.129

한국과학기술정보연구원 (2022). 지능형 과학기술정보 큐레이션 체제 구축. 대전: 한국과학기술정보연구원.

한종엽, 서만덕 (2014). 해양과학기술 분야 연구자의 정보이용행태에 관한 연구. 정보관리학회지, 31(1), 163-187. https://doi.org/10.3743/KOSIM.2014.31.1.163

Gregory, K. (2020). A dataset describing data discovery and reuse practices in research. Scientific Data, 232(2020), 1-11.

Merriam, S. B. (1998). Qualitative Research and Case Study Applications in Education. San Francisco: Jossey-Bass Publishers.

Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. Journal of Documentation, 74(5), 1053-1073.

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

Ajou University Industry-Academic Cooperation Foundation (2018). User Behavior Analysis for

Intelligent Curation Services. Daejeon: Korea Institute of Science and Technology Information.

Han, Jong-Yeop & Seo, Man-Deok (2014). A study on the information usage behavior of researchers in the field of ocean science and technology. Journal of the Korean Society for Information Management, 31(1), 163-187. https://doi.org/10.3743/KOSIM.2014.31.1.163

Kim, Na-Yeon & Jung, Eun-Kyung (2020). (2020). An investigation on data needs and data reuse behavior in the field of social sciences. Journal of the Korean Society for Information Management, 37(4), 1-26. https://doi.org/10.3743/KOSIM.2020.37.4.001

Korea Institute of Science and Technology Information (2022). Construction of an Intelligent Science and Technology Information Curation System. Daejeon: Korea Institute of Science and Technology Information.

Kwon, Na-Hyun, Lee, Jung-Yeon, & Jung, Eun-Kyung (2012). Understanding scientific research lifecycle: based on bio- and nano- scientists' research activities. Journal of the Korean Society for Library and Information Science, 46(3), 103-131. https://doi.org/10.4275/KSLIS.2012.46.3.103

Lee, Lan-Ju & Kim, Su-Jin (2015). A study on the literature review of information use behavior in specialized fields. Journal of the Korean BIBLIA Society for Library and Information Science, 26(2), 129-153. https://doi.org/10.14699/kbiblia.2015.26.2.129

Na, Eun-Yeop (2023). An exploratory study of natural scientists' information seeking behavior when writing a research proposal. Journal of the Korean Biblia Society for Library and Information Science, 34(1), 53-74.

Shim, Won-Sik, Ahn, Hye-Yeon, Park, Gyu-Ri, Song, Sa-Kwang, & Lim, Hyung-Jun (2023). A study on the researchers' needs for open research commons. Journal of the Korean Society for Library and Information Science, 57(4), 209-232. https://doi.org/10.4275/KSLIS.2023.57.4.209

Shim, Yoon-Hee & Kim, Ji-Hyun (2019). A study on the development of research data management service in a domestic university library: focused on the analysis on the needs of researchers affiliated in Seoul National University. Journal of the Korean Society for Information Management, 36(3), 61-80. https://doi.org/10.3743/KOSIM.2019.36.3.061

Yu, Ju-Hyun, Jo, Sang-Min, & Kim, Dong-Hyun (2017). 4th generation R&D paradigm shift and institutional design. Journal of Governance Studies, 12(2), 1-24. https://doi.org/10.16973/jgs.2017.12.2.001